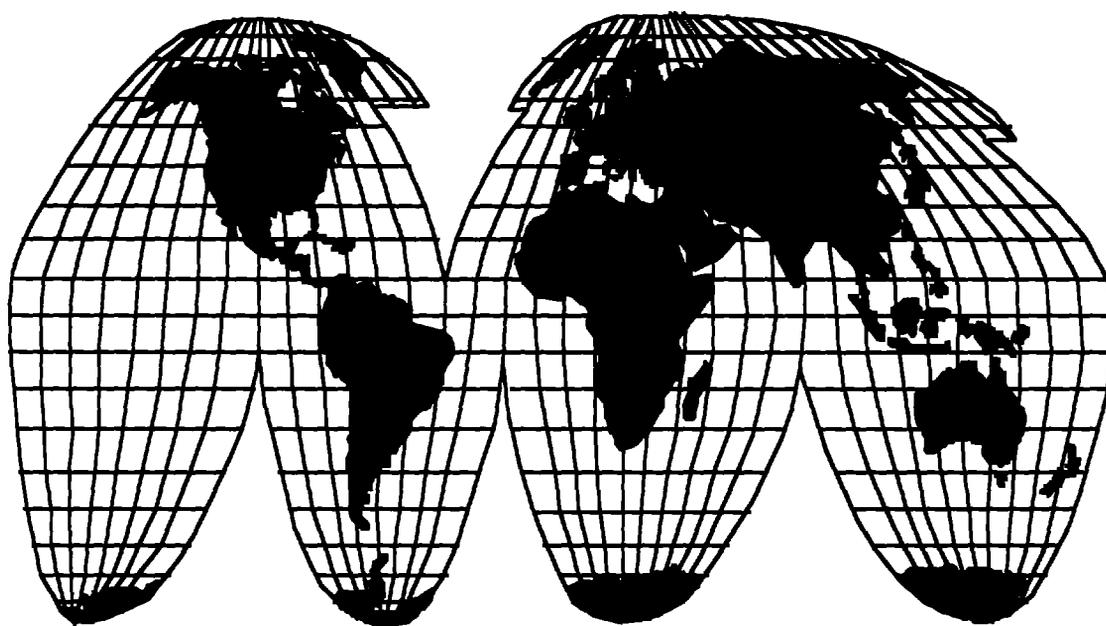


- 15372 -

# **Regulatory Policies and Reform: A Comparative Perspective**

**Claudio R. Frischtak (ed.)  
December 1995**

**Pre-publication Edition**



**Private Sector Development Department  
The World Bank**



**Regulatory Policies and Reform:  
A Comparative Perspective**

Claudio R. Frischtak (ed.)

Private Sector Development Department  
The World Bank



---

# Preface

In the last 20 or so years there has been a continuing debate on the role and importance of the state and private sector for economic development. The demise of the import substitution model, which accorded the state a central position, and the progressive extension of markets and entrepreneurial capacity swung the pendulum toward a somewhat simplistic view: a “minimal state”—concerned, in the economic sphere, fundamentally with the defense of property rights and the enforcement of contracts—would be sufficient to establish the basis for sustainable growth.

In most developing and emerging market economies there has been a realization that the state has overextended its economic reach, and in so doing, has woven a regulatory net that is inconsistent with the requirements of a competitive economy. Agents are thwarted in their search for new opportunities, markets remain thin, and economies tend to be isolated from global resource flows. Deregulation—removing policy and regulatory barriers to mobility and competition—has appropriately become a prerequisite for development.

Yet even in a deregulated environment the efficiency of markets is not always guaranteed. Market outcomes do

not necessarily further the broad public interest. This volume discusses the regulatory mandate of the state in product and factor markets, in environmental protection, and in infrastructure. It suggests that, although there is no need for a large or intrusive state, there is a need for one that can effectively target and correct the most significant market failures and imperfections. State resources should focus in few critical regulatory areas, with relative emphasis depending on the maturity of markets and institutions.

This volume has been produced under the auspices of the Private Sector Development Department of the World Bank. It is part of an effort to facilitate understanding of the preconditions for sustainable economic development in a private sector–led economy. Regulatory policies and oversight mechanisms that function in accordance with the public interest contribute to establishing solid foundations for such an economy. I am confident that the papers offered in this volume will be of interest to both policymakers and the generally informed public.

*Magdi R. Iskander*  
*Director*  
*Private Sector Development*

---

# Contents

Preface	iv	
Acknowledgments	v	
Contributors	vi	
1	The changed role of the state: regulatory policies and reform in a comparative perspective	1
	<i>Claudio R. Frischtak</i>	
2	The instruments of competition policy and their relevance for economic development	16
	<i>R. Shyam Khemani and Mark A. Dutz</i>	
3	Competition policy and institutions in reforming economies	38
	<i>Roger Alan Boner</i>	
4	Competition issues beyond trade liberalization: distribution and domestic market access	69
	<i>Mark A. Dutz and Sethaput Suthiwart-Narueput</i>	
5	Antidumping policy and competition	85
	<i>Sadao Nagaoka</i>	
6	The basics of consumer protection: principles and policies	104
	<i>Eduardo Engel</i>	
7	Beyond the basics of consumer protection	131
	<i>Eduardo Engel</i>	
8	Bankruptcy policy	144
	<i>Izak Atiyas</i>	
9	Labor policies and regulatory regimes	170
	<i>Zafiris Tzannatos</i>	
10	The case of land markets	191
	<i>Antônio Salazar P. Brandão and Gershon Feder</i>	
11	Incentive regulation: market-based pollution control for the real world?	210
	<i>Raymond S. Hartman and David Wheeler</i>	
12	Regulatory policies and reform in telecommunications	236
	<i>Ioannis N. Kessides</i>	
13	Competition and regulation in the railroad industry	255
	<i>Ioannis N. Kessides and Robert D. Willig</i>	
14	Regulatory policies and reform in the electricity supply industry	279
	<i>David M. Newbery</i>	
15	Regulating the power sector	307
	<i>Anthony Churchill</i>	

---

# Contributors

Izak Atiyas, The World Bank  
Roger Alan Boner, Federal Trade Commission  
Antônio Salazar P. Brandão, Fundacao Getulio Vargas, Brazil  
Anthony Churchill, Washington International Energy Group  
Mark A.Dutz, The World Bank  
Eduardo Engel, University of Chile  
Gershon Feder, The World Bank  
Claudio R. Frischtak, The World Bank  
Raymond S. Hartman, UC-Berkeley  
Ioannis N. Kessides, The World Bank  
R. Shyam Khemani, The World Bank  
Sadao Nagaoka, Seikei University, Japan  
David M. Newbery, Cambridge University, UK  
Sethaput Suthiwart-Narueput, The World Bank  
Zafiris Tzannatos, The World Bank  
David Wheeler, The World Bank  
Robert D. Willig, Princeton University

---

# Acknowledgments

This volume is the product of a collective effort. I would like first to thank the authors for their enthusiasm, hard work, and perseverance, in a project whose duration exceeded our original expectations. I hope the results will contribute to the debate on the roles of the state and private sector in economic development. I would also like to thank my colleagues at the World Bank for their support throughout this project. Carl Dahlman and John Page

deserve special thanks, as does the Private Sector Development Department and its director, Magdi Iskander. Amy Chan and Cindy Wong shared the critical role of bridging the geographical gap between authors and editor and monitoring the myriad details necessary to bring this volume to completion. Finally, I would like to thank my family, and Leila in particular, for their support.

# The changed role of the state: regulatory policies and reform in a comparative perspective

Claudio R. Frischtak

During the last forty or so years, the state in most developing countries attempted to emulate a Schumpeterian engine of growth, mobilizing resources to invest directly in productive activities or steering those resources to specific sectors. A complex array of protective policies, promotional instruments, and regulatory controls led to the emergence of industries and entire social segments—entrepreneurs and rentiers. Although industrial growth was the object of policy, industrial resources also became the instrument of employment creation, regional balance, and other equity-related purposes. The quest for industrialization in the context of postwar reconstruction, decolonization, and independence gave political impetus to this pattern of state action; the predominance of import-substitution industrialization models in the development literature provided the intellectual underpinning and the economic rationale for an activist state.

For most countries this model of state action as the vanguard of economic development has become dated for two fundamentally distinct reasons. First, the growth of an entrepreneurial class, the extension of markets, and the reduction of transactions costs—arguably outcomes of state efforts to bestow endowments and stimulate agents to enter emerging markets—lessened the need for either direct or indirect state involvement in productive activities. Moreover, and in contrast to state's earlier role of inducing entry and thereby market dynamism, policy and regulatory activism became associated in many countries with the protection of entrenched incumbents from the challenges posed by new competition. Rent-seeking behavior was stimulated and sanctioned by the instruments of protection, promotion and regulation. The state was increasingly perceived as having been captured by its own creations; and in this process, the state reached its fiscal limits.

Second, there were exogenous factors contributing to the exhaustion of a paradigm of state action characterized by pervasive regulation of economic agents and allo-

cation of their productive activities. The rapid pace of technological progress, the globalization of economic relations, and the increase in the value of information, signifies that the competitive standing of firms and countries is predicated on their ability to respond quickly to market and technological trends. Intrusive regulatory regimes, plodding government agencies, or opaque rules and regulations are fundamentally incompatible with a more open and competitive market environment.

Just as the import-substitution model of industrialization had conferred respectability to an activist state, its criticism called for and became the justification for other models of state action. Advocates of the "minimal state" argued that the state's functions were to administer justice, establish property rights, and enforce contracts.<sup>1</sup> Those searching for an alternative developmental model, inspired on the East Asian "miracle" economies, perceived the state's role more broadly: an agent for "crowding in" investment in areas commanding high social returns; a catalyst for cooperation among social groups; an arbiter of economic conflict; a partner for economic development.

## Groping for a new paradigm

What then should be the scope of action for a contemporary developmental state? At the most basic level, the state should provide a set of rules and market-supportive institutions to assign property rights, enforce contracts, and establish a stable economic environment so that productive activity can flourish. A property law, a judicial system with enforcement powers, and a credible monetary regime are public goods supplied by the state. Stability, predictability, and transparency of policies are the basis of sound government action in the economic sphere. Beyond these precepts, which both a minimalist and a developmentalist perspective would abide by, there are three fairly distinct sets of economic functions that differentiate the modes of state action: investment, policy, and regulation.

In terms of its investment functions, a development-oriented state first would attempt to complement or “crowd in” private investment, by providing physical and other supportive infrastructure.<sup>2</sup> Second, it would target areas that present large externalities and where the private sector would normally underinvest, for example, basic education, preventive health care, and other segments of the social infrastructure. Third, this state would be involved in the production of “merit” goods, which would be available to all members of society regardless of their income.

As for its policy functions, the developmental state is expected to provide a stable and predictable macroeconomic environment. To do so requires an unwavering commitment to defend the national currency from inflation, to maintain the public trust in financial institutions, and to avoid sudden policy shifts (except as necessary to guarantee the foundations of the monetary regime).<sup>3</sup>

From a microeconomic perspective, the presence of well-informed agents contributes to the stability and efficiency of markets. A key policy function of this state therefore is providing information to financial, labor, and other markets. Industrial policy, in particular, is increasingly perceived as being grounded on the dissemination of market-relevant information. Indeed, among the most important of the initiatives of the new developmental state is the creation of bodies which promote a structured exchange of information between the public and private sectors, so as to signal producers that entry is profitable, reduce uncertainty, and make commitments credible.

In this context, while private actors commit to specific projects or performance targets, the government would undertake to remove major obstacles to private sector activity: regulatory and bureaucratic barriers, infrastructure bottlenecks, weaknesses in technology delivery systems, and gaps in education and training. Information and coordinating externalities would provide the economic rationale and underpin these arrangements.

The scope of regulatory activities encompasses two distinct tasks. Effective regulation can rarely be implemented without first removing regulatory obstacles to the improved allocation and use of resources. This chapter describes briefly the rationale and substance of deregulation, and the changes that this process normally entails in both the domestic policy arena and the international trade regime. A modern regulatory regime equally presupposes the buildup of a regulatory system. In this regard, it is necessary to define the objectives of regulatory intervention and establish the analytical basis for an appropriate regulatory regime within which markets can

function efficiently and consistent with the public interest.

The extensive discussion of regulatory regimes presented in this volume stresses certain recurrent themes and outstanding lessons. An attempt has been made to summarize them in this chapter and to glean some of the main principles of a modern regulatory regime. One important conclusion is that the considerable regulatory requirements of a modern economy do not imply a large state; they do demand, however, institutions capable of effectively bringing about deregulation and introducing new regulatory structures, consistent with the needs of a market economy. In establishing effective regulatory regimes and market-supportive rules and institutions, the state delimits its own boundaries; the absence of such mechanisms, by contrast, facilitates the state’s inordinate expansion.

### **The challenge of deregulation**

A complex regulatory maze survives in most countries where the state had an important role in the promotion and protection of economic activity. In the early stages of industrialization, such regimes may have promoted entry and stimulated investment—by helping mobilize resources, providing the institutional means to direct them to areas with important external economies (particularly in infrastructure), while attracting private agents to still thin industrial markets. Yet these same mechanisms and rules came to impose over time barriers to resource mobility and competition.

Driving those regulatory barriers were rules that generally privileged incumbents at the expense of entrants. Information asymmetries also played a role, with insiders often having detailed knowledge of the unwritten norms and the individuals who implemented them. Relatively rigid market configurations emerged, because rents accruing to incumbents became a strong disincentive for such firms to penetrate new, untested arenas. By constraining flexibility in resource allocation and use, and by limiting competition, regulatory barriers—including those restricting the operation of factor markets or resulting from outdated models regulating the supply of infrastructure services—brought about efficiency losses, in both a static and an intertemporal sense. Further, insofar as such barriers benefited incumbents more often than not, they contributed to a less equitable distribution of opportunities, income, and wealth.

Many countries responded to such regulatory constraints through deregulation, that is, by changing the set of rules that framed economic activity, thereby removing

barriers to competition, factor mobility, and firm growth. The object was to increase the flexibility and speed with which economic agents redeployed resources in response to changes in markets and technologies, and ultimately to improve the economy's efficiency and distributional outcomes. In this sense, deregulation has the same basic purpose as efforts to institute a modern regulatory regime.

Deregulation is not a trivial process. It involves reforming or phasing out instruments and mechanisms to which markets and agents have adapted. Moreover, for every piece of legislation and rule—no matter how outdated or incompatible with the public interest—there is a constituency, often allied or gravitating around the implementing agency, ready to defend the *ancien régime*. They become the core of counter-reform. In view of the interests created around the regulatory web, and (often) the lack of knowledge concerning its extension and impact, reform tends to be slow and generally following difficult and protracted negotiations.<sup>4</sup>

Investment licensing, ubiquitous in industrializing countries, provides a useful illustration. Licensing the creation and expansion of capacity has often precluded potential competition, encouraged entry-detering behavior by established producers, and reduced actual competition by constraining supply. Some countries have used licensing to control entry or expansion of multinational corporations into specific segments or ensure that local investors attain a minimum equity share. Not infrequently, licensing simply constituted a protective barrier for any incumbent—national or transnational—against all entrants, irrespective of their national origin.<sup>5</sup> Despite its adverse impact on competition and the creation of new economic opportunities, investment licensing remains an important instrument of policy in numerous countries.

The removal of investment licensing and other regulatory barriers to competition in domestic markets should be high on the regulatory reform agenda of governments and their competition policy agencies. In fact, a major statutory role of competition policy agencies should be to comment and expose—and, with the force of law, change—those policies and regulations that affect competition adversely. In so doing, agencies would be undertaking an advocacy function in the public interest (chapters 2 and 3 for further discussion).

Liberalizing the trade regime—by progressively removing tariff and nontariff barriers, reducing its anti-export bias, and increasing import competition—constitutes the basic agenda for the deregulation of the international trade regime. It complements and reinforces deregulation efforts in domestic markets. One

important class of barriers to international trade is vertical restraints to domestic market access. These restraints—which range from government-granted monopoly rights in distribution to poorly defined property rights which discourage long-term investment in facilities—prevent domestic and international prices from converging, muting the positive impact of trade liberalization (chapter 4).

Antidumping and countervailing duties are an additional constraint to international trade. Their importance has grown significantly with the progressive removal by most countries of traditional instruments of protection, and in view of the legal and political difficulties of reintroducing them. The imposition of antidumping duties for reasons other than predatory dumping should be avoided (chapter 5). The welfare benefits derived from such policies are generally dominated by losses incurred by consumers, and in the longer term, by the disincentive for producers to restructure away from uncompetitive activities.

Regulatory constraints affect not only product markets but the operation of factor markets as well. In case of labor markets, job security regulations that forbid collective layoffs or substantially increase obstacles to adjustment in labor quantities may constitute a major obstacle for increased labor absorption. By undermining labor discipline, such regulations make labor *de facto* more costly. Insofar as they block exit from unprofitable activities, they make it more difficult for firms to eliminate product lines and scrap older plants, thereby hampering modernization of the economy.

Although some have blamed sluggish job creation and low employment levels in the formal sector of several economies on such regulations, others are less sanguine about the case for eliminating them altogether (chapter 9). The reason is that partial equilibrium analysis does not provide the basis for deriving economywide implications. It might help, however, to predict the “positive” impact of labor market deregulation on specific subsectors. Moreover, the very notion of flexibility should be broadened: Employment-at-will, the paradigm of hiring and firing practices associated with employment creation, is just one aspect of job flexibility. Possibly more important for employment and productivity performance is the functional flexibility typically found in the Japanese large firm system. In this perspective, the focus of deregulation should be the work rule rigidities that impede labor from multitasking across a wide variety of functions.

In land markets, deregulation should be centered on the removal of formal restrictions on the exercise of prop-

erty rights that hamper or prevent individuals from transferring and renting land (chapter 10). By constraining more productive farmers from gaining access to land and credit markets, these restrictions have far-reaching implications for efficiency, investment, and growth. A subclass of such restrictions are rent controls, as well as prohibitions on share tenancy or the imposition of upper limits on the landowner's share. They stimulate eviction of tenants and steer landlords to specialize in activities less demanding in supervision and risk. Generally, they are introduced at the cost of output losses, in addition to less investment in land improvements. The effects of rent controls in housing markets are similar: they tend to stimulate eviction and dampen investment.

In physical infrastructure, the wedge of deregulation has been driven by changes in technology, a reexamination of the natural monopoly characteristics of the industry, and a new understanding of the behavior of agents in response to the challenge posed by entry in structurally contestable markets. Technological progress has been most dramatic in telecommunications, with the industry "experiencing the throes of entry, rivalry, and diversity" (chapter 12). Structurally competitive interchange markets permit vertical separation of the industry and deregulation of the prices of interchange services.

The seminal notion of "separating firms vertically in order to segregate portions needing regulation from those that do not because of their degrees of competition or contestability" should equally inform deregulation efforts in the railroad industry (chapter 13). Once vertical separation is attained, deregulation implies freedom of pricing (as well as entry and exit) for competitive activities. Elsewhere, proper regulatory oversight is called for.

Vertical separation should also be considered in the electricity supply industry, where technical progress in conventional generation and potentially competitive entry of new combined cycle gas turbine technology, among other factors, challenged the traditional paradigm of the vertically integrated national (or regional) monopoly. Although high-tension transmission and low-tension distribution remain natural monopolies, generation and supply of end customers are potentially competitive, possibly allowing for vertical separation and deregulation (chapter 14).

A common base for deregulation in the infrastructure industries clearly exists. It calls for identifying and separating monopolistic from competitive elements of the industry, on the presumption that eventual efficiency losses in economies of vertical integration will be more than compensated by gains from deregulation of entry

and prices and introduction of competition in structurally contestable segments of the market. For smaller (or thinner) markets, and systems at an early stage of development, however, a combination of feeble potential competition and large finance requirements would make these reforms less attractive. Nonetheless, deregulation of certain segments of the infrastructure industry seems to be an irresistible trend in many countries, industrial and developing alike. The scope of reform and the content of the new regulatory regime—more than its basic orientation—are the real objects of discussion.

### **Efficiency and equity as regulatory objectives**

Deregulation is not sufficient to ensure that markets perform efficiently and that their outcomes are reasonably equitable. Markets require rules to orient the behavior of agents and institutions to support their development. In particular, although deregulation is often consistent with the goal of distributive justice, it is not instrumental to it.<sup>6</sup> With deregulation comes the need to establish a regulatory framework based on principles of efficiency and distributive justice and in line with the public interest.

On efficiency grounds, there are two major reasons for regulatory activism: nonconvexities in production and consumption, a technological datum, "external" to markets; and market gaps and failures, which are a direct function of high transactions costs. Lowering such costs generally depends on institutional innovations and technological change. Whereas increasing returns (a result of nonconvex technologies or preferences) generate imperfect markets—where few agents compete and the outcomes of which (in terms of prices and quantities) are in most cases suboptimal<sup>7</sup>—positive transactions costs ultimately explain market failures or the nonexistence of markets.<sup>8</sup>

Particularly important from a welfare perspective are externalities. Among the most important externalities for economic activity are those related to the provision of information. High transactions costs preclude the existence of markets for externalities. The well-known result is either an excess or a scarcity of output (and investment) in activities characterized by externalities (negative or positive, respectively). As a market failure, externalities are not exogenously determined: They can be internalized in the presence of supporting institutions and adequate technologies.<sup>9</sup>

Public goods are a special case of externalities. Once produced, they cannot be appropriated; all share in their consumption, even those unwilling to spend resources to access them. The quality of air in a particular geographical area is a classic example of a public good: Its improve-

ment benefits everyone in the area, even those who are unwilling to pay for purer air. Public goods generally are financed by the state (through taxes), which either undertakes the production of public goods directly or under its close regulatory scrutiny.

Government regulatory action is also governed by equity considerations. The public interest is well served when the distribution of endowments among families (and regions) is balanced and individuals have access to a basket of goods and services that ensures their survival with at least minimum dignity. To help ensure the provision of these “merit” goods, governments sometimes provide subsidies; alternatively, governments guarantee the means for their acquisition.

In certain cases, the provision of a good is justified on both equity and efficiency grounds. Basic (and secondary) education is an example. There is an element of “merit” in this service, in the sense that most societies believe that children and youth should receive a minimal education to live with dignity. At the same time, there are strong positive externalities associated with education, because of its impact on productivity. Combined, they suggest systematic underinvestment in basic education were it left completely to the market.<sup>10</sup>

Of course, not only markets but governments as well are subject to failures and imperfections. Consequently, the introduction of any regulation as a market-supportive mechanism must pass two tests: First, the failure or imperfection must be clearly identified; second, its magnitude must be significant enough to dominate government failures. Unless both tests are met, government intervention may generate greater distortions that it attempts to correct. The crux of the regulatory question is first, to evaluate *ex ante* a government’s ability to correct market failures and imperfections and second, to devise interventions that do it effectively on an *ex post* basis. A first step is to identify what should be regulated and how.

### **The focus of regulatory intervention**

The scope of economic phenomena that merit some form of regulatory oversight is potentially broad.<sup>11</sup> The chapters in this volume focus basically on the regulatory requirements of a market economy, in areas that typically demand some measure of regulatory activism.

#### *Anticompetitive conduct in production and distribution of goods and services*

The ability of economic agents to exercise monopoly power is derived from the presence of barriers to competition. These barriers may be natural (as a function, for

example, of economies of scale), strategic (due to the presence of few agents in markets), or policy-generated (erected by anticompetitive instruments of regulation, promotion, and protection of economic activity). Beyond elimination of anticompetitive regulations, removal of those barriers requires an active competition and antitrust policy.

The object of competition policy is to protect the competitive process and encourage competitive behavior, so as to promote economic efficiency (chapter 2). Competition policy entails the design and enforcement of competition law—a code of conduct compatible with the prevailing legal, regulatory, and commercial environment of the country—and a process for adapting this code to changing circumstances (chapter 3).

Competition law should be fundamentally concerned with horizontal restraints—agreements to fix prices, reduce output, allocate customers, or bid collusively—in view of their anticompetitive effects. Regarding vertical restraints—restrictive contractual agreements between suppliers (typically manufacturers) and buyers (typically distributors)—there is a view that their efficiency impact tends to dominate eventual anticompetitive effects or at least that the commercial effects are more ambiguous. Competition authorities should be most concerned when vertical restraints lead to market foreclosure, predation, vertical squeezing, and other practices that harm the competitive process. These are more likely to occur in economies where entry into distribution might be impeded by regulatory and capital market constraints.

The tension between efficiency and anticompetitive effects is particularly pronounced in case of merger policies. Despite the difficulties of evaluating *ex ante* proposed mergers, competition authorities should carefully scrutinize the costs to the economy (and to consumers in particular) and weigh the efficiency claims. Experience shows that very few mergers have a dominant anticompetitive impact; if an efficiencies defense is offered (as most countries do), competition policy agencies usually approve the merger (no more than 1 to 2 percent of all mergers attract enforcement action). These relatively small percentages are not unexpected if there is a credible filter that acts as a deterrent to anticompetitive mergers which clearly fail the efficiencies test. Moreover, remedies such as asset divestiture and supply licensing agreements—more easily available in commercial environments where contracts are efficiently enforced—allow for focused and relatively restrained merger control (chapter 3).

Commercial conduct is as much a product of legal statutes as of their application and improvement through enforcement (chapter 3). The enforcement process is thus critical for the emergence of efficient legal standards. It implies the use of devices such as precedent (to develop enforcement standards), the private right of action (which decentralizes the enforcement of the law), and the availability of independent review by a judicial or quasi-judicial body. It also suggests the advantages of private bearing of litigation costs to promote selective use of the legal system as well as the availability of out-of-court settlement procedures to minimize those costs.

Historically, enforcement has failed less from ill-designed competition statutes than from weakness of the enforcement mechanisms—the competition agency and the judicial system. The competition policy agency should operate with considerable independence, and insulated from political and budgetary interference (chapter 2). At the same time, the possibility of regulatory capture is small, because private firms interact only infrequently with the agency (chapter 3). Despite the advantages of judicial enforcement—greater transparency, the progressive development of legal standards, and its decentralized nature—the absence of an effective judicial system may call for a limited competition law, enforced through an administrative mechanism, while legal reform is under way. In this perspective, the scope of enforcement would expand only gradually, after including initially the most overt forms of anticompetitive behavior which can be judged on a *per se* basis. Simple and transparent rules enforced judiciously, combined with a strong educational role and advocacy function by the competition policy agency, may be regarded as the hallmark of sound competition law and policy for reforming economies.

#### *Consumer-unfriendly behavior*

Consumer protection policy does not generally conflict with competition policy; on the contrary, it is arguable that they are complimentary. Consumer protection is justified in view of the incentives producers have to withhold information about the true characteristics of products and services, and the difficulties that individual consumers have in asserting their rights, in the absence of specific legislation and protection mechanisms. The object of consumer policies is both to protect consumers by improving the environment within which consumption decisions are made and to promote a change in consumer behavior, through education and information (chapters 6 and 7). Better-informed consumers and more balanced consumer-producer relations help increase the density

and efficiency of markets. Moreover, a reduction in power and information asymmetries between consumer and producer increases the likelihood of higher-quality, competitively priced products. A market populated by firms that compete fiercely for consumers who are both well aware of what they buy and protected by adequate legislation, provides an important basis for an internationally competitive industry.<sup>12</sup> A market populated by consumers protected in their basic rights helps improve distributional outcomes, in particular for countries with high levels of illiteracy and poverty.

Consumers generally encounter three types of goods in the marketplace, categorized according to the ability to ascertain quality prior to purchase. Because most consumers have sufficient information to verify the quality of search goods on an *ex ante* basis, this category is not an object of regulatory concern. In case of experience goods, consumers learn about their quality *ex post*. Despite the information asymmetry, sellers often have sufficient incentives to provide guarantees (as a signal of high quality) or information (particularly for repeat purchases), or to post a bond, making regulatory intervention unnecessary. Information about experience goods is also provided by third parties, such as product-testing and standards organizations or consumer magazines. In some cases, government subsidies may be desirable in view of the strong positive externalities.

Finally, for credence goods, quality is seldom ascertainable, providing incentives for opportunistic behavior by sellers. The protection of consumers' health and safety, in particular, requires government to take a more active stance, through a combination of liability law and direct regulation. While liability law should deter firms from placing consumers at avoidable risks, it is more appropriate as an instrument to compensate consumers for damage inflicted. Regulation, by contrast, if well designed, can be an efficient mechanism for prevention of damage. In addition, a liability law requires the use of a normally expensive legal apparatus and is commonly subject to abuse. Care should be exercised that high transactions costs related to enforcement of such a law do not outweigh its benefits. In any case, the inclusive nature of liability law is an advantage over regulatory mechanisms, the scope of which is necessarily limited.

Consumers' economic interests are also an object of protection. Governments often attempt to protect consumers from economic abuse by regulating certain aspects of adhesion contracts, and legislating against commercial fraud and unfair sales practices. In addition, governments specify minimum informational require-

ments for product advertising and labeling. Such requirements are sometimes regarded as an alternative to direct regulation (chapter 6). The provision of information has important limitations, however, particularly when a significant proportion of the population is functionally illiterate or has low levels of education. Thus, improving the informational content of goods and services should be regarded as an important part of a broader menu of regulatory, market-based, and voluntary measures.

#### *Obstacles to restructuring and exit*

Efficient restructuring faces important obstacles: weak discipline, due either to limited competition, soft budget constraint (via subsidies and transfers), or contractual imperfections (making it difficult to assert ownership rights); constrained factor mobility, due to distortionary policies (that inhibit flexibility in the allocation of factors in response to rapid market shifts), and lack of basic rules and institutions (to provide redundant workers with retraining and information on new jobs, or rapid capacity reduction in structurally depressed industries); and scarce managerial, informational, and financial resources, which limit the ability to conduct efficient restructuring. A mixture of policy reform, and an appropriate set of rules and institutions, is therefore required to inject efficiency in restructuring decisions.<sup>13</sup>

Bankruptcy law and policies are instrumental in enhancing capital mobility and facilitating timely exit. There are two somewhat different perspectives on the proper role of bankruptcy: it can be viewed either as a means to enforce debt contracts or as a mechanism of restructuring and exit (chapter 8). Bankruptcy policies are therefore concerned, on the one hand, with orderly disposal of assets (thereby internalizing a coordinating externality for creditors); on the other, they are directed toward recontracting and restructuring of debts on a basis that leaves the firm as a going concern in case of a temporary liquidity crisis. These approaches to bankruptcy find correspondence in the still fundamentally distinct (through slowly converging) models in the United States and the United Kingdom. The first takes a debtor-oriented approach to bankruptcy, whereas the second places emphasis on the “creditor’s bargain” (in the French system, the primacy is neither of the debtor or the creditor, but the court).

Bankruptcy laws, together with debt collection laws, have an additional fundamental economic role in terms of long-term development. By providing the underlying code required for enforcing contractual instruments between debtors and creditors, bankruptcy legislation

and policies are instrumental in sustaining credit markets, and financial markets more broadly. In this perspective, it is critical that credit contracts be adequately enforced and creditors’ rights protected. A system providing excessive bargaining power to debtors would exacerbate problems of credit rationing, already prevalent in developing countries; upholding the “creditor’s bargain,” by contrast, would help deepen financial intermediation (chapter 8).

The scope of bankruptcy reform generally includes outdated legislation (which in many countries fails to distinguish between an enterprise and its owners and managers); the limited processing capacity of the court system (among other judicial system constraints); and the lack of effective supervision and regulation of the banking system, which simultaneously stimulate excessive risk taking and a lax attitude toward debt collection.

#### *Inequities and inefficiencies in labor markets*

This market has very special characteristics. What is being transacted are not objects, but human labor (or labor time). Some degree of regulation therefore appears inevitable. Almost all countries make illegal certain contractual arrangements (such as debt peonage) or work conditions that are incompatible with health or human dignity. Laws on minimum wages are also broadly regarded as necessary for reasons of distributive justice, although they may lead to undesirable adjustments in quantities (through unemployment).

The regulation of labor markets is also justified on efficiency grounds. Static gains can be reaped when workers have more information on available jobs and firms more information on workers’ skills. Similarly, institutional arrangements and incentive structures that stimulate firms to invest in their employees and workers to care about the competitiveness of their productive unit, bring economic advantages from an intertemporal perspective.

At the macroeconomic level, efficiency depends on a regulatory and institutional framework that allows for flexibility in quantities and prices. Cross-country experience suggests that employment growth and wage flexibility are compatible both with the system of “employment-at-will” under individual contracts and with the Scandinavian and German models of collective labor contract. The fundamental advantage of the latter is that industry and labor need per force to internalize the macroeconomic impact of their contractual decisions, for most workers are either covered by the contract or follow its guidelines. An unrealistic rise in nominal wages would be accommodated by higher inflation levels (chapter 9).

Although there is no commonly agreed typology of regulatory interventions in labor markets, one could attempt a categorization by stating that whereas some interventions affect “the institutional process of labor exchange,” others affect labor demand or supply. The former include labor rights and standards, and conflict resolution, intermediation, and information provision mechanisms, as well as contractual arrangements for wage coordination. More direct interventions on the supply and demand schedules for labor take the form of wage and benefit regulations, job security legislation, and social insurance and assistance provisions. These are illustrative of the range of regulations normally found in labor markets.

Although excessive or inappropriate regulation is identifiable in specific sectors or industries, neither theory nor evidence suggests that multisectoral labor regulations should either be totally removed or remain in place untouched. Reform in the sense of removing certain regulations, changing others, and introducing new ones on an economywide basis, will necessarily depend on a critical assessment of the impact on efficiency and equity and of the direct and indirect costs to society.

#### *Failures in the operation of land markets*

What is the role of government in land markets? On the one hand, government should spearhead the removal of restrictions that distort the operation of land markets (see above). All constraints to land sales and rentals (including prohibitions on sharecropping agreements) should be removed and urban rent controls eliminated. At the same time, ill-conceived land development policies and legislation act as a powerful incentive for speculation at the fringe of large cities and thereby reduce agriculture output. Such policy should be reformed.

The efficient operation of land markets also requires a supportive regulatory regime. Land markets tend to fail most pronouncedly in periurban areas and at the agricultural frontier, where property rights are poorly defined or inadequately enforced. The two pillars of a new regime would be a land law—construed as “a system of predictable market rules,” including those related to expropriation of land for public projects—and a land administration that provides technical information and border adjudication services, resolves conflicts, enforces property rights, and performs valuations and assessments for tax purposes (chapter 10). The land administration should avoid establishing rules which reduce incentives for compliance with registration requirements (such as restricting registration to plots above a minimum size); it should be organized to provide technical assistance to local governments and communities.

#### *Environmental damage*

There is growing international consensus on the importance of controlling pollution and protecting the environment. Firms (and individuals) generally do not internalize the environmental costs of their activities because most common resources, such as air and water, are either free or grossly underpriced. Thus the need for regulation.

There are basically two ways of regulating activities that damage the environment (chapter 11). The first is to set a price that reflects the true scarcity value of environmental services—either directly, through taxes or license fees set in proportion to pollution levels, or indirectly, through the market. The market approach would require the allocation of scarce and potentially tradable “pollution rights” among producers. Setting up such markets may not be simple, however, mainly because the number of participants is likely to be small.

The second means of reducing environmental damage is to impose quantitative restrictions for maximum permissible pollution levels, independent of the price firms would be willing to pay to pollute. Such restrictions are particularly warranted in the case of “stock pollutants,” which have long-term health effects.

Yet, as argued in this volume, price or quantity-based regulations are inherently insufficient, unless incentive structures to stimulate compliance are also in place. Regulators act as the principals for society, establishing and enforcing performance goals for firms, which act as their agents. In this relationship, there are well-known problems of asymmetric information, as well as inconsistency between the objectives and motives of principals and agents. For one thing, regulators systematically lack information on the actual implementation of regulations; for another, the desired conduct sought from firms is not a mere outcome of imposed normative rules of behavior. Incentive-based regulations are needed therefore to make it beneficial and economically attractive for firms to reveal information and comply with regulations.

In an incentive-based regulatory regime, the regulator would ask firms to report their effluent levels. Those coming forward, in a process of self-selection, would likely be in compliance and exceeding their targets. Such desired behavior would be rewarded. Conversely, unreporting firms, most likely those that fall short of their target effluent levels, would be scrutinized. An incentive regulation system so designed would be expected to converge into a tradable permit system, with its desirable properties of holding emission levels to target on a cost-minimizing basis: with firms out of compliance and inefficient abaters would find it cost-effective to pay those that surpass their targets.

*Exercise of natural monopoly power*

The regulation of economic activity has been historically associated with the control of natural monopolies, once pervasive in the infrastructure sectors. The production characteristics of infrastructure services explains why monopolistic production appeared as a state of nature. Supply is effected through a dedicated network, made up to a large extent of fixed facilities, whose costs are for the most part irrecoverable and sunk; and where efficiency in the operation of the system and in investment decisions are critically dependent on intranetwork coordination, due to the necessary interconnectedness of the network elements. On the consumption side, most infrastructure services are private in nature; they are both excludable and rival, in the sense that delivery technology allows users to be excluded, and growing congestion implies positive marginal costs to supply additional consumers.

Although supply and demand characteristics are not uniform across different types of infrastructure services, broadly speaking market failures in infrastructure services seems to relate mostly to the production side, a result of coordinating externalities, on the one hand, and increasing returns to scale and sunk costs, on the other. While increasing returns allows competition only “among few”, the presence of sunk costs may constitute binding entry barriers, muting potential competition. Both the presence of externalities and limited competition may call for regulatory oversight. On the demand side, the merit good nature (or the essentiality of supply) of many infrastructure services (such as water and electricity) or even the captivity of users, may justify a measure of regulatory intervention.

Technological progress and a better understanding of the functioning of markets, particularly for the supply of infrastructure services, have put in question traditional models of natural monopoly regulation. Although complete deregulation may still not be possible for most services, it is certainly an alternative for competitive segments, so long as “vertical separation” itself is feasible. Even if separation is not advisable on economic grounds, the scope of regulatory reform in infrastructure remains broad, as underscored by the discussion of chapters 12–15.

*Structural issues.* The fundamental structural question that needs to be addressed in reforming the infrastructure services concerns the vertical separation of the industry: the circumstances that justify divestiture and the optimal design of a restructured system. Separating elements previously integrated in an interconnected whole require the definition of access rules to bottleneck facili-

ties—which stand between provider and end consumer—in order to minimize the probability of foreclosure by the owner or controller of the facility. The additional issue that requires response is therefore how to ensure network access to nonintegrated competitors—be it in telecommunications, railroad transportation or electricity generation, among other infrastructure industries - so that the most efficient providers of services can effectively reach end users. Alternatively, if integrated producers are regionally based, the question is how to guarantee that previously “closed” territories are open to competition, both from other “regional” and from nonintegrated firms.

The definition of access rules would be a precondition for the competitive workings of the system, either by enabling integrated producers to compete in nontraditional markets, outside their territories, or by allowing deverticalized producers access to consumers. Such access rules, by determining to a large extent the nature and intensity of entry, end up shaping the structure of industry and the competitive behavior of firms.

*Pricing.* Public interest regulation in infrastructure is made more difficult from two basic trade-offs facing policymakers. First, between structural flexibility afforded by divestiture and economies of vertical integration; and second, between pricing efficiency and financing requirements of an expanding system. In the pricing of services, the critical issue is how to reconcile the requirements of economic efficiency encapsulated by marginal cost pricing, with a tariff level and structure which guarantees adequate revenues for investment (both incumbents’ and entrants’) and long-term growth of the industry. The other major pricing question concerns network access charges, both under an integrated industry or otherwise, as there is no guarantee that bottleneck facility operators will not exercise market power, and charge above the sum of incremental and opportunity costs they face.

*The regulatory regimes.* In view of the fact that much of the assets in infrastructure industries are durable, long-lived, and immovable, it is absolutely essential that a predictable and stable regulatory system be in place. This system must per force balance the claims of different stakeholders. Investors’ lack of security and confidence, would preclude successfully tapping private sources of finance; at the very least, governments would need to provide guarantees to offset perceived risk and uncertainty.

The need to take into account the claims of different constituencies in the regulatory process is, in fact, central to the stability of a regulatory regime. In this sense, regulation is inherently a political operation (chapter 15).

Technical solutions will be an input, albeit a critically important one, into a process that has an essential political dimension, and which will ultimately determine the ability of regulatory bodies to introduce and sustain economically efficient structural and pricing choices.

That is not to say the complex structural and pricing problems of regulating the infrastructure industries have been solved. To the contrary, the discussion of chapters 12–15 underscores the need not only to undertake what in many cases are bold structural reforms. But moreover to ensure that regulatory practice is consistent with contemporary economic analysis and experience of natural monopoly regulation, for those segments of the industry that must remain under regulatory oversight.

*Telecommunications.* The analysis of telecommunications is illustrative of the range and complexity of regulatory reform issues facing the infrastructure industries (chapter 12). In telecommunications, vertical separation is increasingly attractive as interchange (and value-added) services become structurally competitive. Moreover, despite potentially significant joint economies in the operation of local and interchange services, separation of local and long-distance services might be the most interesting structural option in view of the need to minimize the probability that the local monopoly carrier prevents entry of providers of services which rely on the basic network to access their client base.

To ensure freedom of entry, access to essential or bottleneck services is required; regulating such access in face of a vertically integrated carrier may stretch the regulatory capabilities of most countries. Thus vertical separation may be needed, even in the presence of scope economies. Access prices must at least cover the long-run incremental cost of the use of the network by the entrant; to this should be added, in case of a vertically integrated network, the opportunity cost incurred by the incumbent.

While the interchange market is increasingly competitive, with remaining natural monopoly elements being undermined by rapid technological change, increasing returns to scale in network design and management, and the presence of network or subscriber externalities, may still justify continuing regulation of local services. This despite the fact that mobile (wireless) communications and convergence of voice data in video technologies are dissipating the natural monopoly element of local markets. Nevertheless, it appears that over the short and medium term, regulatory oversight of local exchange services may still be needed.

For the local services, regulation could be based on “social contracts” that delimit a set of core activities that

remain regulated and the constraints that must be met in providing associated services (chapter 12). In exchange, the regulator would deregulate other competitive or non-essential services. Price caps would be an example of social contract regulation, insofar as the so-called aggregate “index-ceilings”—adjusted periodically to the rate of inflation minus a pre-specified productivity gain—are placed on a group of services (“baskets”). Service prices can depart from the index as long as the basket as a whole does not.

A pricing structure consistent with the public interest would be composed of two parts: a subscriber-specific fixed access charge to recover the marginal non-traffic sensitive costs of connection to the network, varying according to location, income and other subscriber characteristics; and an additional charge directly proportional to the subscriber’s usage of the network. In opposition to a flat rate for local services, a measured service (not only for interchange but also local calls), by differentiating between access and local usage, would be an effective way of taking into account network externality. If necessary, it would allow for the underpricing of local access charges, thereby explicitly recognizing the externality provided by the additional subscriber, most significant at least for low levels of household telephone penetration.

It is well known that pure marginal cost pricing is not consistent with revenue requirements of firms operating under increasing returns, as in telecommunications. Some loss in efficiency is therefore inevitable. Ramsey prices would provide the basis for establishing a tariff structure and a pricing regime that minimizes welfare losses (chapter 12). Each service would be priced at a mark-up over marginal cost inversely related to the elasticity of demand for that service, and in proportion to its value to the consumer, thereby concentrating the departures from marginal cost pricing in the steepest or most inelastic part of the demand curve. Prices stand above marginal costs (otherwise investors would be drawn away from the industry) but below the monopoly outcome, thus rewarding investors without gouging consumers, which are charged according to the “value of service” consumed.

*Railroads.* The key regulatory issues in the railroad industry—historically one the most regulated sectors of the economy—is not dissimilar from telecommunications: how to arrive at an economically efficient pricing regime, responsive both to the public interest and consistent with the railways’ financial requirements; and what would be the optimal structural configuration of the industry.

Despite the fact that the industry is characterized by “indivisibilities, pervasive economies of scale and scope, and small numbers competition,” with extensive capital sums having to “be sunk in way and structures and in a variety of ancillary facilities in order to create new rail lines,” there are strong competitive pressures on railroad services from intramodal, intermodal, geographic and product competition (chapter 13). As a result, it is posited that rates for traffic subject to competition should be deregulated and the railroad offered freedom in pricing.

For the regulated segments of the industry, rates must be chosen which make for adequate revenues (in the sense of providing a return on net investment equal to the current cost of capital, that is, the return available on alternative investments with similar risk characteristics). As in telecommunications, Ramsey prices would provide the basis for an economically sound rate structure for the industry, consistent with the public interest.

Although regulators have difficulty in accepting the notion that price discrimination is ultimately beneficial to all consumers, experience suggests that maintaining zero economic profits under uniform pricing in face of very different demand elasticities will make the industry less competitive and lead ultimately to its demise as core clients are driven away.

Despite the attractive characteristics of Ramsey prices, it is argued that it would be extremely cumbersome, if not outright impossible, to calculate marginal costs and demand elasticities for every railroad service or movement. Further, Ramsey prices appear unable to protect captive shippers from the exercise of market power by the railroad. Finally, the services volume offered by the railroad under Ramsey prices is less than would have been offered under marginal cost pricing. Thus, “it may be feasible” for railroad and shippers to enter into individualized voluntary contracts that will leave both parties better off, with final prices following between Ramsey and marginal cost values.

A critical question is how to arrive at a rate ceiling which protects captive shippers. Contrary to ceilings derived from fully distributed costs, with the averse impact on the railroads operations and costing decisions, the stand-alone cost (SAC) constraint (representing the minimum costs of a hypothetical alternative to the service provided by the incumbent railroad, and thus the maximum it could levy shippers without substantial traffic diversion) provides the basis for a set of economically rational ceilings in the absence of competition. The underlying rationality of this rate-setting criteria is that the railroad has no incentive to engage in cost-padding

and other practices to increase rates, insofar as such cost increases are unrelated to the SAC. Furthermore, by applying only to those services which the railroad has monopoly power, the SAC does not “interfere with the railroad’s incentives to aggressively pursue additional traffic...” (chapter 13). Finally, stand-alone costs can be calculated on the basis of engineering and other data, be periodically updated, and be structured in a way that lends itself to price-capping incentive regulation.

The basic structural alternatives to the monolithic railway are competitive access and vertical separation. In case of competitive access, competing railways would operate in a given market over a particular fixed facility, thereby having mutual rights of access to trackage. An alternative option would be vertical separation of the industry, taking advantage of the possibility of segregating the competitive and monopoly elements in supply. In this case, track, roadbed and other non-exclusive fixed facilities are the property of one owner, whereas the operation and marketing of the services are undertaken by competing firms. Separation might, however, be effected at the marketing stage, with competing entities (“retailers”) marketing the railways services to shippers.

Which is the dominant or most attractive alternative is a nontrivial question: while the basic obstacle to competitive access are the incentives of bottleneck holders, inimical to efficiency and competition, it is argued that separation tends to create “serious coordination problems, loss of economies of scope, and otherwise unnecessary transaction costs” (chapter 13). Separation would be particularly attractive “where a dense and extensive rail network permits many operators to function” (permitting active or potential competition among rail operators or retailers), and “a mature and well developed set of fixed facilities” is available, so that no new major infrastructure investments are required, thereby avoiding major incentive problems (among the infrastructure provider, operators and shippers) over the risks and rewards associated with such investment. Otherwise, competitive access may be the preferred choice, as long as pricing and terms of access lead to adequate compensation for the “landlord” railway, and efficient operations of the “tenant.”

*Electricity industry.* Similar complex structural and pricing questions characterize the electricity supply industry as governments face the option of separating its potentially competitive parts from those which have strong natural monopoly characteristics (transmission and distribution). This suggests separating transmission from generation, thereby creating a market for bulk elec-

tricity, “the key reform enabling competitive pressure to be put on generation” (chapter 14). The importance of guaranteeing access to the transmission system by new generators, and the dangers of introducing rigidities into the system by having incumbent monopoly generators own the transmission system through a “club” arrangement, are most incisive lessons from recent experience, due to difficulties for the regulator to intervene and change contracts among private parties. Thus, it is argued that short of high-quality regulation, “continued public ownership of the transmission system” may be the best option.

Pricing the grid services on an economic basis, by striking an adequate balance between efficiency and financing requirements, is nontrivial. What drives away potential investors is not only the level of access charges or even the difficulty of gaining access to the grid, but the risks associated with investments which are not only fixed but mostly sunk, and returns to which long term. In this perspective, despite the strong efficiency properties of price capping and similar pricing regimes, they become of questionable application to countries which need to attract large volume of investments in the sector, in view of the risks associated with the frequent regulatory reviews which comes with price capping.

The other extreme is a pricing structure that supports a predetermined rate of return. Though ensuring the availability of low cost finance to underwrite investment, it is well known to promote inefficient investment choices. Moreover, rate-of-return regulation increasingly attracts prudential regulatory reviews, which injects instability into the earnings profile of the supplier and increases the perceived risk of the investment. In either pricing regime is it is necessary that regulatory reviews be designed so as to maintain investor confidence and continued support for the system of regulation.

There are, of course, alternative institutional solutions to the risk-related regulatory problem. The U.S. response is a set of independent regulatory agencies within the frame of “constitutional guarantees to a fair rate of return...upheld by an independent legal system that protects property rights” (chapter 14). This stands in contrast with the European Continental experience of “a regulatory compact in which the costs to the government of intervening to impose tighter regulations outweigh the benefits in terms of lower prices and short run voter support”, approaching thereby a system of self regulation. Strict self regulation is, however, quite unusual. As the above discussion suggests, creating a system of effective regulatory oversight in line with the public interest is a

fundamental requirement for the sustainability of regulatory reform in the infrastructure industries.

### **Principles of regulation**

This volume suggests that there is a set of economic activities that need to be regulated insofar as they are the outcome of markets characterized by imperfections, failures, and gaps. Although there are no universal rules, certain principles help ensure that in attempting to correct one distortion, others (perhaps worse) are not introduced.

First, regulatory controls that make use of price vectors and market mechanisms are generally more effective. There are substantial gains from engineering a market or emulating one, thereby providing a decentralized resource-allocating mechanism for activities subject to regulation.

The problem is that markets for externalities, or for economic activities populated by few actors—circumstances that justify the imposition of regulatory limits—tend to be either nonexistent or extremely thin. To establish such markets, or provide them with greater density, property rights must be defined and efficient exchange mechanisms created. These, however, are only necessary but not sufficient conditions for the functioning of markets. When markets are absent and cannot be engineered, it is desirable for governments to use taxes, subsidies, fines, and other instruments that emulate prices, so that firms may respond through an unconstrained process of cost-minimization.

Second, there are circumstances in which price vectors—either market-generated or exogenously determined by government—are insufficient and must be complemented by quantitative restrictions. For example, such restrictions would be required in the case of a monopoly in a noncontestable market that does not respond to signaling from a regulatory agency, or when it is determined that the maximum tolerable level of pollution for a given pollutant is close to zero.

Third, in setting up regulatory mechanisms, the complexity of the task as well as the availability of technical, managerial, and administrative resources should be taken into account. Regulatory capacity is not built overnight. It is a process involving adequate technical training, the accumulation of tacit knowledge through trial and error, and a progressive narrowing of the information gap between the regulator and the regulated entity. In addition to specific technical knowledge, effective regulation requires a great amount of information, some of which is to be disclosed by the producer. A priori the regulator has no way to know if the information supplied is correct, in

particular because it is largely observable only by the producer. Thus the importance of mechanisms which allow regulators to learn from experience and of incentive structures which make it in the interest of the regulated firm to reveal information.

Fourth, the success of regulatory policies depends on a competent judicial system and a set of technical and administrative mechanisms for regulatory analysis and the enforcement of regulatory decisions. The perceived weakness of government agencies should not deter governments from undertaking effective market-supportive regulation. The process of structuring such agencies and staffing them with technically and administratively competent people is slow, but competence on the part of government and maturity from markets are fundamentally based on learning by doing. It is preferable for the government to mark its presence in the regulatory arena modestly and build it up through incremental steps, rather than not at all.

Fifth, given the tentative and incremental nature of a young regulatory regime, universal rules, easily applicable, are preferable. Administrative decisions must be transparent and subject to judicial review. Discretionary decisionmaking, particularly at an early stage of regulatory learning, may lead to gross mistakes and a societal backlash. Both the process of deregulation and that of establishing a new regulatory framework must be carried out in an open manner. Economic agents and society at large must understand the objectives, methods, and hoped-for results of a government's attempts to build a regulatory framework to support an open, competitive, and socially just economy.

Over time, with enough experimentation and deft adaptation, rules can grow in complexity and their implementation can accommodate a degree of discretion. Nevertheless, it should be kept in mind the importance of a stable regulatory regime; the need for its adaptation to changing circumstances has to be cautiously compromised so as to minimize regulatory uncertainty.

Finally, effective regulation is not just predicated on technical information—capturing capabilities (and the experience) of the regulator. It is also dependent on the involvement of civil society in the regulatory process. Increasingly, in all aspects of regulation, the sustainability of the regulatory regime depends on the degree of inclusiveness so as to provide credibility and thereby reduce uncertainty of regulatory decisions. By ensuring broad participation, regulatory mechanisms should not be able to deliver technically efficient and economically sound decisions, but to effectively resolve legitimate social conflicts, consistent with the public interest.

### Concluding remarks

The new developmental state will continue to exercise a decisive influence on the process of industrialization and long-term growth. However, its role will be distinct from the one the state played in the postwar years. A diminished involvement in directly productive activities will contrast with a greater emphasis on social infrastructure; on maintaining a stable macroeconomic environment; and on establishing common development objectives with the private sector, by which government policies and programs will be carried out. The state will also be required to undertake bold reforms in the regulatory arena: to remove the maze of regulatory controls, promotional mechanisms, and protective regimes that constituted the instruments of state action under the old paradigm and introduce policies, rules, and institutions that support efficient markets and improve their distributional results.

In particular, the focus of state attention will be the more complex problems of skills acquisition, rules crafting, and institution building, necessary to allow markets to work and to compensate for failures and imperfections. In this sense, the new developmental state will again be concerned with the basic problems of economic development: establishing property rights, enforcing contracts, removing barriers to trade, and building up the regulatory mechanisms to address problems of market failures and anticompetitive practices.

Yet even in the more developed economies, regulatory mechanisms to address problems of market failure and the exercise of market power were only introduced relatively late and not at once. Economies that did not have the advantage of a long period of market buildup are now facing the substantial task of clearing the way for markets to function and supporting their growth, while curbing the exercise of market power. Thus it will hardly be a trivial task for the new developmental state to bring markets into existence, remove restrictions on transactions, and improve their contractual context—at the same time compensating for the incomplete and imperfect nature of markets.

### Notes

1. See, for example, Robert Nozick, *Anarchy, State and Utopia*, Basic Books, New York, 1974. See also Friedrich Hayek, *The Constitution of Liberty*, University of Chicago Press, Chicago, 1960 (particularly chapter 15).
2. An econometric investigation on the determinants of real private investment presents strong evidence on the positive impact of public investments. A one percent increase in the ratio of public invest-

ment to GDP expands the private investment–GDP ratio by 0.257 percentage point, far outstripping the impact of other variables. See Luis Servén, and Andrés Solimano, “Economic Adjustment and Investment Performance in Developing Countries: the Experience of the 1980s,” in Vittorio Corbo, Stanley Fisher, and Steven Webb, eds., *Adjustment Lending Revisited: Policies to Restore Growth*, The World Bank, Washington, D.C., 1992, Table 7.9.

3. The perception of uncertainty by investors (even if investors are risk neutral and their risks diversifiable) is a major obstacle for long-term economic growth. It is particularly critical in the context of adjustment programs, where investment response has proven to be often unexpectedly slow and weak due to the incomplete credibility of policy reforms. On this point see Luis Servén and Andrés Solimano, “Private Investment and Macroeconomic Adjustment: A Survey,” *The World Bank Economic Observer*, Vol. 7, No.1 (January 1992).

4. Not infrequently, the agents favored by certain rules argue that the decision to invest in a specific area was taken in response to implicit government guarantees and directly under its orientation. How should the government therefore remove its support from an activity that was once a priority to the point that the State “commanded” private investment and regulated its allocation? Government officials, on the other hand, responsible for the design and implementation of the old regime, saw as their responsibility assuring the survival of firms that undertook what were once priority projects.

5. For a more detailed discussion, see Claudio Frischtak et al., “Competition Policies for Industrializing Countries,” World Bank Policy and Research Series No. 7, Washington, D.C., 1989.

6. Justice, in the Rawlsian sense, would require that all individuals have access to a set of income-generating endowments allowing for life with a minimum of dignity. See John Rawls, *A Theory of Justice*, Harvard University Press, Cambridge, Mass., 1971.

7. As in the cases of non-regulated natural monopolies, cartelized (cooperative) oligopolies, or when firms’ strategic variable are quantities (Cournot oligopolies). In case of Bertrand oligopolies (with firms competing aggressively in prices), the probability that outcomes will be competitive is considerably greater. Finally, in case of bilateral monopolies or oligopolies, the outcome is undefined, as bargaining substitute markets.

8. See Kenneth J. Arrow, “Political and Economic Evaluation of Social Effects and Externalities,” in J. Margolis, *The Analysis of Public Output*, National Bureau of Economic Research, New York, 1970.

9. Take the case of pollution. By specifying a maximum permissible level of pollution, distributing “pollution permits” and letting markets flourish by allowing such permits to be traded, is possible to simultaneously reduce the levels of pollution to a pre-established maximum (the aggregate of all permits), while stimulating firms to decrease their contribution to pollution. The distribution of (lim-

ited) pollution “rights,” and the ability to market such rights, ensure that firms take into consideration the negative externality associated with pollution, without, however, being deflected from their optimizing objectives.

10. Education does not have public good characteristics; individuals can easily be excluded from its consumption. Not all activities characterized by externalities are of public good nature. On the other hand, the subprovision of education in the absence of government regulation suggests a market failure, which appears unable to signal to both producers and consumers of the service the true (social) price and profitability of the activity. Note that such failure is not directly related to the question of equity. Even under a state of distributive justice, the supply and demand for education would generate prices and quantities that fail to maximize welfare. Inequity obviously makes the situation worse. Individuals with income below the minimum acceptable fail to consume enough education also from a merit good perspective, in addition to the adverse impact on economic efficiency.

The underconsumption and underproduction of education in the absence of government regulation result from both its positive externalities and its merit good aspect. Note finally that from these considerations it cannot be immediately inferred what is the most adequate mechanism for intervention, except that being a merit good, the service should be consumed by all children. If the provision of education will be undertaken directly by the government or just supported by scholarships in a regime of free choice, will depend on the market response to the non-regulated demand for education.

11. This volume has not dealt with the regulatory requirements of financial market stability and efficiency. It should be clear, however, that printing money, and controlling its volume, are activities classically regulated by the government. More than a “veil,” money becomes the anchor that allows the real economy to function without turbulence (except for the noise inherent in changes in technologies and markets). An essential function of government is thus to guarantee monetary stability. Financial markets, through which money circulates and is “created” (via credit), has a central role in this process. The perception that market institutions are strong is as important to the stability of the economy as the credibility of government agencies that regulate financial flows. In their absence, demonetization occurs, and the real sector collapses. Thus the enormous positive externality derived from solid financial institutions and stable financial markets; and the need for prudential regulation, and the obligation that institutions disclose information on transactions and portfolios. There is, on the other hand, a significant anticompetitive bias in financial sector regulations. They deter entry (through rules such as minimum capital requirements and limits on permissible activities); control exit (by injecting resources in institutions with contaminated portfolios); systematically intervene in sectoral prices (such as interbank rates); and, more generally, regulate competition among financial institutions, in order to

minimize market turbulence. In this sense, there is a nearly inevitable "trade-off" between micro and macroeconomic efficiency, the first stimulated by competition in the financial system; the second supported by its stability.

12. See Michael Porter, *The Competitive Advantage of Nations*, London, Macmillan, 1990.

13. For a detailed discussion of these rules and institutions see, for example, Izak Atiyas, Mark Dutz and Claudio Frischtak, "Fundamental Issues and Policy Approaches in Industrial Restructuring," Industry Series Paper No. 56, Industry and Energy Department Working Paper, The World Bank, April 1992.

### References

Arrow, Kenneth J. 1970. "Political and Economic Evaluation of Social Effects and Externalities." In J. Margolis, ed., *The Analysis of Public Output*. New York: National Bureau of Economic Research.

Atiyas, Izak, Mark Dutz, and Claudio Frischtak. 1992. "Fundamental Issues and Policy Approaches in Industrial Restructuring." Industry Series Paper 56. Industry and Energy

Department Working Paper, World Bank, Washington, D.C.

Frischtak, Claudio, and others. 1989. "Competition Policies for Industrializing Countries." World Bank Policy and Research Series 7. World Bank, Washington, D.C.

Hayek, Friedrich. 1960. *The Constitution of Liberty*. Chicago: University of Chicago Press.

Nozick, Robert. 1974. *Anarchy, State, and Utopia*. New York: Basic Books.

Porter, Michael. 1990. *The Competitive Advantage of Nations*. London: Macmillan.

Rawls, John. 1971. *A Theory of Justice*. Cambridge, Mass.: Harvard University Press.

Serven, Luis, Andrés Solimano. 1992a. "Economic Adjustment and Investment Performance in Developing Countries: The Experience of the 1980s." In Vittorio Corbo, Stanley Fisher, and Steven Webb, eds., *Adjustment Lending Revisited: Policies to Restore Growth*. Washington, D.C.: World Bank.

———. 1992b. "Private Investment and Macroeconomic Adjustment: A Survey." *The World Bank Research Observer* 7(1):95–114.

# The instruments of competition policy and their relevance for economic development

R. Shyam Khemani and Mark A. Dutz

The objective of this chapter is to discuss the conceptual underpinnings of competition policy, its instruments and limitations.<sup>1</sup> By “competition policy,” we mean those government measures that directly affect the behavior of enterprises and the structure of industry. An appropriate competition policy, as we define it, includes both (1) policies that enhance competition in local and national markets (such as liberalized trade policy, relaxed foreign investment and ownership requirements, and economic deregulation), and (2) competition law (also referred to as antitrust or antimonopoly law) designed to prevent anticompetitive business practices by firms and unnecessary government intervention in the marketplace.

While most developing and emerging market economies<sup>2</sup> have put in place numerous policy measures that fall under the rubric of competition policy, few subject these policies to a clearly articulated, coherent set of objectives. This reflects a diversity in views on the proper objectives of competition policy. Should it seek only to foster economic efficiency or should it take up broad public interest issues (such as balanced regional development and employment)? Should it promote only consumer welfare or total economic welfare? And is the active promotion of competition by government conducive to industrial growth and international competitiveness?

A critical question is whether countries need a competition law to complete their national economic policy framework. Is liberalization of international trade sufficient to promote competition? Similarly, is competition law essential, desirable, or only worth considering after other, more urgent policy measures have been introduced? Even if competition law is considered desirable in the abstract, the question remains whether the probability of improper enforcement, misuse of bureaucratic power, or regulatory capture is so high in developing and emerging market economies that the expected costs of such legislation outweigh the possible benefits.

This chapter argues that a well-designed competition law should be accorded a central place in the national economic policy framework. A good competition law is essential to economic growth and development—especially in newly industrializing and emerging market economies. Policymakers in these countries have too often relied on state enterprises, regulatory controls, promotional instruments, and trade restrictions to spur industrial development. Measures such as entry and capacity licensing, subsidies, targeted investments, and trade barriers have led to concentration of industry and sheltered domestic markets. In this environment, the profitability of some industries has been three to five times that in open markets (Frischtak, Hadjimichael, and Zachav 1989). Not surprisingly, monopolies and large stakeholders tend to wield significant influence in such economies.

In contrast, competition law, by preventing artificial barriers to entry and facilitating market access, complements and buttresses other policies that promote competition. Appropriate competition policy requires more than the enforcement of competition law. It entails conditioning the business environment in such a way that competition constrains the exercise of market power by firms,<sup>3</sup> improves static and dynamic efficiencies, and promotes reallocation of resources from lower- to higher-value uses. This requires no less than a paradigm shift in how competition authorities function. That is, competition law and policy must be applied not only reactively but proactively against anticompetitive situations—whether these emanate from business practices or from government policies.<sup>4</sup> In short, a competition advocacy function is as important as the law enforcement function of competition agencies.<sup>5</sup>

The first section of this chapter discusses the objectives of competition policy. Differing views on proper objectives reflect different schools of thought in industrial organization economics. While a commonly shared

objective is to promote the efficient use of resources, tensions exist between schools on the choice and the application of different policy instruments. The section also examines the case for an active, state-directed industrial policy to spur rapid industrialization and promote international competitiveness. Here, the available evidence is mixed at best. But cross-country experience suggests that vigorous domestic and international rivalry between firms is an important ingredient in successful industrial performance.

The second section discusses the principal structural and behavioral provisions of competition law. The debates between the different schools of thought surrounding these provisions suggest priorities for the implementation of competition law in developing and emerging market economies. The third section examines the argument that liberalized international trade is sufficient to promote competition—especially for small economies—and that a competition law is not needed. This view is rejected because various factors can segment domestic and international markets and permit anticompetitive business practices

The fourth section discusses issues arising from increasing internationalization of competition law and policy. The globalization of markets and increasing number of transborder corporate mergers and acquisitions, joint ventures, and strategic alliances suggest that conflicts can arise in how competition policy is applied in different jurisdictions. Multilateral approaches toward resolving international competition issues are also examined. The last section concludes on the role and importance of competition policy in the structural adjustment process in developing and emerging market economies. Throughout the paper, we identify policy areas where further research is warranted.<sup>6</sup>

This chapter suggests that the principal objective of competition policy should be to maintain and encourage market competition in order to promote economic efficiency. Interfirm rivalry with regard to price, quality, and service form the crux of competition. Competition leads to improved productivity and long-run commercial success. But a distinction must be drawn between protecting the competitive process and protecting competitors. Focusing on economic efficiency is more likely to foster consistency in economic decisionmaking than the pursuit of multiple objectives under the rubric of “public interest.”

While industrial market structure is an important determinant of business conduct and economic performance, policy emphasis should be on firm conduct, not firm size. Large firm size does not automatically confer

market power. Barriers to entry play a key role in enabling large firms to *exercise* that power. By working to eliminate or reduce barriers to entry, policymakers can dramatically enhance competition. One of the focal points of competition law and policy should therefore be the removal of barriers to new competition. A deliberate policy of protecting domestic markets and fostering large firms does not guarantee rapid industrial growth and international competitiveness and may actually hinder them. Firms that do not compete in their home markets are unlikely to be able to compete in international markets. These messages apply to industrial as well as developing and emerging market economies.

### **Objectives of competition policy**

Although Canada and the United States enacted competition legislation in 1889 and in 1890, respectively, most other industrial countries enacted such legislation after World War II.<sup>7</sup> Many countries in fact have had longstanding provisions against monopolistic exploitation in their constitution or other laws that remained largely unexplored. More recently competition law has gained popularity in developing and emerging market economies (table 2.1). Many of these inherited a tradition of extensive government ownership and regulation, as well as stiff barriers to trade. Frequently, reciprocal barriers between countries have segmented domestic and world markets and led to monopolies, oligopolies, and inefficient production. Throughout, concern about low or falling productivity and competitiveness, coupled with fiscal deficits and the poor track record of government intervention, has prompted moves toward more market-oriented policies. Policymakers have included competition policy in their reforms to guard against the exercise of market power by newly privatized or deregulated firms.

Various objectives have been ascribed to competition law at different times in different countries. A survey of legislation and policy statements reveals that the most common (the principal or “core”) objectives are maintaining free competition and protecting or promoting effective competition.<sup>8</sup> Policymakers have viewed these objectives as synonymous with preventing unreasonable private restraints on competition. In many countries, competition policy is also aimed at improving market access by reducing barriers to entry. In the European Union, free competition has been equated with promoting economic integration. As a result, rules on competition have been used as policy instruments to foster deregulation, privatization, trade liberalization, and other measures that can enhance the mobility of resources.

TABLE 2.1  
**Adoption of competition law in developing and emerging market economies: selected examples**

Country	Year law enacted
<b>Africa</b>	
Côte d'Ivoire	1993
Ghana	Legislation initiated
Kenya	1988
Morocco	Legislation initiated
Senegal	Legislation initiated
South Africa	1979
Zambia	Legislation initiated
Zimbabwe	Legislation initiated
<b>Asia</b>	
India	1969
Korea, Republic of	1980
Pakistan	1970
Philippines	Legislation initiated
Sri Lanka	1987
Thailand	1979
<b>Latin America and the Caribbean</b>	
Argentina	1919, 1946, 1980 (revisions under way)
Brazil	1962, 1994
Chile	1959, 1973
Colombia	1959, 1992
Ecuador	Legislation initiated
Jamaica	1993
Mexico	1993
Venezuela	1991
<b>Formerly socialist countries</b>	
Belarus	1992
Czech and Slovak Republics	1991
Poland	1990
Russia	1991

Why should competition be preserved and fostered? Different schools of thought in industrial organization economics provide different answers. The most influential and distinctive of these are as follows:<sup>9</sup>

- *The structuralist school* emphasizes the interaction between market structure and collusive and exclusionary business practices by firms that enable them to exercise market power and persistently earn excess profits. The structuralist position is rooted in the “high market concentration–low competition” paradigm, which holds that firms operating in oligopolistic industries with large market share are more likely to coordinate their pricing and output or to unilaterally engage in anticompetitive behavior.<sup>10</sup> The structuralists view allocative efficiency as important but also perceive deconcentration and improved income distribution as valid objectives of competition policy.

- *The Chicago school* has evolved largely as a reaction to the structuralist viewpoints. Economists associated with this school argue that collusion is difficult to practice profitably in all but the most highly concentrated industries and is therefore not a serious problem (see Stigler 1968). Where competition is restricted, collusion arises primarily because of barriers to entry created by government.<sup>11</sup> These economists advocate one unequivocal goal for competition policy—the pursuit of economic efficiency. In most instances, they view exclusionary practices of firms as motivated by pursuit of economic efficiency. Failure to consider economic efficiency, they say, distorts the basic intent of competition policy. As a result, they favor a minimalist approach toward the administration of competition policy. Competition law, in particular, should be restricted to preventing of collusion, especially price fixing agreements (see Bork 1978 and Posner 1969).

- *The statist or industrial policy school* argues that the competitive market system is an outdated economic institution embodying misguided values that have contributed to the deterioration of economic performance.<sup>12</sup> Markets often fail to guide investments to industries that would generate high growth—and governments must therefore “lead the market” by identifying strategically important industries. Closer integration of business and government is needed to ensure that firms are large enough to compete.<sup>13</sup> Accordingly, this school believes that competition policy should be set aside since it hinders domestic firms’ ability to compete against foreign firms and to graduate from national to international markets.<sup>14</sup>

The arguments advanced by these schools have provided rationales for countries to adopt different approaches to competition policy. Some countries (such as Canada, Colombia, Mexico, New Zealand, and the United States) have emphasized economic efficiency to varying degrees, others (such as Australia, France, India, the United Kingdom, and some emerging market economies) the impact of competition on the broad public interest as well. This second group of countries has identified such issues as employment, diffusion of economic power, community welfare, and regional development as matters that competition authorities must consider. But the pursuit of these multiple goals suggests that competition policy may become subject to political compromise. It is not easy to balance multiple and often conflicting objectives. Instead, it is preferable to pursue noneconomic objectives through separate government policies and instruments. This approach increases the likelihood of developing a consistent competition policy and makes administration more transparent and accountable.

Even if there is general agreement that competition policy should primarily strive to promote economic efficiency, the question arises whether this means static or dynamic efficiency or both. Moreover, should the goal be to maximize consumer welfare or total (consumer plus producer) welfare?

Consumer welfare increases when firms expand output and consumers pay lower prices. Producer welfare increases when firms can earn higher profits by charging higher prices. A total welfare approach entails evaluating the actual and potential net gains to both consumers and producers. For example, reduced output may increase prices, firm profits, and producer welfare while lowering consumer welfare. However, total economic welfare may increase if the gains in producer welfare are greater than the reduction in consumer welfare plus associated dead weight losses, and if profits are reinvested to gain dynamic efficiencies in terms of better production technology or improved product quality and service, which lead to increased output and consumption.<sup>15</sup> A stringent consumer welfare maximization approach to competition policy may foster static efficiencies at the cost of dynamic efficiencies. To promote the latter, a more permissive policy toward large firms, mergers and acquisitions, and interfirm cooperative arrangements in research and development may be necessary. This might shift a larger proportion of welfare gains to producers, constituting an incentive for industries to restructure.

Some economists, mostly associated with the statist or industrial policy school, have argued that an overly activist competition policy forces companies to think short-term—to the detriment of their long-term competitiveness. Moreover, policymakers in industrial as well as developing and emerging market economies have argued that the pressures of global competition require a domestic business environment that fosters the development of a few large enterprises (including financial-industrial conglomerates) that can act as “national champions” and “engines of growth.” They point to the high post-war growth rates of Germany, Japan, and the East Asian economies relative to those of the United Kingdom and the United States, which supposedly underscores the importance of government’s active role in protecting and promoting selected enterprises and industries.

*Firm size and economic performance: What does the empirical evidence suggest?*

Many studies have been conducted on the relationship between firm size and such measures of economic performance as profits, productivity, exports, and research

and development expenditure. Although the voluminous literature is based almost exclusively on industrial country data, it permits some broad conclusions (see Scherer and Ross 1990).

Analysis of a large sample of industries in a number of countries has found a positive association between the size of leading firms or industry concentration, on the one hand, and profits on the other.<sup>16</sup> But there are two opposing views on the interpretation of these results and on the emergence of industrial concentration.<sup>17</sup> One, along structuralist lines, holds that the positive relationship between industry concentration and profits is indicative of monopolistic pricing. According to this interpretation, high levels of concentration result from anticompetitive business practices that lead to resource misallocation. The opposing view, advanced by the Chicago school, holds that the positive relationship reflects superior competitive performance by leading firms. According to this view, in the absence of government-erected barriers to entry, high levels of concentration and profits can be maintained only if the leading firms constantly strive to be innovative and efficient.

No unequivocal statement can be made about the implications of firm size and industry concentration solely on the basis of their relationship with profits. Further research on the underlying factors is needed, particularly on barriers to entry and new competition.<sup>18</sup> A large number of supplementary studies have been conducted on the determinants and effects of industrial concentration. Broadly, they suggest the following:

- Economies of scale are an important determinant of industrial concentration. But most efficiency gains are achieved at a relatively small size of operations, and that is increasingly the case with technological change and the introduction of modern production methods such as CAD-CAM and flexible manufacturing.
- Cross-country studies that have included Japan and the Republic of Korea have found high industrial concentration to be hostile to technical efficiency (see Scherer 1965 and Caves 1992). They have also found that international competition has a more limited effect on efficiency than the domestic competition.
- No systematically positive relationship has been found to exist between firm size and profits or export activity. Instead, firm-specific characteristics unrelated to size determine performance in these areas.
- Some evidence suggests a positive relationship between industrial concentration and productivity growth, but one with an important link to research and development activity. Industries with greater research and

development expenditures tend to have greater productivity increases. Within industries, however, research and development intensity tends to increase with firm size up to an intermediate level and then decline, except in a few industries such as chemicals and petroleum products. But smaller firms show a higher output of patents per dollar spent on research and development (see Scherer 1965 and Bound and others 1984).

- Conglomerates have emerged for a variety of reasons: maximizing growth, reducing risk through diversification, and lowering the cost of financing through internal reallocation of funds. But the performance of conglomerates has lagged behind that of specialized firms, in part because of the high cost of coordinating diverse, unrelated economic activities. Diversification by large enterprises has also been found to reduce technical efficiency. The recent trend is toward divesting unrelated business and focusing on core areas.

If firms or industries are seen as having unexplored economies of scale or other advantages of size, a question that needs to be answered is why these are not being exploited. If the domestic market is small, these economies can be attained through exports—as in Singapore, Sweden, and Taiwan (China).

At the same time, government efforts to promote large firms and conglomerates pose significant risks. Large, sheltered firms may engage in rent-seeking or wield undue political and economic influence. While rent-seeking further accentuates resource misallocation and is an obstacle to structural adjustment, the influence of large firms and conglomerates can undermine emerging democratic institutions and the independence of government economic policymaking.

*The East Asian economies: What has been the experience?*

In East Asia, a unique mix of government intervention and competitive discipline has been applied to foster industrial growth.<sup>19</sup> The mix of strategies varies across the 23 economies in East Asia that have experienced higher-than-average growth. But one common denominator is a high degree of interfirm rivalry and exposure to competition, both domestic and international. During Japan's development phase, although its industrial sectors were protected from foreign competition, a large domestic market supported enough firms to ensure vigorous competition in both domestic and international markets. In Taiwan (China), tax incentives direct investments into key industries, but firms are exposed to international competition. Most Taiwanese firms, including export-oriented

ones, tend to be relatively small and to compete intensely with one another. In Korea, which has permitted large industrial conglomerates to emerge, strong government policies have supplied the discipline usually imposed by market forces. Conglomerates must compete for government subsidies, for access to credit, and for foreign exchange. Strict export performance standards are set and enforced on the basis of international prices, which provide an objective yardstick for gauging success.

In developing these and other industrial strategies, the East Asian economies have established cooperative government-business relationships. Private sector participants have contributed to and agreed to abide by a transparent set of rules and procedures that they strictly observe.<sup>20</sup> Most of the East Asian economies quickly abandoned their initial policies of trade protection and import substitution. Few entry and exit controls, licenses, or other regulations inhibit industrial activity. A stable macroeconomic environment and consistency in government policies toward foreign investment, technology, and taxation have also aided progress. Clearly articulated and transparent industrial and trade policies have minimized rent-seeking behavior. While specific industrial sectors may enjoy protection, individual firms do not. All sectoral participants receive equal treatment.

Michael Porter, in *The Competitive Advantage of Nations*, has observed that:

Few roles of government are more important to the upgrading of an economy than ensuring vigorous domestic rivalry. Rivalry at home is not only uniquely important to fostering innovation but benefits national industry. . . . In fact, creating a dominant domestic competitor rarely results in international competitive advantage. Firms that do not have to compete at home rarely succeed abroad. Economies of scale are best gained through selling globally, not through dominating the home market. (1990, p. 662)

Porter based his conclusion on firm- and industry-specific studies across a number of countries that included the East Asian economies of Japan and Korea as well as Germany and Sweden.

### **Main provisions of competition law**

Competition laws generally contain conduct and structural provisions relating to business activity, as well as additional procedural provisions covering administration and enforcement. An advocacy role in promoting com-

petition objectives in government policymaking is often part of the law.

#### *Horizontal restraints and collusion*

Horizontal restraints on trade—explicit or implicit agreements entered into by enterprises in the same market for their mutual benefit—are the core area of concern in existing competition laws. The term “collusion” generally refers to agreements between enterprises intended to restrict output and raise or fix prices. These agreements clearly are detrimental to competition.<sup>21</sup> Other types of agreements that allow participants to benefit from coordination are more ambiguous in their effects.

There is a consensus among economists and legal professionals as well as among competition law practitioners that agreements between enterprises to fix prices, to reduce output, to allocate customers, or to bid collusively are anticompetitive. Just as independence and rivalry among competing suppliers in a market lie at the core of competition, prohibitions against collusion are central to an appropriate competition law. Countries moving toward market-oriented principles from a more controlled and regulated environment have the greatest need for penalties against cartel-like behavior. Here agreements between enterprises to divide markets may have been accepted practice in the past. In such environments the competition agency has a critical role to play in changing both the mindset of enterprise managers and the code of conduct of firms, so that independent offers for the business of buyers become the norm rather than the exception.

Cartels generally involve attempts by two or more domestic enterprises to carve up and reduce competition in local markets. But cartels also are possible between two or more foreign firms anxious to share the domestic market of specific countries.<sup>22</sup> Without laws prohibiting such practices, small, open economies have little recourse against the anticompetitive behavior of increasingly global corporations. Applying domestic competition law to foreign firms does, however, raise enforcement problems related to extraterritoriality and access to evidence.

While there is consensus that collusion is the most serious violation of competition laws, in practice it can be difficult for competition law enforcement agencies to enforce tough rules against this behavior. In the absence of concrete evidence of formal cartel agreements, enforcement agencies generally rely on the testimony of participants in collusive schemes or their associates. It is therefore important for countries to include a provision in their competition law offering participants immunity

for cooperation in uncovering and prosecuting anticompetitive behavior. It also may be desirable, in cases of collusion, to allow injured parties to sue for damages sustained (or a multiple thereof).

Much conceptual work relying on dynamic game theory has recently been done to understand how oligopolistic firms might be able to collude in a *noncooperative* manner (“tacit collusion”). To sustain collusion without overt agreements, firms must credibly threaten to punish those that deviate from implicit agreements. But they can retaliate only when they learn that a firm has deviated. Since cheating on cartel agreements is not always observable, price wars may occur even when firms successfully collude much of the time (Green and Porter 1984; Abreu, Pearce, and Stachetti 1990). So, evidence of occasional price wars does not prove the absence of collusion.

Because sustaining collusion becomes more difficult as the number of firms increases, the best defense against tacit collusion lies in facilitating entry into markets. This has important policy implications. By reducing barriers to international and domestic trade, competition agencies can make tacit collusion increasingly difficult. As the number of firms increases, so do the costs of organizing collusion, and it becomes harder to detect cheating. This makes firms more likely to deviate.

Given the difficulty of proving tacit collusion, competition agencies should focus on attacking those practices that make tacit collusion easier. For example, industry trade associations may collect detailed information on the transactions of their members or permit members to cross-check price quotations, reducing secrecy and facilitating collusion. This specific function should be discouraged where collusion is likely, though caution should be exercised not to discourage the many other useful services that trade associations provide. Highly similar products also facilitate collusion, since differentiation in quality, durability, costs, and other attributes makes uniform price agreements more difficult to reach and monitor. For more heterogeneous goods, agreements on product specifications and standards as well as rule-of-thumb pricing help firms to reduce informational complexity and better coordinate strategies.<sup>23</sup> Competition agencies need to be aware of steps that oligopolists may take to promote collusion. Vigilance, even more than penalties, can serve as a strong deterrent.

Special considerations arise in countries experiencing significant changes in prices or in which price signals remain distorted because of government controls. Where inflation is high and persistent or relative prices are changing frequently because of market liberalization,

enterprises may find it difficult to coordinate a pricing strategy. In these situations, collusion is more likely to occur in the form of market sharing, either along geographic lines or based on consumer demand characteristics. Past government price controls and other policy-based interventions may facilitate collusion, by providing a “focal point” for coordination or by posing barriers to entry. It is important for competition law agencies in such environments not to focus exclusively or even primarily on potential price-fixing arrangements but to broaden their monitoring activities to include all types of market-sharing arrangements.

Not all forms of cooperation between enterprises are necessarily harmful to competition. An active debate is under way on the desirability of certain kinds of interfirm cooperation that may promote efficiency and dynamic change. Proponents of the industrial policy school, for example, have advocated export cartels, arguing that it is in a country’s interest to allow firms to fix prices or outputs for export or to divide export markets. Many competition law statutes exempt such agreements, as long as the export cartel does not restrain domestic competition. The rationale for permitting export cartels is that they facilitate penetration of foreign markets, transfer income from foreign consumers to domestic producers, and improve the trade balance. The GATT remains relatively silent on export cartels, having no jurisdiction over cartels that are not government-sponsored or government-sponsored cartels that fix prices (just as taxes on exports are permissible).<sup>24</sup> In practice, however, there are relatively few goods for which a national export cartel could have an appreciable effect on price, as shown by the unsatisfactory performance of commodity cartels and the inherent difficulties in coordination.<sup>25</sup> Furthermore, in these as in all cooperative agreements, there remains the risk that information will be shared that will facilitate collusion in the domestic market. As a practical matter, competition agencies in developing and emerging market economies should be vigilant whenever domestic enterprises in concentrated industries cooperate with the stated motive of forming an export cartel.

In addition to export cartels, countries have sometimes permitted other types of cartels (recession/depression, rationalization, specialization, research and development, standardization). The justification for these cartels has been that they allow participants to attain advantages that would otherwise be unavailable, such as achieving scale economies, sharing risk and information, and overcoming coordination problems. To facilitate industrial restructuring, including the capacity reduction

in declining industries, production cartels have been permitted in some countries as short-term measures to raise or at least stabilize prices so as to finance adjustment.<sup>26</sup> It is critical for the competition agency to monitor the effect of such cartels on competition and ensure that exemptions apply for a finite period. Exemptions also are sometimes granted for the sharing of statistics, for research joint ventures that allow firms to share the risks, expenditures, and benefits associated with a specific research project, or for other projects in which coordination can avoid wasteful duplication or take advantage of participating firms’ complementary assets. In all these cases, tension is bound to arise between the possibility of efficiency gains and the risk of horizontal collusion. In light of this, there is need for research to identify appropriate arrangements that enhance efficiencies but do not form a basis for output reduction. In the meantime, close scrutiny of all information-sharing agreements appears to be warranted.

#### *Vertical restraints and market access*

Vertical restraints—contractual agreements between suppliers and purchasers in separate upstream and downstream markets—raise fewer concerns for competition policy than horizontal agreements. These contractual agreements, which provide a variety of means to integrate decisionmaking, can include:

- Restrictions on pricing (for example, *resale price maintenance*, a provision dictating a final downstream price).
- Restrictions on market partner choice (*exclusive dealing*, when the buyer is induced to deal only in one seller’s products, and *refusal to deal or supply*, when buyers encounter difficulties in obtaining products from suppliers, usually aimed at coercing them into adopting an anti-competitive practice).
- Restrictions on location choice (*exclusive territory*, when the geographic territory in which the buyer may resell is limited).
- Restrictions on purchase choice (*tying*, when the availability of one product is contingent on the purchase of other goods or services, and *full line forcing*, a form of tying in which the manufacturer requires the distributor to carry all of its products).
- Restrictions on other business practices (such as trade name, brand logo, store setup, and form and content of advertising, as in *franchising agreements*).<sup>27</sup>

Whether or not vertical restraints pose an anticompetitive concern and should be subject to competition law has been a controversial issue in recent years. The controversy centers on whether vertical agreements, though

on the face restrictive, may nonetheless promote economic efficiency by overcoming market failures.

Competition policy concerns arise because vertical restraints can enhance the market power of major participants, either by facilitating collusion (reducing rivalry among established firms in upstream or downstream markets) or by permitting market foreclosure (excluding existing or new competitors from the market). It has been argued that resale price maintenance, for example, can help to sustain cartel prices by making it easier for upstream members of the cartel to monitor cheating. It has also been argued that vertical agreements that restrict downstream intrabrand competition (between retailers selling the same brand) may also decrease upstream intrabrand competition by making wholesale price cuts less attractive (Rey and Stiglitz 1988). While more research is needed in this area, there is a consensus that vertical restraints pose a risk of collusion only to the extent that they make collusion easier, but that they will be ineffective unless market conditions are already favorable to horizontal agreements. This underlines the need for competition officials to monitor conditions that permit horizontal agreements.

Recent theoretical literature in this area suggests that a much stronger case can be made for competition policy concern when vertical agreements have foreclosure effects. A dominant firm may be able to force competitors out of the market or prevent new entry by foreclosing access to other upstream or downstream markets. For example, through long-term contracts with downstream distributors, an upstream enterprise can create entry barriers for competing firms. When most of a downstream market is locked into trading with one supplier, other potential suppliers may find the residual market too small to enter. This view has been challenged on the ground that distributors would be unwilling to sign contracts that strengthen the seller's monopoly position; the only reason they might be willing to sign such contracts is because they enhance efficiency (Posner 1976; Bork 1978). But Aghion and Bolton (1987) have shown that an incumbent manufacturer and a distributor have common interests. By signing a contract, they achieve monopoly power over other potential entrants by making entry difficult or unattractive. By sharing the extra rent made possible by a long-term contract, a dominant upstream enterprise can induce distributors to accept the contract. Similarly, a dominant downstream distributor can create barriers to entry for competitors by sharing excess earnings with an upstream manufacturer (Comanor and Rey forthcoming). An enterprise can also raise rivals' costs by obtaining exclusive

rights to inputs or distribution channels (for example, to particular retailers or retail locations). An upstream enterprise with a large market share may be able to increase the distribution costs of rivals and thereby deter entry—for example, by using long-term contracts with exclusive dealing provisions to tie up the best locations.<sup>28</sup>

While these arguments on the whole suggest that vertical restraints with market foreclosure effects should be taken seriously as potential anticompetitive practices, a great deal of recent theoretical work has been undertaken to explain these practices on efficiency grounds. The debate mirrors the tensions among different schools of thought in industrial organization economics generally. Proponents of the Chicago school are the most ardent defenders of the efficiency rationale. As a result of their work, it is now generally acknowledged that vertical restraints can have strong pro-competitive effects, by more closely aligning the incentives of upstream and downstream firms. These restraints overcome a market failure whenever they internalize externalities that arise when downstream enterprises ignore the effect of their actions on upstream profits. Restraints can, for example, overcome a downstream moral hazard problem and improve the supply of retail service and quality to consumers. If there is no assignment of exclusive territories, for example, retailers may be inhibited by fear of free-riding from undertaking promotional or other customer service efforts, leaving all worse off. Similarly, manufacturers may refrain from investing in distribution-related facilities that increase sales in the absence of an exclusive dealing provision.

Since vertical restraints may either promote or reduce economic efficiency, the general competition policy prescription for industrial countries is to avoid simplistic *per se* rules. In general, policymakers and watchdogs should consider market structure when determining whether a particular vertical restraint is acceptable from a competition point of view. Where no market power exists, or where interbrand competition is sufficiently strong, the net welfare effect of vertical restraints should be positive. Therefore it is desirable to establish different rules and enforcement guidelines, depending on the state of competition in the upstream and downstream markets. For example, small, new franchisers should receive quasi-automatic permission to include vertical restraints in agreements, while franchisers with a dominant position in their market should be required to provide a more detailed justification on efficiency grounds.<sup>29</sup>

For developing and emerging countries more research is needed on the effect of vertical restraints. The current

state of research does not allow us to confidently propose even rule-of-thumb guidelines. For example, it may be argued that smaller economies should be more active in enforcing the law against vertical restraints because market structure is typically more concentrated, competition among distributors limited, and entry more difficult. Since opportunities for contract renegotiation allow new entrants to compete, it may be prudent in such economies to bar contracts between manufacturers and distributors for periods longer than two or three years, especially those that are exclusionary.<sup>30</sup> On the other hand, it is extremely difficult to distinguish between contracts aimed at exclusion and those concluded for innocuous purposes such as risk sharing. Risk and uncertainty, as well as information and capital market imperfections, problems that long-term contracts can better address, are likely to be much more acute in developing and emerging market economies.<sup>31</sup>

#### *Abuse of dominant market position*

An area of concern to competition authorities is that dominant firms may be able to adopt anticompetitive business practices to maintain or enhance their position.<sup>32</sup> Such practices may be aimed at preventing the entry or inducing the exit of competitors or modifying the business behavior of actual or potential competitors. These practices may include exclusive dealing, market foreclosure through vertical integration, tied selling, the control of scarce facilities and vital inputs or distribution channels, price and nonprice predation, and interfirm contractual arrangements. Provisions against such monopolistic conduct—known as abuse-of-dominance, or AOD, provisions—form an important part of competition law.

AOD provisions also are one of the most controversial areas of competition policy. One issue is what constitutes dominance. Different countries specify different market share thresholds.<sup>33</sup> There is no critical market share or concentration level that clearly signals the degree of competition and monopoly in a market. Moreover, there is disagreement about how to delineate the relevant size of the market.<sup>34</sup> And, as mentioned above, the debates among economists have changed the general view on whether market dominance by a firm or group of firms necessarily implies monopolistic conduct. In the absence of barriers to entry, it is difficult for large or dominant firms to exercise market power. Moreover, firms may be large because they are more efficient and innovative. Large firm size may be a consequence of vigorous competition rather than of monopolistic practices. A policy that tends to impede growth or force the break up of large firms may therefore prevent economic efficiency.

It has been argued that both price and nonprice predatory behavior by large firms to induce the exit of competitors is not sustainable over the long term.<sup>35</sup> In the absence of barriers to entry, exiting firms can, in theory, reenter the market once incumbent dominant firms raise their prices. This possibility should act as a check on pricing and output policies, keeping them competitive. Moreover, as mentioned above, many allegedly anticompetitive business practices, such as exclusive dealing and other types of vertical restraints, can be justified on the grounds that they enhance efficiency.

Various arguments in the economics literature suggest, however, that large incumbent firms can benefit from “first mover advantages,” the presence of sunk costs, and asymmetric information. To maintain and strengthen their market position, these firms may engage in “strategic entry deterrence” by raising the costs of rival firms, building reputation, and sending signals about being tough to competitors.<sup>36</sup>

Given the range and the lack of resolution of these issues, AOD cases in many jurisdictions have appropriately focused on the economic effects of particular practices. Thus, competition authorities have focused on the practices that dominant firms adopt to prevent competitors from expanding or from entering a market rather than on the exercise of market power. Competition law and its practice have led to a list of factors that can help authorities determine whether a dominant firm is preventing new competition.<sup>37</sup> This list, which includes price and nonprice predation, market foreclosure, and vertical squeezing, is of course not without controversy. We posit that superior competitive performance entails gaining advantages over competitors rather than actively creating disadvantages for them. While distinguishing between these two types of business conduct is not always easy, the latter tends to misallocate managerial and other resources and should thus be resolutely discouraged.

In industrial economies, dominant firms have tended to emerge as a result of market forces. By contrast, in developing countries such firms often have arisen as a result of government industrial policy and private rent-seeking behavior, while in transition economies they have come about almost exclusively because of central planning and administrative fiat. This raises the question of whether countries should actively pursue a policy of breaking up large firms in order to create the structural basis for increased competition. And, if so, do they run the risk of reducing economic efficiency? Although a definitive answer requires further study, such risks are likely to be low. Thus, competition authorities in devel-

oping and emerging economies may have more freedom to adopt structural solutions to correct competition problems stemming from large dominant firms than their counterparts in industrial countries.

### *Mergers and acquisitions*

Determining whether large firms promote efficient resource allocation or lead to exercise of market power is no easy task—particularly in developing and emerging market economies with small home markets and high levels of domestic concentration. Although firms in these countries may be large relative to the domestic market, they may still be too small to achieve internationally competitive scales of production. As a result, even large firms may have to grow larger to fully exploit potential efficiencies.

Competition authorities in such situations have conflicting responsibilities: to encourage competition but also to promote economic efficiency. One possible solution to this dilemma is to foster free trade. But as discussed more fully in the next section, factors other than tariff and nontariff barriers to trade can still segment domestic and international markets. Pro-competition policies, such as deregulation, privatization, and the removal of restrictions on foreign investment and ownership, can also foster the entry and exit of firms and the mobility of resources. But introducing such policies does not absolve competition authorities from the need to review corporate mergers, acquisitions, and restructuring to ensure that they do not lessen competition.

A variety of motives underlie corporate decisions to engage in mergers and acquisitions (see Ravenscraft and Scherer 1987 and Steiner 1975). These motives tend to differ by the type of merger. There are three main kinds:

- Horizontal mergers among two or more firms in the same line of business and in the same market. Such mergers reduce the number of competing firms and increase market concentration. The motives could be to acquire market power or to increase efficiencies.
- Vertical mergers among firms engaged in different (upstream–downstream) stages of production and marketing. This type of merger activity may foreclose sources of inputs or distribution channels to competitors, or result in efficiencies by reducing transaction and other costs through internalization of different stages of production–distribution.
- Conglomerate mergers among firms in diversified or unrelated businesses. The motive for these mergers may be to reduce risk and exploit various financial and other economies. However, they could also give rise to cross-

subsidization and reciprocal arrangements to limit competition in different markets.

While competition concerns arise primarily in the case of horizontal mergers in industrial countries, in developing and emerging market economies all three types of mergers and acquisitions may pose competition problems. But, as we discuss below, this risk has to be seen in the context of the relative overall priority of enforcing different provisions of competition law in these countries. Moreover, many of the analytical problems posed by a small number of market participants and oligopolistic interdependence are more likely to occur in small economies than in larger ones.

One of the difficulties in analyzing horizontal mergers is evaluating the extent to which prices before the merger reflect the exercise of market power. For example, a leading firm may suppress its prices and profits by frequently adding capacity, thus reducing opportunities for higher-cost producers or producers of substitutes to enter the market.<sup>38</sup> Or the leading firm may have raised its prices to the maximum the market will bear to realize supra-competitive profits. In this situation, a horizontal acquisition is less likely to permit the firm to increase prices further under prevailing demand conditions.<sup>39</sup> The merger may not make matters worse, but it may nevertheless entrench existing producers and impede the development of competition. In contrast, if existing prices reflect high costs, an efficiency-generating merger may lower prices in the absence of collusion by the surviving firms. Even if incentives to increase prices or reduce output are lacking, however, the reduction in the number of competing firms may increase oligopolistic interdependence, causing organizational “slack” and other “x-inefficiencies” to set in (see Leibenstein 1966). Monopolies and oligopolies can lead not only to high prices but to high costs. Independent firm behavior and rivalry act to minimize costs.

Another difficulty lies in evaluating the tradeoff in a horizontal merger between the reduction in competition and the potential gains in economic efficiency. Williamson (1968) has demonstrated that a very modest decrease in costs can offset the adverse economic welfare effects of a large increase in prices following an anticompetitive merger. This condition holds across a wide spectrum of demand elasticities. But these results depend on what assumptions are made about pre- and postmerger market power and the choice of model of price formation.<sup>40</sup> The theoretical underpinnings of this “tradeoff” approach for evaluating mergers have also been criticized by some economists (see Cowling, Stoneman, and Cubbin 1980, De Prano and Nugent 1969, and Jackson 1970).

There are also practical difficulties in defining and quantifying what constitutes acceptable efficiencies. Generally speaking, cost reductions arise from economies of scale and scope as well as savings from the integration of production facilities, diversification, rationalization, and financial economies. But efficiencies from improvements in product quality, introduction of new products and innovations, and increased product choice and service cannot always be translated into price and cost terms. A related issue is who should bear the onus of gathering, analyzing, and presenting such information. Competition authorities are unlikely to be able to independently collect this information. In several jurisdictions it has become the convention for the merging parties to supply the information. But since the merging firms have a vested interest in completing their transaction, they are likely to exaggerate the efficiency gains. Moreover, before the merger, the firms are unlikely to have accurate product, cost, and other sensitive business information about one another. For these reasons, efficiency arguments in merger cases should be carefully scrutinized and sometimes heavily discounted.

Despite these difficulties, competition authorities should weigh efficiency claims in their decisions. For one thing, taking efficiency effects into account provides a useful check on a purely structural approach to horizontal mergers. In Canada and the United States, efficiencies can provide grounds for exempting otherwise anticompetitive mergers from competition laws. Efficiencies are also an important factor in the evaluation of mergers in the European Union, Italy, and the United Kingdom. And several developing economies, including Colombia and Mexico, have explicitly incorporated efficiency considerations in their merger review process.

"Vertical interdependencies" between buyers and sellers of intermediate products also have a bearing on the efficiency question. In addition to high levels of seller concentration, high levels of buyer concentration may exist simultaneously. This leads to bilateral oligopoly (monopoly)—oligopsony (monopsony) problems, which have been described in the industrial organization economics literature as being "indeterminate with a vengeance" (Scherer 1980, p. 299). Vertical mergers resulting in increased integration may solve interfirm bargaining problems, giving rise to significant efficiencies. But they may also increase the incidence of bilateral monopolies and create new barriers to entry by foreclosing input sources and distribution channels that competitors might otherwise use. New firms may have to enter the market at more than one stage of the product cycle, a considerably more difficult and expensive proposition.

Preserving an adequate number of input sources and distribution channels is critical to maintaining and encouraging competition. This fact is supported by findings by Kwoka (1979) and Kwoka and Ravenscraft (1986) that the presence of the third or fourth largest firm in a concentrated industry significantly reduces the price or profit margins of the other large firms. And Scherer and others (1975) in a study of multiplant operations in Canada—an economy with small, highly concentrated markets—found that:

Consumers, and particularly industrial buyers, exhibited a strong preference for having at least two alternative supply sources, even if it meant fragmenting what would otherwise be a natural monopoly and causing unit costs (although not necessarily price) to be higher. Industrial buyers of paints, bottles, cement, steel, and batteries evidently value both the security against total interruption of supplies and the bargaining power conferred by being able to play one producer off against the other. (p. 134)

In many horizontal merger cases, the combined market share of the merging parties erodes following the merger as customers shift purchases to other firms. These findings imply that competition authorities should consider the number of competing firms and alternative sources of supply available to customers as well as potential gains in efficiency if they hope to preserve competition. In some instances potential efficiencies from eliminating the cost of suboptimal operations may not be as large as empirical studies on economies of scale would suggest. In other cases, instruments such as consent decrees, limits on long-term contracts, and nondiscriminatory supply guarantees may be necessary. Competition authorities may find it advisable to limit resolution of merger cases to the enactment of other pro-competition policies, such as the reduction of tariffs or quotas. Canadian competition law contains provisions for lowering customs duties if they are likely to facilitate anticompetitive practices. Such measures can help foster development of markets in emerging market economies, where most interfirm relationships, horizontal as well as vertical, have been developed through administrative fiat rather than through efforts to maximize profits or promote efficiency. Liberalization of trade along with increased exposure to domestic and international competition will force firms to reconfigure their operations along more efficient lines and prevent them from gravitating back to comfortable but inefficient relationships.

The concentration of economic wealth among a small number of families and groups in developing countries coupled with high levels of industry concentration also raise competition problems with respect to conglomerate mergers. While conglomerates, in theory, emerge to maximize growth or diversify risk, the presence in developing countries of a small pool of managerial talent or underdeveloped capital markets may be their real causes. Today, in the industrial countries, conglomerates are divesting unrelated businesses to focus on “core” activities. In the developing world, however, this phenomenon has yet to occur. Given the small number of corporate players and limited pool of managerial talent in developing countries, interlocking directorates are common. These can encourage the sharing of information and the coordination of anticompetitive strategy. Members of conglomerates are often inclined to become suppliers and customers of one another. Interdependent relationships that limit access by competitors to input sources and distribution channels may consequently emerge. Perhaps more important, conglomerates that loom large in the economy are in a position to engage in rent-seeking behavior.

#### *Administration and enforcement*

Whether a competition law is desirable or even enforceable in developing and emerging market economies is the subject of an active debate. The main concern is that an inappropriately administered and enforced competition law can be worse than no law at all. By creating an additional mechanism for intervention in the economy, the law provides opportunities to stifle rather than enhance competition. This argument presumes that government institutions, both national and local, may lack competence or, worse still, that they may be hostage to special interests.<sup>41</sup> Opponents of a competition law claim that most of these countries lack the specialized staff institutional knowledge and tradition of analytical rigor needed to ensure beneficial interventions. They worry that laws may be applied too intrusively or in such a way that they impinge on the very freedom and rewards that markets are supposed to bestow.<sup>42</sup> Moreover, by establishing a competition agency, competition law may set up another target for rent-seeking interest groups. Opponents worry that the agency or courts may take actions against efficient but politically unpopular firms and industries or may promote the private interests of individual enterprises.<sup>43</sup>

Arguments concerning governance and administrative capacity are clearly relevant for all areas of government involvement. While these arguments undoubtedly have some validity, they have as much validity for industrial as

for developing and emerging market economies. It is therefore instructive to review how other countries have dealt with these concerns.

Some countries rely solely on the criminal (penal) or civil codes and courts to discover abuses; others rely on government ministries and administrative tribunals; and still others rely on both. In some jurisdictions price agreements and certain cartels are officially sanctioned and regulated; in others they are prohibited. The United States has multiple competition agencies and both federal and state antitrust laws. Most other countries have placed this responsibility under the sole authority of the central government. Countries considering adopting or substantially revising existing competition laws need to address these issues.

Another key concern is who should enforce the law: a public agency, private parties, or some mix of the two. Naturally, the choice of an enforcement agency is likely to affect enforcement effectiveness. Granting sole enforcement power to an organ of the government invites problems related to lack of transparency, corruption, and misuse of power, problems that reliance on the legal system does not necessarily alleviate. In many countries court proceedings are so slow and cumbersome that they render laws ineffective. Incentives for private enforcement, such as multiple damage awards, might seem preferable to dependence on public enforcement. But this approach carries the risk of increasing nuisance litigation or other “legal blackmail.” In competition cases, as with other commercial and business disputes, matters must be resolved expeditiously. Delays increase costs, which can deter investment, raise barriers to entry, and induce economic distortions.

No research has been conducted on the most efficient organizational structure for the administration and enforcement of competition policy. In most countries the institutions and procedures in place have evolved as a result of specific cases, legal precedents, experience, and even historical accidents. A tradeoff clearly exists between the benefits of rules and the ability to respond flexibly to changing conditions. Insulation from “capture” by political or business interest groups is also critical to effectiveness. Moreover, since countries differ in their institutional capacity, it is an open question whether they should also differ in their mix of rule-based and discretionary decision-making approaches.

The experience of industrial countries with long-standing competition regimes suggests certain principles that ought to be observed in designing an institutional framework for the effective implementation of competition policy.<sup>44</sup> The basic principles are:

- The competition policy agency should be independent and insulated from political and budgetary interference.
- The agency needs to be accountable. One way to ensure this is to require that it publish an annual report and submit it to the legislature or a special committee.
- The competition law should separate the investigative, prosecutorial, and adjudicative functions. This prevents the competition policy agency from becoming investigator, judge, prosecutor, and jury rolled into one.
- The process should incorporate a system of checks and balances that provides for the right of appeal, the right to review decisions, and access to information on legal and economic interpretation of the law. Administrative procedures and regulations should be transparent.
- Proceedings and the resolution of cases should be expeditious to avoid unnecessary business-related costs. But commercially sensitive business information should be safeguarded.
- To deter anticompetitive business practices, the law should provide for penalties, including meaningful fines, and other remedial measures.

In addition, the competition policy agency should be granted a statutory role in formulating policies affecting competition. The agency should also be granted the right to comment on policy changes. A competition advocacy role lets an agency counter or at least minimize the adverse effects of rent-seeking behavior found in all countries to a degree, but particularly in developing and emerging market economies. Given limited administrative capacity and enforcement experience in these economies, some commentators view this role as the most important if not the only worthwhile function a competition enforcement agency can perform (Rodriguez and Williams 1994). They argue that competition advocacy can reduce the possibility of misapplying competition law. Yet both the competition advocacy and enforcement functions of an agency are important. The correct mix will depend on the stage of development of the country in question. It is often advisable to first grant a competition agency an advocacy role, next enforce the rule-based provisions of the law, and only after accumulating sufficient experience apply discretionary provisions such as those dealing with abuse of dominance and mergers and acquisitions.

### **Trade policy as competition policy**

#### *The "trade liberalization alone" approach*

In many small developing and emerging market economies, where industry concentration facilitates anti-

competitive business practices, competition authorities face a policy dilemma. If they follow a vigorous structural approach, they may block or impede potential efficiencies. Eastman and Stykolt (1960) and Bhagwati (1965) have suggested that this problem can be resolved by increasing the exposure of domestic markets to international trade. In an integrated, perfectly competitive world market with no barriers to trade, domestic monopolists or oligopolists lose their ability to exercise market power irrespective of actual imports' share of the domestic market, in view of the threat of potential import competition.

Proponents of this approach believe that, following appropriate liberalization of international trade, imports alone will temper market power. They argue that free trade makes competition law largely irrelevant in small, open economies. Markets that are local, such as services and distribution, are assumed to be easy to enter on a small scale and therefore intensely competitive (Godek 1992).

The view that imports limit market power gains support from studies that find differing degrees of convergence between domestic and international prices in the face of trade liberalization and a negative relationship between price and cost or profit margins and imports.<sup>45</sup> But some recent empirical studies suggest that effects of trade liberalization may be less significant than previously thought, raising questions about the true effect of trade liberalization on competition.<sup>46</sup>

#### *The insufficiency of trade liberalization as a guarantor of competition*

The pro-competitive effects of tariff reductions may be diluted if the import supply curve is upward sloping, that is, if it is not perfectly elastic. This occurs when increased demand for imports can be met only at higher prices or when imports are comparatively insensitive to changes in domestic prices. It is possible to construct alternative theoretical models where imports meet only a small part of domestic demand and are provided by fringe suppliers facing a residual demand curve that domestic oligopolists have already factored into their pricing strategies (see Perry and Porter 1985 and Ross 1988a). In addition, early free trade models date from an era of fixed exchange rate regimes. In an environment of floating exchange rates, if domestic firms fail to rationalize high-cost operations and improve productivity, the domestic currency is likely to depreciate, offering new protection from import competition.

Furthermore, trade policy consists of more than tariff policy. Quotas, voluntary export restraints (VERs), and

antidumping and countervailing duties are among the instruments that governments can wield to limit import competition. While a true “free trade” policy would require that all such measures be removed, in reality this never happens. Indeed, as import tariffs are liberalized, the pressure to invoke other measures only increases. Over the past few years, as more developing countries have liberalized trade policies, a simultaneous movement has occurred to put in place systems of protection against dumping and subsidies. By the end of the 1980s, more than 30 developing countries had become signatories or observers of the GATT antidumping code, and in 1980–92 more than 2,000 antidumping and countervailing duty actions were taken.<sup>47</sup>

From a competition policy perspective, existing antidumping law as sanctioned by the GATT is *not* structured to promote economic efficiency. The current law provides a process through which domestic firms injured by lower-priced goods from abroad can seek (and often obtain) protection in the form of duties on the foreign goods. The law is intended to protect domestic competitors rather than the competitive process. From a political economy perspective, therefore, developing and emerging market economies would be better off if they did not replicate the full U.S. or EU arsenal. Given differences in country and market size, that arsenal is likely to cause greater harm to smaller economies, through the mechanism of reciprocity. Policymakers should look for alternative instruments to help domestic industries injured by foreign competition.

A number of alternative actions offer ways both to deal with discriminatory pricing and to enhance competition. A common feature of these options is restraint in the exercise of traditional antidumping measures. For example, parties to free trade agreements should replace antidumping provisions with harmonized competition laws (though how best to achieve such harmonization in a specific country requires further research). In countries with antidumping laws in place, national competition agencies can play a critical role in ensuring that enforcement of antidumping does not excessively constrain competition. Competition agencies should promote the use of binational or supranational panels and consultations. And they should explore the replacement of antidumping measures with an expanded GATT-sanctioned safeguards code.

Even if trade barriers, such as antidumping duties, are eliminated, other factors can impede the pro-competitive effect of trade liberalization. First an increasing share of economic activity in developing as well as industrial coun-

tries relates to nontradable goods and services. These include high weight-to-value products with high transport costs (such as cement and steel), perishables (such as food), and legal, financial, and other services.<sup>48</sup> Second, in the absence of effective competition, domestic firms can raise prices up to the international price plus transport costs and still keep out imports. Third, interfirm contractual arrangements and vertical integration may prevent the development of new sources of inputs or new distribution channels. This problem has been cited by many American firms as limiting their ability to gain access to markets in Japan and forms part of the “framework” and “structural impediments initiatives” discussions under way between the U.S. and Japanese governments. Fourth, international cartels may divide up markets through price-fixing or geographic market-sharing agreements. And importers and foreign firms may find it more profitable to become parties to domestic anti-competitive arrangements than to compete. Indeed, foreign firms have often been charged with just this offense in several jurisdictions. In concentrated industries such as pharmaceuticals, petro-chemicals, and telecommunications equipment, where the total number of firms worldwide is small, such arrangements are particularly common. Fifth, differences in income, tastes, and culture and in product safety, consumer protection, and technical standards may also separate markets.

For these and other such reasons, liberalized trade cannot effectively substitute for competition law. The two should be viewed as complementary. Both can help a country to elicit the maximum benefits obtainable from specialization and scale economies. Competition law can lower barriers to trade and investment and thereby enlarge markets by improving access to profitable business opportunities. Chapter 15 of the North American Free Trade Agreement between Canada, Mexico, and the United States explicitly recognized this view, and to enhance business and investor confidence, Mexico has decided to modernize its competition law.

Mexico’s decision highlights an important final point. In a world where multinational companies have grown accustomed to operating under competition laws, the absence of competition law or a poorly designed one in a country can act as a barrier to trade and foreign investment. As local enterprises begin to operate more in international markets, a national law in harmony with competitors’ laws makes it easier for them to adapt. It also spares foreign firms any additional hurdles in their business activities.<sup>49</sup>

**International dimensions of competition policy**

Increased globalization of markets has focused attention on the need to reconcile competition policies. It also underscores the need to integrate competition policies with those governing trade and investment. Over the past decade, foreign direct investment (FDI) and technology transfers among firms rose faster than world trade, world production, or domestic investment. Mergers and acquisitions, joint ventures, and strategic alliances have become preferred ways to enter foreign markets (see Safarian 1993). Countries vary in how they approach and administer competition policy. They differ in policy objectives, legal systems, institutional arrangements, information-filing requirements, time frames, and procedures for reviewing cases. As transborder transactions increase among firms, friction among systems is bound to arise (see Ostry 1990, forthcoming). Since differences among systems increase transaction costs and impede entry, greater harmonization of competition policy would increase global economic efficiency.

An empirical example illustrates these points. The merger between Gillette and Wilkinson—two multinational razor blade manufacturers—led to a simultaneous review of the transaction by authorities in 14 jurisdictions.<sup>50</sup> The focus and nature of the reviews varied extensively across the 14 jurisdictions. While in the end the deal was allowed to precede with minor alterations prompted by different authorities, the delays, procedures, and compliance costs the firms experienced were far from trivial.

Until recently these kinds of issues arose primarily among industrial countries, where the bulk of FDI has occurred. But developing and emerging market economies are now beginning to confront similar issues.<sup>51</sup> A number of bilateral and multilateral approaches to the enforcement of competition law have been forged by different countries. Canada and the United States have signed a memorandum of understanding (MOU) and a mutual legal assistance treaty. MOUs have been drafted though not yet finalized between Canada and the European Union and the United States and the European Union. Twenty-four member nations are signatories to the OECD Council Recommendations Concerning Cooperation on Restrictive Business Practices (1986). The UNCTAD Set of Multilaterally Agreed Equitable Principles and Rules for the Control of Restrictive Business Practices (1980) is another multilateral agreement. But these arrangements are designed to facilitate cooperation and information sharing. They do not involve harmonizing or developing common principles for administering and enforcing competition law.

Should such measures be developed and expanded to cover the larger world trading community? Sir Leon Brittan, former vice president of the European Commission, believes so. He writes:

The next GATT Round should include restrictive business practices and cartels on its agenda. The aim should be to draw up common rules, lay down the principle that restrictive arrangements are not enforceable at law and that governments are responsible internationally for the implementation of these rules and procedures. . . . For mergers, common rules should also be established, as well as a common commitment to enforce them. (1992b, p. 108)

Underlying this viewpoint is the desire to promote consumer welfare by ensuring market access. In an address to the World Economic Forum, Sir Leon Brittan (1992a) stated that we must “ensure that the benefits of liberalization are more fully passed on to the people who really count, the citizen/consumer.”

Others share this vision. In 1993, for example, an international group of lawyers under the auspices of the Max Planck Institute published a draft international antitrust code.<sup>52</sup> Yet there is considerable skepticism that a world consensus could be reached on competition or antitrust policy.<sup>53</sup> The very diversity of factors causing friction in trade and competition matters makes it difficult to agree on a single solution.

One suggestion that has been made is to strengthen GATT rules aimed at enhancing market access and competition and to then have GATT signatories enforce them more aggressively (see, for example, Hoekman and Mavroidis 1993 and Davidow 1994). But the GATT deals only with government policies. Anticompetitive business practices and arrangements by private entities are not directly subject to its rulings. Moreover, many market restrictions, cartels, and discriminatory practices, particularly those prevalent in developing and emerging market economies, are permitted or supported by government.

While the general objectives of the GATT and competition law are congruent, liberalized trade alone cannot ensure competition. A poorly designed and implemented competition law, or its absence, can be a barrier to entry. Problems arising from nontraded goods and services, restrictive distribution systems, transborder market-sharing agreements and cartels, and the use of antidumping actions can also be impediments to trade.

A reasonable agenda might be to foster the harmonization of competition law and policy and to establish a

common code among regional trade blocs. With the exception of the ASEAN region and the European Union, trade blocs have been accounting for a decreasing share of world exports while intraregional exports have been increasing (see de Melo and Panagariya 1992, especially pages 16–17). Foreign direct investment has also exhibited an increasing regional focus. The principal areas that need to be addressed at the regional level range from adopting competition law where none exists<sup>54</sup> to amending laws to strengthen them to effectively apply existing provisions against collusion, vertical restraints, and excessive mergers and acquisitions.<sup>55</sup> Reducing the number of areas subject to exemptions would also be a worthy goal. Even within this limited multilateral context, these tasks represent major challenges for international cooperation and policy formulation.

### Conclusion

This overview of competition policy has touched on a number of issues that developing and emerging market economies should weigh heavily as they strive to promote industrial growth. Although many of these countries have liberalized trade and investment policies as part of a program of market-oriented economic reforms, less consensus exists regarding the value and proper role of competition policy. We hold that a well-drafted competition law is an important policy measure that governments should undertake in order to strengthen market forces and that an efficiency-oriented competition law is advisable. We conclude that high levels of industrial concentration are not necessarily inimical to competition. Rather, the policy emphasis should be on business conduct as opposed to firm size. Empirical support for policies designed to promote competitiveness by limiting competition and fostering large firms, is scant. Moreover, governments have not proved to be better able to pick “winners” than the market. Generally, where governments have intervened in this manner, the costs have been high. Developing and emerging market economies that already have large fiscal deficits can ill afford such an approach toward promoting industrial growth. Those who point to the “East Asian miracle” as an example of a successful government-led industrial policy do not adequately recognize the critical role that domestic and international rivalry has played in fostering superior economic performance.

Although trade liberalization promotes competition in domestic markets, it cannot ensure it. We have identified various impediments that can dampen the pro-competition effects of import competition. Indeed, the absence of an effective competition policy can itself act as a bar-

rier to trade. Trade liberalization and competition policy play complementary roles in strengthening market forces and ensuring that their benefits flow to consumers.

We examine and reject the view that the administration and enforcement of competition law itself must inevitably become a source of intervention in the market, corruption, misuse of bureaucratic power, or cause of market distortions. All of these risks can be dealt with through institutions that incorporate accountability, transparency, checks and balances, and clear rules and procedures. The design and implementation of competition law, and the mix of policy instruments and enforcement priorities must, however, reflect the institutional endowments and technical capacity of countries at different stages of economic development.

Finally, we note that increases in foreign investment and international trade heighten the possibility of friction between competition policies in different jurisdictions. A number of industrial organization economists and lawyers have recently called for the development of an international code to resolve problems that may arise between countries. We share the view that developing a multilateral code in this area may be difficult. We suggest that the alternative of developing a competition policy code within regional trading blocs such as the ANDEAN and ASEAN countries may be a more appropriate starting point—especially in light of increased regional trade and regional economic integration.

### Notes

1. For a discussion of institutional aspects and country experience, see Boner in this volume on economies in transition and Boner and Krueger (1992). A useful discussion of implementation issues can be found in Pittman (1992).

2. We use the term “emerging market economies” as being synonymous with “transitional economies” to refer to former centrally planned countries (for example, Poland and Russia). In these economies virtually all production and consumption was state-directed and there were no markets in the conventional sense. That is not the case for “developing or newly industrializing” economies (for example, India, Philippines, and several countries in Latin America and Africa), where, despite the prevalence of a large public sector, a market-oriented private sector also exists—one that governs daily purchasing decisions by ordinary consumers.

3. The concept of market power is central to the administration and enforcement of competition law and policy. It refers to the ability of a firm, unilaterally or in collusion with others, to profitably raise price and lower output over a significant period without competitive response from other existing firms or potential entrants into the market.

4. See Frischtak, Hadjimichael, and Zachav (1989) and Geroski and Jacquemin (1985) for arguments for reduced barriers and increased mobility of resources.
5. Competition advocacy has more recently evolved to become part of the role of competition agencies. During the 1970s, the U.S. Antitrust Division and Federal Trade Commission actively fostered and participated in the deregulation of several key economic sectors: transportation, telecommunications, and energy. Similar initiatives were soon taken by the Canadian, German, and U.K. competition authorities. A unique feature of the Canadian competition law is that it gives officials of the Bureau of Competition Policy a formal statutory right to appear before tribunals, commissions, and committees to put forward pro-competition views in the search for policy solutions that interfere least with market processes. Competition advocacy has also been incorporated in the role of competition authorities in several developing and emerging market economies, including Colombia, Hungary, and Poland.
6. See also Kühn and others (1992) for an assessment of the theoretical literature and unresolved issues.
7. This section draws on Khemani (1993).
8. See Khemani (1993) and related proceedings of the Round-Table Discussion of the Competition Law and Policy Committee, OECD.
9. See Audretsch (1985) for a fuller discussion.
10. The emphasis on the role of market structure was developed primarily by the Harvard economist Edward Mason during the 1930s and subsequently formalized by Bain (1968) and Kaysen and Turner (1959). Today, few (if any) structuralist economists would regard market structure as an exclusively exogenous and sufficient determinant of firms' ability to earn excess profits. There is now greater focus on advertising, research and development, contractual arrangements, preemption of input sources and distribution channels, and other types of pricing and output policies that may affect market structure while aiming at excluding competition.
11. Baumol, Panzar, and Willig (1982), who are not associated with the Chicago school, also emphasize the role that barriers to entry play in limiting competition. They suggest that reducing barriers makes markets "contestable" and that an industry consisting of one firm or a few firms may be efficient.
12. These views have been particularly espoused by Thurow (1980a and 1992).
13. Galbraith is also critical of the adverse impact of competition laws; see Galbraith (1967).
14. These views have recently come to the forefront of U.S. policy. See Krugman (1994) for a critical analysis of some of the underlying arguments.
15. See Williamson (1968).
16. See Scherer and Ross (1990), p. 415n; also see the survey by L. W. Weiss, "The Concentration-Profit Relationship and Antitrust," in Goldschmid, Michael-Mann, and Weston (1974).
17. Goldschmid, Michael-Mann, and Weston (1974) provides a good synthesis of the relevant points of issue.
18. Bain (1956), for example, conducted research on the role of such barriers as scale economies, product differentiation, and the absolute cost advantages of firms.
19. The ensuing discussion draws on World Bank (1993). For differing views, see Fishlow and others (1994), Singh (1994) and Lall (1994). These views are consistent with the need to avoid the suppression of competition and to ensure interfirm rivalry.
20. These are administered by highly trusted, skilled, and technically competent public service administrators with low incidence of corruption.
21. Collusion formalized in an explicit agreement is generally referred to as a cartel.
22. In a recent case, Sumitomo Chemical Co. Ltd. of Japan and Chemagro parent Bayer AG of Germany were fined more than US\$900,000 each for participating in a conspiracy to divide market share for chemical insecticides in Canada. The Canadian Bureau of Competition Policy's investigation was initiated after a voluntary disclosure by a third party involved, which took advantage of a recently introduced Canadian policy of offering immunity for cooperation in uncovering serious anticompetitive behavior. See Antitrust and Trade Regulation Report 65 (November 25, 1993), p. 691. Washington, D.C.: Bureau of National Affairs. It is not known how common such agreements are in developing and emerging market economies, but it is safe to say that the increasing internationalization of corporations makes them more likely.
23. An example of rule-of-thumb pricing is basing-point pricing, in which a mill price is established and customers at different destinations are charged the announced mill price plus freight. Such practices aid coordination by restricting attention to a single price and reduce informational delays, thus hastening retaliation. A colorful example of collusion is a bid-rigging conspiracy in the electrical equipment industry (high voltage switchgears). Firms coordinated bids on government procurement contracts by agreeing on which firm would be the lowest bidder (and rotating so that each firm would win an agreed upon number or value of contracts). The firms used the phases of the moon to determine which would submit the low bid. See Scherer and Ross (1990), chapters 7 and 8.
24. The next round of GATT World Trade Organization negotiations is, however, expected to focus on competition policy. See the section below on the international dimensions of competition policies.
25. There is also the concern, from a global welfare point of view, that export cartels by one country may evoke retaliation, with everyone better off if no one uses them.
26. For a review of arguments for production cartels in the context of industrial restructuring, see Atiyas, Dutz, and Frischtak (1992), especially pp. 28-32.

27. For overviews of the theoretical literature on vertical restraints, see Kühn and others (1992) and Katz (1989).
28. For a general discussion of circumstances in which such strategies allow increased exercise of market power, see Krattenmaker and Salop (1986).
29. This recommendation is one of the main findings of a recent study on franchising undertaken by the OECD (1993). The Russian competition law (Law on Competition and Limiting Monopolistic Activities in Commodity Markets, of March 22, 1991, amended by Law of July 15, 1992) imposes exactly this sort of market power “screen” on vertical agreements [Article 6 (2)], while permitting an efficiency defense in exceptional cases [Article 6 (3)].
30. In particular, dominant manufacturer contracts that restrict distributors to distributing a small volume or none of the products of competing manufacturers should be viewed with extreme suspicion by competition agencies. An illustrative example is the Borsod Brewery case in Hungary, in which a regional beer monopolist facing entry forced its distributors to agree that 95 percent of the beer they distributed would be that of the monopolist.
31. See Dutz and Suthiwart-Narueput in this volume.
32. The term “abuse of dominant” market position has been explicitly incorporated in competition legislation in several economies, including Canada, the European Union, Germany, and several formerly centrally planned economies. The counterpart provisions/concepts in the United States are those dealing with monopoly and attempts to monopolize a market.
33. In Germany and the United Kingdom a market share of 33 percent is considered dominance, and in Australia, a 60 percent share. In many other jurisdictions a firm is considered dominant when it accounts for 40 percent or more of the relevant market. But establishing the market share of a dominant firm is only the starting point of a competition case. The presence of other structural and conduct factors—such as the conditions of entry that firms face and the pricing-out policies of the dominant firm—have to be examined. As the phrase “abuse of dominant market position” implies, dominance itself is not illegal. It is the abuse of market position by a dominant firm that may call for antitrust action.
34. For an overview, see American Bar Association (1986).
35. The economic literature on the rationality and effectiveness of predatory pricing is in a state of flux. Although the practice can be as costly to the predator as to the victim, predators may nevertheless adopt the practice to “soften” up rivals for future acquisition or to send signals to existing and potential competitors to condition their business behavior. See OECD (1989) for a summary of the issues, concepts, and operational approaches.
36. See, for example, Milgrom and Roberts (1982), Kreps and Spence (1984), Kreps and Wilson (1982), and Selten (1978).
37. Schmidt (1983) compares German, European, and U.S. policy toward market-dominating enterprises.
38. This was the view held by the court in the monopolization case of *United States v. Aluminum Co. of America (ALCOA)* 148 F.2d 416 (2d CH 1945).
39. This has often been referred to as the “cellophane trap” arising from *United States v. E. J. DuPont de Nemours and Co.*, 351 U.S. 377 (1956). The court failed to accept the premise that DuPont had market power in the “flexible wrapping material” market at prevailing prices (which were high) because of available (though inferior) substitutes, such as waxed paper. See also American Bar Association (1986) and White (1987).
40. Cowling, Stoneman, and Cubbin (1980) have argued that there is a potential inconsistency in the application of competition policy to mergers and dominant firm monopoly, if this “tradeoff” approach is adopted. In effect, mergers would be judged in terms of the direction of change in welfare (whether it is raised or lowered) while monopolies would be judged in terms of the level of social welfare relative to the competitive alternative. In other words, even though a merger may improve social welfare compared with the pre-merger situation, the outcome is still worse than the situation that would prevail under competition. The tradeoff approach is viewed as a less stringent approach to mergers. But it should be noted that in jurisdictions where this approach has been adopted, such as Canada and the United States, it is not sufficient to demonstrate that a merger would result in efficiencies that augment consumer or total welfare. It must also be demonstrated that these efficiencies cannot be attained through other means, such as competition, joint ventures, and licensing.
41. For a general critique of competition law enforcement and administration, see Godek (1992).
42. Commenting on competition policies inaugurated in Eastern Europe and other emerging market economies, the American Bar Association recently noted its “concern that the law might be applied too intrusively. The abuse of dominance law could, if applied unwisely, effectively restore price control under the guise of antitrust.” See American Bar Association (1992).
43. A private interest theory of competition law, suggesting that U.S. antitrust enforcement was seldom in the public interest and often used to protect particular competitors at the expense of competition and efficiency, has been advanced in Shughart (1990). In the U.S. context, it has been claimed that “the Federal Trade Commission investigations are undertaken at the behest of corporations, trade associations and trade unions whose motivation is at best to shift the cost of their private litigation to the taxpayer and at worst to harass competitors.” See Posner (1969).
44. These principles are designed to provide a system of checks and balances, by ensuring due process of law with provisions for appeals and additional review of case facts and decisions.
45. See, for example, Esposito and Esposito (1971); also see de Melo and Urata (1986), which reports these findings for Chile. Similar results have been obtained by researchers for Canada, West Germany, and the United Kingdom, among others.

46. On this last point, see Globerman (1990). See also the empirical work by Leamer (1988) and the review by Fishlow (1990).

47. See Finger (1993) and Low and Subramanian (1993). Non-OECD countries that have recently introduced or reactivated antidumping or countervailing duty statutes include Argentina, Bolivia, Brazil, Chile, China, Colombia, Egypt, India, Indonesia, Israel, Jamaica, Malaysia, Mexico, Morocco, Peru, the Philippines, South Africa, Thailand, Trinidad and Tobago, and Venezuela. Two main motivations appear to be responsible for this trend. First, as governments lower tariffs and nontariff barriers, they come under increasing pressure from adversely affected import-competing domestic producers. Antidumping duties are a convenient substitute mechanism for providing protection to firms that claim injury. Second, with increased trade liberalization, many countries joined the GATT for the first time in the 1980s. Following the example of such economies as the United States and the European Union, these countries adopted the full arsenal of provisions permitted by the Antidumping Code to ensure that they respond to similar protective trade practices by other governments and that domestic firms have a set of protective remedies similar to those of their foreign counterparts. Many antidumping cases by developing countries appear to have been brought at least in part as retaliation—counterpressure in support of domestic exporters facing antidumping cases in other countries.

48. The General Agreement on Trade in Services (GATS), a part of the Uruguay Round, would help in this regard by dismantling “within border” regulatory barriers. Almost a quarter of world trade is service-related. See Broadman (1994).

49. This point has been made in the context of Turkey in Dutz (forthcoming). Whether or not Turkey joins the European Union, an EU-harmonized competition law rather than one distinctly Turkish would both help Turkish exporters adapt to EU practices and encourage European firms to trade with and invest in Turkey.

50. See Whish and Wood (1993). The study reviews eight other case examples involving such firms as Westinghouse Electric-ABB, Matsushita-MCA, Renault-Volvo, and Fiat-Ford.

51. See UNCTAD (1993) and citations therein for limited information on Africa, East Asia, and Latin America. During the spring of 1993 the Russian State Antimonopoly Committee and several countries of the former Soviet Union entered into a memorandum of understanding on mutual assistance and cooperation on inter-jurisdictional antimonopoly problems (press release, Moscow, May 7, 1993).

52. See Antitrust and Trade Regulation Report special supplement (August 19, 1993).

53. See interview with U.S. Assistant Attorney General for Antitrust Ann Bingaman, reported in Antitrust and Trade Regulation Report (Fall 1993). Washington, D.C.: Bureau of National Affairs. Similar views have been expressed by James Rill, Ms. Bingaman’s predecessor, in meetings of the OECD

Committee on Competition Law and Policy, and by Davidow (1994).

54. For example, in the ANDEAN region, Bolivia and Ecuador have no competition law, nor do many of the ASEAN countries.

55. See Antitrust and Trade Regulation Report (1993) for more details on these and other areas suggested by the Max Planck Institute; also see Fox (1994).

## References

- Abreu, D., D. Pearce, and E. Stachetti. 1990. “Toward a Theory of Discounted Repeated Games with Imperfect Monitoring.” *Econometrica* 58: 1041–63.
- Aghion, P., and P. Bolton. 1987. “Contracts as Barriers to Entry.” *American Economic Review* 70: 388–401.
- American Bar Association. 1986. *Horizontal Mergers: Law and Policy*. Monograph 12. Antitrust Law Section. New York.
- . 1992. “Introduction and Recommendations of ABA Antitrust Law Section’s Special Committee on International Antitrust.” *Antitrust and Trade Regulation Report* 62: 171.
- Anderson, R., and S. D. Khosla. 1994. “Competition Policy as a Dimension of Economic Policy: A Comparative Perspective and Agenda for the Future.” Discussion paper. Bureau of Competition Policy, Hull, Quebec.
- Antitrust and Trade Regulation Report*. 1993. “Draft International Antitrust Code as a GATT-MTO-Plurilateral Trade Agreement.” Special supplement. Max Planck Institute, Munich and Washington, D.C.
- Atiyas, I., M. Dutz, and C. Frischtak. 1992. “Fundamental Issues and Policy Approaches in Industrial Restructuring.” Industry and Energy Department Working Paper, Industry Series 56. World Bank, Washington, D.C.
- Audretsch, D. 1985. “The Four Schools of Thought in Antitrust Economics.” Working paper. International Institute of Management, Berlin.
- Bain, J. S. 1956. *Barriers to New Competition*. Cambridge, Mass: Harvard University Press.
- . 1968. *Industrial Organization*. New York: John Wiley.
- Baumol, W., J. Panzar, and R. Willig. 1982. *Contestable Markets and the Theory of Industry Structure*. San Diego: Harcourt Brace Jovanovich.
- Bhagwati, J. 1965. “On the Equivalence of Tariffs and Quotas.” In R. Baldwin, ed., *Trade, Growth and the Balance of Payments*. Chicago: Rand McNally.
- Boner, R., and R. Krueger. 1992. *The Basics of Antitrust Policy*. World Bank Technical Paper 160. Washington, D.C.
- Bork, R. 1978. *The Antitrust Paradox: A Policy at War with Itself*. New York: Basic Books.
- Bound, J., C. Cummius, Z. Griliches, B. H. Hall, and A. Jaffe. 1984. “Who Does Research and Development and Who Patents.” In Zvi Griliches, ed., *Research and Development*,

- Patents and Productivity*. National Bureau of Economic Research Conference Report. Chicago: University of Chicago Press.
- Brennan, T. 1988. "Understanding 'Raising Rivals' Costs." *Antitrust Bulletin* 2: 95–113.
- Brittan, Sir Leon. 1992a. "A Framework for International Competition." Address to World Economic Forum. Davos, Switzerland. February 3.
- . 1992b. *European Competition Policy: Keep the Playing Field Level*. Brussels: Centre for European Policy Studies.
- Broadman, H. G. 1994. "GATT: The Uruguay Round Accord on International Trade and Investment in Services." *World Economy* 17: 281–92.
- Carlton, D., and J. Perloff. 1989. *Modern Industrial Organization*. Glenview, Ill.: Scott Foresman/Little Brown.
- Caves, R. E. 1992. *Industrial Efficiency in Six Nations*. Cambridge, Mass.: MIT Press.
- Caves, R. E., and M. E. Porter. 1977. "From Entry Barriers to Mobility Barriers: Conjectural Decisions and Contrived Deterrence to New Competition." *Quarterly Journal of Economics* 97: 247–21.
- Caves, R. E., M. E. Porter, A. M. Spence, and J. T. Scott. 1980. *Competition in the Open Economy: A Model Applied to Canada*. Cambridge, Mass.: Harvard University Press.
- Comanor, W., and P. Rey. Forthcoming. "Competition Policy towards Vertical Restraints in a Global Economy." In *Competition in a Global Economy*. Center for International Studies, Toronto.
- Conrath, C. W., and B. T. Freeman. 1994. "A Response to the Effectiveness of Proposed Antitrust Programs for Developing Countries." *North Carolina Journal of International Law and Commercial Regulation* 19: 233–45.
- Cowling, K., P. Stoneman, and J. Cubbin. 1980. *Mergers and Economic Performance*. Cambridge: Cambridge University Press.
- Davidow, J. 1994. "The Whys and Wherefores of GATT Competition Rules." Meeting on Does an International Antitrust Code Belong in the OECD or GATT, District of Columbia Bar, International Law Section, Washington, D.C., February 4.
- de Melo, J., and A. Panagariya. 1992. "The New Regionalism in Trade Policy." An Interpretive Summary of a Conference. Centre for Economic Policy Research, London, and Trade Policy Division, Country Economics Department, World Bank, Washington, D.C.
- de Melo, J., and S. Urata. 1986. "The Influence of Increased Foreign Competition on Industrial Concentration and Profitability." *International Journal of Industrial Organization* 4: 287–304.
- De Prano, M. E., and J. R. Nugent. 1969. "Economies as an Antitrust Defense: A Comment." *American Economic Review* 59: 947–53.
- Dutz, M. A. 1993. "Enforcement of Canadian Trade Remedy Laws: The Case for Competition as an Antidote for Protection." In J. M. Finger, ed., *Antidumping: How It Works and Who Gets Hurt*. Ann Arbor: University of Michigan Press.
- . Forthcoming. "Competition Law and Its Relevance for Turkey." In R. Erzan, ed., *Competition Policies for Turkey*. New York: Macmillan.
- Eastman, H. C., and S. Stykolt. 1960. "A Model for the Study of Protected Oligopolies." *Economic Journal* 70: 336–47.
- . 1967. *The Tariff and Competition in Canada*. Toronto: Macmillan.
- Elizinga, K. G. 1985. "New Developments on the Cartel Front." *Antitrust Bulletin* 29: 3–26.
- Esposito, L., and F. Esposito. 1971. "Foreign Competition and Domestic Industry Profitability." *Review of Economics and Statistics* 53: 343–53.
- Finger, J. Michael, ed. 1993. *Antidumping: How It Works and Who Gets Hurt*. Ann Arbor: University of Michigan Press.
- Fisher, F. M. 1979. "Diagnosing Monopoly." *Quarterly Review of Economics and Business* 19: 7–33.
- . 1989. "Games Economists Play: A Non-cooperative View." *RAND Journal of Economics* 20: 113–24.
- Fishlow, A. 1990. "The Latin American State." *Journal of Economic Perspectives* 4: 61–74.
- Fishlow, A., C. Gwin, S. Haggard, D. Rodrik, and R. Wade. 1994. *Miracle or Design? Lessons from the East Asian Experience*. Washington, D.C.: Overseas Development Council.
- Fox, E. 1994. "Antitrust and the Next GATT Agenda." Comparative Competition and Trade Policy Project. School of Law, New York University, New York.
- Frischtak, Claudio, Bitá Hadjimichael, and Ulrich Zachav. 1989. "Competition Policies for Industrializing Countries." Policy, Planning, and Research Series 7. World Bank, Washington, D.C.
- Fudenberg, D., and J. Tirole. 1986. *Dynamic Models of Oligopoly*. New York: Harwood Academic Publishers.
- Galbraith, J. K. 1967. *The New Industrial State*. Boston: Houghton Mifflin.
- Geroski, P., and A. Jacquemin. 1985. "Industrial Change, Barriers to Mobility and European Industrial Policy." *Economic Policy* 1: 169–218.
- Gilbert, R. 1986. "Preemptive Competition." In J. Stiglitz and F. Mathewson, eds., *New Developments in the Analysis of Market Structure*. Cambridge, Mass.: MIT Press.
- . 1989. "Mobility Barriers and the Value of Incumbency." In R. Schmalensee and R. Willig, eds., *The Handbook of Industrial Organization*. Amsterdam: North Holland.

- Globerman, S. 1990. "Trade Liberalization and Competitive Behavior: A Note Assessing the Evidence and the Public Policy Implications." *Journal of Policy Analysis and Management* 9:80-88.
- Godek, P. E. 1992. "One U.S. Export Eastern Europe Doesn't Need." *Regulation* 20:19-22.
- Goldschmid, H. J., H. Michael-Mann, and J. F. Weston. 1974. *Industrial Concentration: The New Learning*. Boston: Little Brown.
- Gorecki, P., and W. T. Stanbury. 1984. *The Objectives of Canadian Competition Policy 1888-1983*. Halifax: Institute for Research on Public Policy.
- Green, J., and R. Porter. 1984. "Noncooperative Collusion under Imperfect Price Information." *Econometrica* 52: 87-100.
- Hay, G. A., and D. Kelley. 1974. "An Empirical Survey of Price-Fixing Conspiracies." *Journal of Law and Economics* 7: 13-38.
- Hazeldine, T. 1991. "Trade Policy as Competition Policy." In R.S. Khemani and W. T. Stanbury, eds. *Canadian Competition Law and Policy at the Centenary*. Halifax: Institute for Research on Public Policy.
- Hoekman, B. M., and P. C. Mavroidis. 1993. "Competition, Competition Policy and the GATT." Policy Research Working Paper 1228. World Bank, Washington, D.C.
- Jackson, R. 1970. "The Consideration of Economies in Merger Cases." *Journal of Business* 43: 439-46.
- Jacquemin, A., and M. Slade. 1988. "Cartels, Collusion and Horizontal Merger." In R. Schmalensee and R. Willig, eds., *The Handbook of Industrial Organization*. Amsterdam: North-Holland.
- Katz, M. 1989. "Vertical Contractual Relations." In R. Schmalensee and R. Willig, eds., *The Handbook of Industrial Organization*. Amsterdam: North-Holland.
- Kaysen, C., and D. Turner. 1959. *Antitrust*. Cambridge, Mass.: Harvard University Press.
- Khemani, R. S. 1993. "Objectives of Competition Policy." OECD, Paris. DAFPE/CLP (92) 2/Rev 1.
- Khemani, R. S., and W. T. Stanbury, eds. 1991. *Canadian Competition Law and Policy at the Centenary*. Halifax: Institute for Research on Public Policy.
- Krattenmaker, T., and S. Salop. 1986. "Anticompetitive Exclusion: Raising Rivals' Costs to Achieve Power over Price." *Yale Law Journal* 96: 209-95.
- Kreps, D., and M. Spence. 1984. "Modeling the Role of History in Industrial Organization and Competition." In G. Fewel, ed., *Contemporary Issues in Modern Microeconomics*. London: Macmillan.
- Kreps, D., and R. Wilson. 1982. "Reputation and Imperfect Information." *Journal of Economic Theory* 27: 253-79.
- Krugman, P. 1994. "Competitiveness: A Dangerous Obsession." *Foreign Affairs* 73: 28-44.
- Kühn, K. U., and others. 1992. "Competition Policy Research: Where Do We Stand?" Occasional Paper 8. Centre for Economic Policy Research, London.
- Kwoka, J. E. 1979. "The Effect of Market Share Distribution on Industry Performance." *Review of Economics and Statistics* 61:183-89.
- Kwoka, J. E., and D. J. Ravenscraft. 1986. "Cooperation vs. Rivalry: Price-Cost Margins by Line of Business." *Economica* 53: 351-63.
- Lall, S. 1994. "Industrial Policy: The Role of Government in Promoting Industrial and Technological Development." *UNCTAD Review/United Nations Conference on Trade and Development* 65-89. Geneva.
- Leamer, E. E. 1988. "Cross-Section Estimates of the Effects of Trade Barriers." In R. C. Feenstra, ed., *Empirical Methods for International Trade*. Cambridge, Mass.: MIT Press.
- Leibenstein, H. 1966. "Allocative Efficiency vs. 'X-Efficiency.'" *American Economic Review* 56: 392-415.
- Low, P., and A. Subramanian. 1993. "Trade Protection in Agriculture: A Special Case?" Mimeo. International Economics Department, World Bank, Washington, D.C.
- Milgrom, P., and J. Roberts. 1982. "Predation, Reputation and Entry Deterrence." *Journal of Economic Theory* 27: 280-312.
- Ordovery, J., and D. Wall. 1988. "Proving Entry Barriers: A Practical Guide to the Economics of New Entry." *Antitrust* 2: 12-17.
- OECD (Organization for Economic Cooperation and Development). 1989. *Predatory Pricing*. Paris.
- . 1993. *Competition Policy and Franchising*. Paris.
- Ostry, S. 1990. *Governments and Corporations in a Shrinking World*. New York: Council on Foreign Relations.
- . Forthcoming. "Globalization, Domestic Policies and the Need for Harmonization." In *Competition in a Global Economy*. Center for International Studies, Toronto.
- Perry, M., and R. Porter. 1985. "Oligopoly and the Incentive for Horizontal Merger." *American Economic Review* 75: 219-27.
- Pittman, Russell. 1992. "Some Critical Provisions in the Antimonopoly Laws of Central and Eastern Europe." *International Lawyer* 26: 485-503.
- Porter, M. E. 1990. *The Competitive Advantage of Nations*. New York: The Free Press.
- Posner, R. A. 1969. "The Federal Trade Commission." *University of Chicago Law Review* 37:47-89.
- . 1976. *Antitrust Law: An Economic Perspective*. Chicago: University of Chicago Press.
- . 1981. "The Chicago School of Antitrust Analysis." *Corporate Practice Commentator* 22: 583-610.
- Ravenscraft, D., and F. M. Scherer. 1987. *Mergers, Sell-offs and Economic Efficiency*. Washington, D.C.: Brookings Institution.
- Rey, P., and J. E. Stiglitz. 1988. "Vertical Restraints and Producers Competition." *European Economic Review* 32: 561-68.

- Rodriguez, A. E., and M. D. Williams. 1994. "The Effectiveness of Proposed Antitrust Programs for Developing Countries." *North Carolina Journal of International Law and Commercial Regulation* 19: 209–32.
- Ross, T. W. 1988a. "Movements towards Free Trade and Domestic Market Performance with Imperfect Competition." *Canadian Journal of Economics* 21: 507–24.
- . 1988b. "On the Price Effects of Mergers with Free Trade." *International Journal of Industrial Organization* 6: 233–46.
- Safarian, A. E. 1993. "Foreign Direct Investment and International Cooperative Agreements: Trends and Issues." Center for International Studies, Toronto.
- Salop, S. 1979. "Strategic Entry Deterrence." *American Economic Review* 69: 335–38.
- Scherer, F. M. 1965. "Firm Size, Market Structure, Opportunity and the Output of Patented Inventions." *American Economic Review* 55: 1097–1125.
- . 1977. "The Posnerian Harvest: Separating Wheat from Chaff." *Yale Law Review* 86: 974–1002.
- . 1980. *Industrial Market Structure and Economic Performance*. 2d ed. Chicago: Rand McNally College Pub. Co.
- Scherer, F. M., and David Ross. 1990. *Industrial Market Structure and Economic Performance*. 3d ed. Boston: Houghton Mifflin.
- Scherer, F. M., and others. 1975. *The Economics of Multiplant Operation*. Cambridge, Mass.: Harvard University Press.
- Schmalensee, R., and R. Willig, eds. 1989. *The Handbook of Industrial Organization*. Amsterdam: North Holland.
- Schmidt, I. 1983. "Different Approaches and Problems in Dealing with Control of Market Power: A Comparison of German, European and U.S. Policy towards Market-Dominating Enterprises." *Antitrust Bulletin* 28: 417–60.
- Schumpeter, J. A. 1942. *Capitalism, Socialism and Democracy*. New York: Harper Press.
- Selten, R. 1978. "The Chain Store Paradox." *Theory and Decision* 9: 127–59.
- Shepherd, W. 1984. "Contestability vs. Competition." *American Economic Review* 4: 572–87.
- Shugart, W. F., II. 1990. *Antitrust Policy and Interest-Group Politics*. Westport, Conn.: Quorum Books.
- Singh, A. 1994. "Growing Independent of the World Economy: Asian Economic Development since 1980." *UNCTAD Review/United Nations Conference on Trade and Development* 91–103. Geneva.
- Steiner, P. O. 1975. *Mergers: Motives, Effects and Policies*. Ann Arbor: University of Michigan Press.
- Stigler, G. 1968. "A Theory of Oligopoly." In George J. Stigler, ed., *The Organization of Industry*. Housewood, Ill.: Richard D. Irwin.
- Thurrow, L. C. 1980a. "Let's Abolish the Antitrust Laws." *The New York Times*, October 19.
- . 1980b. *The Zero Sum Society*. New York: Basic Books.
- . 1992. *Head to Head*. New York: W. Morrow & Co.
- UNCTAD. 1993. *Concentration of Market Power and Its Effects on International Markets*. Geneva and New York.
- Whish, R. P., and D. P. Wood. 1993. "OECD Merger Process Convergence Project." OECD, Paris.
- White, L. J. 1987. "Antitrust and Merger Policy: A Review and Critique." *Journal of Economic Perspectives* 1: 13–22.
- Williamson, O. E. 1968. "Economies as an Antitrust Defense." *American Economic Review* 58: 18–36.
- . 1977. "Economies as an Antitrust Defense Revisited." *University of Pennsylvania Law Review* 125: 699–736.
- . 1984. "Credible Commitments." *Antitrust Bulletin* 29: 33–76.
- World Bank. 1993. *The East Asian Miracle*. New York: Oxford University Press.

# Competition policy and institutions in reforming economies

Roger Alan Boner

Competition policy, long a feature of developed market economies, in recent years has been adapted to numerous reforming economies.<sup>1</sup> Competition policy encourages efficiency by creating and preserving the competitive process.<sup>2</sup> In practice, it strives to ensure that competition between private parties is not unduly impeded by state or private actions. Frequently, private suppliers react to reforms by taking steps to avoid competition.<sup>3</sup> In addition, governmental bodies occasionally impede competition by imposing unnecessarily restrictive regulations. The focus of this chapter is on the implementation and enforcement of competition policy in reforming economies.

Competition policy, often referred to as “antitrust,” consists of competition *law* and competition *advocacy*. Competition law and advocacy are designed to correct market failures resulting from, respectively, private and regulatory impediments to competition. Competition law consists of an enforceable legal code applying to commercial tactics and transactions involving private enterprises. This code usually prohibits commercial conduct that would conflict with specific national goals (such as efficiency or the preservation of economic freedom).

In contrast, competition advocacy refers to public analysis and comment by a competition agency regarding the competitive effects of laws, regulations, and other actions of state bodies. Such advocacy provides an institutional means of exposing and perhaps correcting the competitive harm caused by unnecessarily restrictive regulations. In the developed economies, competition advocacy is usually an advisory device; that is, a competition agency, acting as advocate, provides regulatory analysis and recommendations, but does not compel.<sup>4</sup> In contrast, competition advocacy in certain reforming economies (for example, Hungary and Kazakhstan) operates with the force of law.<sup>5</sup>

## Tailoring competition reform to a national economy

The competition laws of certain industrial economies, particularly Germany, Japan, the United States, and the European Union (EU), have often served as models for reforming economies. The new competition laws in Eastern Europe, for example, are based largely on the laws of Germany and the European Union, whereas Argentina, Jamaica, New Zealand, the Philippines, and Venezuela have used the U.S. and the EU statutes.<sup>6</sup> In addition, the legal standards of industrial nations are sometimes prescribed for reforming economies.<sup>7</sup> This prescription can be sound only to the degree that a specific legal standard (for example, an unconditional, per se prohibition of price fixing) would be effective and efficient in different national commercial environments.<sup>8</sup>

The competition laws of reforming economies are seldom a direct translation of legislation in the industrial countries.<sup>9</sup> Not only do economic goals differ greatly, but so does the effectiveness of national institutions. Broadly speaking, the legal and enforcement standards of competition law should be designed so that one mechanism (the law) is applied only where it improves on an alternative mechanism (the market).<sup>10</sup> That is, a court order or enforcement decree should be imposed only if it produces a benefit—such as an increase in efficiency or consumer welfare or an improvement in private-sector development—that would not occur otherwise.<sup>11</sup>

The relative effectiveness of markets and legal or regulatory systems varies across nations, which correspond to differences in the legal and enforcement standards of competition law. In addition, reforming economies (almost by definition) operate in a commercial, legal, and regulatory environment characterized by rapid change; the “optimal” (that is, efficient) set of legal standards governing commercial conduct must change in response. For this reason, devising a new competition law for a reforming economy is not merely a matter of writing sensible

legal standards. Instead, it is more important for new competition law to provide an enforcement process from which efficient legal standards can be expected to emerge.

Some scholars fear that the enforcers of embryonic competition laws will make mistakes by developing nonsensical or counterproductive legal standards regarding competition. Thus, they recommend against using competition law to promote economic development.<sup>12</sup> Judging from the enforcement histories of developed economies, such mistakes are a natural and unavoidable feature of competition law. However, the important question is not whether enforcers will make mistakes but rather, whether they will learn from and correct these mistakes. This form of adaptability flows from the structural aspects of enforcement mechanisms: Does the mechanism make use of precedent, the private right of action, and the availability of independent review? These devices decentralize and stabilize the enforcement of a new competition law and ensure that enforcement is responsive to a changing commercial environment.

The nature and efficiency of the enforcement process result largely from the institutional arrangements for resolving legal disputes; these arrangements strongly influence the development and flexibility of legal standards of commercial conduct. In Argentina, Brazil, Mexico, and the Philippines, seemingly reasonable competition laws have not been strongly enforced. Price fixing, although ostensibly illegal, is observed in each country. The failure to enforce has resulted more from institutional deficiencies than from ill-designed competition statutes, and recent efforts to reform the law have focused on strengthening enforcement mechanisms.

### **Basic concepts: market power and dominance**

The enforcement and commercial effect of a competition law depend greatly on the goals and concepts on which the law is based. Competition laws in the developed economies have been based primarily on two concepts: market power and dominance. The market power concept has been best developed in the United States, the dominance concept, in the European Union and its member states, particularly Germany. These concepts also have been incorporated in the competition laws of many reforming economies.<sup>13</sup> For example, recent competition laws proposed for Argentina and for the Philippines are based largely on market power, whereas the new law of the Russian Federation is based on dominance.

Whether based on market power or dominance, competition laws apply to similar forms of commercial con-

duct and transactions. For example, most laws address price fixing and other horizontal restraints, and resale price maintenance and other vertical restraints. Many address mergers and other corporate transactions. Nevertheless, laws based on market power tend to be enforced in different circumstances, and with a different commercial effect, than those based on dominance. The reason is that market power and dominance, although related, are not identical concepts. Market power refers to the ability of a supplier to exert a lasting influence on the market price or restrain the market output of a specific good or service. In general, enforcement actions based on market power tend to prohibit unilateral or concerted conduct that would result in higher prices for some good or service.<sup>14</sup> The dominance concept encompasses market power but goes further: A large enterprise may be considered dominant if it can restrict or foreclose the commercial opportunities of smaller rivals or trading partners, even if doing so has no effect on the exercise of market power.

The practical distinction between market power and dominance is revealed in enforcement actions. Consider merger control. Competing enterprises often merge in order to rationalize duplicative distribution assets and personnel and thus reduce the costs of the merged enterprise. If the market share of the merged enterprise is small, a competition law based on market power would likely find no threat to competition. Moreover, such a law would view reductions in distribution costs favorably: These would encourage lower prices by the merged firm and, due to competition, by its competitors. Both benefit consumers. Thus, under such a law the merger would be permitted. Yet a law based on dominance might prohibit the termination of distributors as an abuse of dominant position, thereby preventing merger-related cost reductions and competitive advantages.<sup>15</sup>

### *Commercial effects*

Broadly speaking, competition laws based on market power promote short-run allocative efficiency, in the sense that the legal prohibitions are designed to raise consumer welfare. Thus, commercial conduct and corporate transactions are judged by their effects on *buyers*. Commercial actions that improve or do not harm consumer welfare are judged to be legal; those that significantly reduce consumer welfare are usually found to be illegal. In contrast, competition laws based on dominance go further, by postulating and defending a legal right to commercial opportunity. This right is held by enterprises. Thus, commercial conduct and corporate transactions are

sometimes judged with respect to their effect on *suppliers and distributors*. As mentioned, conduct threatening the commercial viability of small enterprises is sometimes treated as an abuse of dominance, even though it may not bring higher prices to buyers. Such enforcement actions do not promote short-run economic efficiency; at best, they may enhance the commercial development of small enterprises.<sup>16</sup>

Competition laws based on dominance result in greater scrutiny of enterprises that are large in absolute terms (as measured by total assets or sales). In contrast, the market-power approach responds to the *relative* size of a supplier, measured in terms of its market share in a market for a specific good or service. A firm that is small in absolute terms may nevertheless possess market power and be liable under a competition law based on market power. Thus, the market-power approach imposes legal liabilities on both large and small firms, whereas the dominance approach imposes greater liabilities on large firms.

#### *Impact on technical efficiency*

In many cases, a transaction or form of conduct allows an enterprise to reduce its costs—in other words, to improve its *technical* efficiency. Absent an effect on market power, cost-reducing actions encourage lower prices by the enterprise and, due to competition, by its competitors. Both benefit consumers. Moreover, cost-reducing actions enhance the competitiveness of national enterprises in international markets.

Industrial and reforming economies often attempt to encourage technical efficiency by incorporating an efficiencies defense in various provisions of the competition law. This defense allows private suppliers to immunize concerted agreements and corporate transactions that would otherwise be prohibited. The defense is available in two forms. It may be sufficient for the parties to demonstrate that an agreement or transaction is necessary to achieve specific efficiencies. Alternatively, the parties may be required to show that efficiencies exist *and* will benefit consumers (through lower prices, better products, or better service).

#### *Dominance, growth, and efficiency*

Dominance-based enforcement is often criticized for protecting small, inefficient firms at the expense of preventing efficient conduct and transactions by large firms. This criticism frequently applies to cases involving predation, vertical restraints, vertical foreclosure, and conglomerate mergers.

Is dominance an appropriate basis for competition law

in a reforming economy? It is not clear—and certainly has not been proved—that promoting short-run efficiency necessarily leads to long-run economic growth.<sup>17</sup> There is increasing empirical evidence that small firms grow faster than do large firms, which suggests not necessarily that small firms are more efficient (in a static sense), but that they may be more responsive to new opportunities (see Evans 1987). In addition, some developing countries exhibit a “missing middle” in the firm-size distribution: There are numerous small and very large enterprises but relatively few medium-size enterprises. This suggests that enterprise growth is strangled once a firm advances beyond a minimum size.<sup>18</sup> To the extent that anticompetitive restraints originate with large firms, a dominance-based competition law, preventing privately generated restraints to trade and protecting the commercial rights of small firms, may enhance private sector development and long-run growth in a reforming economy.<sup>19</sup>

#### *Enforcement standards*

Fundamentally, a competition law provides standards of commercial conduct. Enforcement standards depend on the effectiveness of institutions, such as the competition agency and the judicial system. In mature market economies, competition investigations often must assess whether market forces are sufficient to prevent competitive harm. For example, if a merger between competitors were to result in higher prices, would new suppliers enter the market in response, thereby forcing prices back down? Or, can a small firm or entrant, the target of price predation by a large rival, finance the resultant losses through the capital market? Other things being equal, competition law enforcement should be relatively restrained in the presence of well-developed markets. In many cases, market mechanisms provide buyers, small enterprises, and investors with numerous commercial alternatives that negate prospective harm to competition.<sup>20</sup> Thus, where market mechanisms are effective, a competition law need not intervene.

However, in many reforming economies, market mechanisms and institutions may be absent or otherwise less effective.<sup>21</sup> In this event, a specific competitive concern, though implausible in a mature market economy, may be plausible or even likely. Other things equal, relatively interventionist enforcement is called for where market institutions are undeveloped.<sup>22</sup> Particularly in states with strong judicial systems, this is reflected both in strict legal standards (for example, definitions of dominance that do not weigh entry and exit conditions) and in the use of invasive legal remedies such as administrative prac-

ing orders.<sup>23</sup> As markets develop, enforcement authorities will increasingly encounter evidence that prospective competitive harm is unlikely or implausible in the presence of increasingly efficient markets.<sup>24</sup>

This suggests that enforcement should be more active in reforming economies than in industrial economies. But weaknesses in enforcement institutions (for example, a scarcity of personnel and other enforcement resources, corruption, or the lack of accurate commercial information) would support the opposite conclusion, since imperfect institutions may bring greater harm than imperfect markets.<sup>25</sup>

Institutional deficiencies raise a related concern: that enforcement will be inconsistent with market-oriented reform or biased in favor of specific interest groups. For example, a new competition agency might inhibit the use of mergers and joint ventures as a means of industrial restructuring, chill aggressive price competition through enforcement against predatory pricing or price discrimination, impede foreign competition by taking a strong stance against the use of vertical restraints by foreign suppliers, or reinstate a discredited system of price controls through administrative pricing orders.<sup>26</sup> Enforcement actions such as these can undermine economic liberalization and entrench the economy in its inefficient, pre-reform state.<sup>27</sup> To address this concern, recent competition reform efforts, particularly in Argentina, Mexico, and Taiwan (China), have expended considerable effort in improving the effectiveness of enforcement institutions.

These comments illustrate several important points regarding the implementation of new competition laws. First, competition law is inherently a *customized item*; it must fit national goals as well as the specific legal, commercial, and regulatory environment of the nation. Second, competition reforms are most effective if implemented within a *broader menu of market-oriented reforms*. Third, efficient legal standards are *necessarily dynamic, not static*. Not only must competition reform provide a code of conduct for the near future, it must also provide a *process* for adapting that code to the more distant future. Last, effective competition reform is often accompanied by significant *legal or administrative (regulatory) reform*.<sup>28</sup>

#### *Competitive analysis and competition advocacy*

The laws and regulations of the state strongly influence the vitality of competition by affecting the scope and size of antitrust markets, the level of concentration among incumbent suppliers, the conditions of entry and exit, and the costs and efficiency of incumbent or potential suppli-

ers.<sup>29</sup> In response, competition agencies often serve as an "advocate of competition." Such advocacy involves providing public and expert comment on the competitive aspects of laws, regulations, and other actions of national and local government bodies. Whereas law enforcement is generally concerned with the *preservation* of competition, advocacy is often concerned with the *creation* of competition.

Competition agencies have played significant roles in regulatory policymaking and have tended to support deregulation in reforming economies. The best recent example comes from the Republic of Korea, where the Fair Trade Commission in 1987 began evaluating reform of the licensing and permit system, price and quantity regulations, procedural restraints on commercial activity, and regulations on business line and geographic domain [see Fair Trade Commission (Republic of Korea) 1992, p. 30]. These regulations caused numerous, serious distortions. By 1991 the commission had established priorities and formulated a regulatory plan outlining 56 market reforms covering 21 industries.<sup>30</sup> A variety of new laws have been required to support liberalization in Korea, and the commission has provided recommendations on 33 legislative drafts, 21 of which have been accepted.<sup>31</sup>

There is a striking difference in competition advocacy between the industrial and reforming economies. In the former, advocacy is nearly always an advisory function; that is, the competition agency can only advise and cannot compel other governmental bodies with respect to economic regulations and policies. Naturally, this form of advocacy is often ignored. In contrast, competition advocacy has the force of law in some reforming economies.<sup>32</sup> In these countries, anticompetitive regulations that violate the law can be vacated, and penalties can be imposed on the agencies or individuals responsible for these regulations.

In a reforming economy, raising competition advocacy from an advisory to a law enforcement function offers substantial benefits. Properly applied, it facilitates the implementation of market reforms and helps ensure that reforms are not weakened by subsequent regulatory action at the regional or lower levels of government. Moreover, competition advocacy may sometimes provide the only meaningful constraint on intrusion by the state into commercial affairs.

Yet there are limits to competition advocacy: Without political support, no competition agency can guide economic reform. Thus, an agency can do little if political actors decide to abandon or weaken economic reforms.<sup>33</sup> But if economic reforms enjoy political support, compe-

tion advocacy can ensure that the actions of all levels of government are consistent with the chosen reforms (see Kovacic 1992).

In light of the heavily regulated environments of most reforming economies, competition advocacy provides an important institution for guiding and ensuring the orderly withdrawal of the state from commercial affairs. For this reason, the role of a competition agency as advocate broadly supports economic reform and is at least as important as its law-enforcement role—if not more so.

### **The antitrust regulation of conduct**

Competition laws prohibit commercial practices, contracts, and agreements that significantly lessen competition, strengthen the exercise of market power, or constitute an abuse of a dominant position. The legal standards vary across nations due to analytic uncertainties regarding commercial effect and to differences in the goals served by national competition statutes. Broadly speaking, antitrust enforcement against restrictive agreements has been considerably weaker in the reforming than in the industrial economies. This can result from many factors, most frequently from a lack of familiarity (among the private sector, enforcers, and judges) with a new competition law and from structural imperfections in enforcement institutions.

#### *Horizontal restraints*

An agreement among competing suppliers to limit or restrict, *inter alia*, pricing, investment, capacity expansion, product differentiation or advertising, is referred to as a horizontal restraint. There are numerous kinds of horizontal restraints; many are regarded as injurious and undesirable in market economies. Horizontal price fixing, for example, is nearly always prohibited in industrial economies, and violations frequently attract severe civil and criminal penalties. In contrast, other horizontal restraints, such as agreements to develop standards for the quality of products, are widely believed to enhance both competition and efficiency and are legal in most industrial economies.

Many scholars regard an unconditional (*per se*) prohibition against horizontal price fixing to be *the* centerpiece of an antitrust law. This reflects the strict (*per se*) prohibition imposed in most industrial economies, an enforcement posture resulting from the cumulative experience of thousands of cases examining price fixing and other horizontal agreements.<sup>34</sup> Only rarely have these investigations uncovered evidence of genuine social benefits.<sup>35</sup> On these grounds, reforming economies are often advised to

give high priority to the prosecution of price fixing.<sup>36</sup> Nevertheless, many new competition laws treat horizontal restraints more leniently, via rule-of-reason prohibitions, exemptions for restraints used by small enterprises, or weak enforcement.<sup>37</sup>

There is little doubt that lenient treatment of price fixing leads to its frequent use. Moreover, weak enforcement frequently degenerates into nonenforcement. In the Philippines, domestic cement suppliers have openly conducted monthly, public meetings to allocate cement-marketing territories.<sup>38</sup> This conduct is clearly illegal under Philippine antitrust laws, and such naked violations have resulted in a public perception that the law is not enforced. In an attempt to correct this situation, in 1992 the Senate of the Philippines received a proposed statute that would explicitly vacate the rule of reason as a legal standard under competition law.<sup>39</sup> In Australia, South Africa, and the United Kingdom, the process for granting discretionary exemptions to horizontal restraints has resulted in widespread evasion of the exemption process and in numerous, unregistered anticompetitive agreements among competitors (see Pengilley 1983, pp. 891–98). Similarly loose exemption procedures have been adopted in several Eastern European economies.

There are two reasonable options for reforming economies to address horizontal restraints so that price fixing and other clearly anticompetitive practices are prohibited, and potentially efficient restraints are permitted. First, statutory language may specify generic classes of restraints and apply different legal standards for each class. Mexico's new law does this by distinguishing between "absolute monopolistic practices" and "relative monopolistic practices" (see *Federal Law of Economic Competition*, chapter II, articles 9–10). The former encompass price fixing and similar agreements to reduce output, allocate sales territories, or rig bids in public auctions. These are almost always injurious and are therefore subject to a *per se* prohibition. Relative monopolistic practices encompass agreements on sharing technology and facilities, the competitive effects of which are more ambiguous. Thus, these are subject to a rule-of-reason prohibition for which a complainant must establish that the restraint harms competition, or imposes a selective or unfair commercial advantage.

Second, the competition agency can exempt horizontal agreements that may offer demonstrable efficiencies. In this regime, an agreement is immune from legal challenge only if the parties notify and receive the approval of the competition agency. Absent such notice, a horizontal agreement could be subject to severe legal penalties. For

example, Taiwan (China) has adopted the German system of discretionary cartelization to permit horizontal agreements that offer specific technical advantages to participants.<sup>40</sup> The application process requires participating suppliers to provide considerable information to the commission, including an analysis of cost savings and the resultant effect on prices, output, and the sectors with which the applicants trade. Exemptions are limited to three years, allowing the Fair Trade Commission to monitor and, if necessary, impose conditions on or vacate agreements that do not deliver the stated benefits (see Liu 1993, pp. 154–55).<sup>41</sup>

The competition laws of Mexico and Taiwan (China) use slightly different devices to encourage efficient horizontal agreements. In each, the threat of litigation imposes legal liabilities on the participants to unregistered or unreasonably restrictive horizontal agreements. In Taiwan (China), the Fair Trade Commission may temporarily remove these liabilities, whereas in Mexico the courts or the Competition Commission must define the scope and severity of the prohibition through precedential rulings. These devices have advantages and disadvantages with respect to enforcement,<sup>42</sup> but each offers the important advantage of clarity. The competition statutes of both Mexico and Taiwan (China) are quite specific regarding the forms of horizontal agreements that are subject to the law. Differences in legal treatment respond to differences in the potential efficiencies available from each form of agreement. In contrast, other competition statutes (for example, those of Kazakhstan, Philippines, and Ukraine) are vaguely worded, and their interpretation and implementation has been relatively slow. One would expect this to contribute to the widespread use of horizontal price fixing in these countries.

The more lenient treatment of price fixing in reforming economies most likely reflects a different mix of statutory goals, in particular a relatively greater concern for the interests of national suppliers relative to the interests of consumers. In this context, the competitive concerns raised by horizontal agreements are slightly different from those most often raised in the industrial economies, the principal concern being whether an agreement creates competitive disadvantage among national enterprises. Moreover, one might propose that price fixing would be in the interest of large national enterprises and should therefore be legalized.<sup>43</sup>

Nevertheless, even absent concerns for consumer welfare,<sup>44</sup> a prohibition against price fixing promotes the national economic interest. Price fixing in national wholesale or industrial markets directly harms national enter-

prises that buy in these markets. These buyers experience rising costs and may become uncompetitive, particularly if they sell into the international market. In addition, horizontal price fixing agreements are seldom confined to price alone; most often, participating suppliers must devise some means of excluding potential competitors attracted by the higher profits made possible by price-fixing. In short, price fixing distorts the comparative advantage of national enterprises and raises impediments to international commerce.

#### *Vertical restraints*

Vertical restraints are contracts that restrict the conduct of parties in a buyer–seller relationship, for example, a manufacturer and the distributors of its products. These restraints are usually judged by whether they facilitate the exercise of market power or result in competitive asymmetries among distributors. Broadly speaking, the commercial effects are more ambiguous for vertical than for horizontal restraints.<sup>45</sup> As a result, legal standards vary more across nations. Most industrial nations apply a *per se* prohibition to resale price maintenance, price discrimination, and territorial restraints, with other vertical (nonprice) restraints judged under various rule-of-reason standards.<sup>46</sup>

In the industrial economies, case studies suggest that individual vertical restraints seldom facilitate the exercise of market power among manufacturers. Although the restraints often reduce intrabrand competition, they seldom affect the substantial interbrand competition occurring in unconcentrated markets. Even in reforming economies, vertical restraints may enhance both competition and efficiency. For example, in *Pigi*, the Competition Commission of Greece examined an exclusive dealing arrangement between a manufacturer and 45 of 6,000 retail outlets in Greece (see Christoforou 1990, pp. 57–58). Under this arrangement, *Pigi*'s share of national (diaper) sales increased from 12 percent to 34 percent in three years, which shows that allowing vertical restraints by smaller suppliers can raise output and enhance competition. Nevertheless, the commission found the restraints to be illegal.<sup>47</sup> The legal standard used in this case follows the strict standard employed in the European Union, where vertical restraints are viewed as incompatible with a common market.

In reforming economies, where markets are often highly concentrated and entry into distribution often impeded by regulatory and capital-market constraints, vertical restraints are more likely to restrain trade and enhance market power. By imposing vertical restraints on

national distributors, national manufacturers may be able to raise the costs of entry by foreign manufacturers, inhibit direct foreign investment, and insulate the national market against the potential benefits of foreign competition.

Consider the experiences of Japan and Chile. In Japan vertical restraints have until recently been effectively exempt from competition law, resulting in a rigid, inefficient distribution system controlled by domestic suppliers.<sup>48</sup> In trade negotiations between Japan and the United States, these rigidities have been cited as structural impediments inhibiting the efforts of foreign suppliers to sell in Japan.<sup>49</sup> In contrast, the strict legal prohibitions against vertical restraints in Chile have resulted in a flexible and independent distribution sector controlled neither by large domestic suppliers nor by foreign suppliers.<sup>50</sup> This accounts in part for the success of Chile in opening its economy to foreign trade and competition. To the degree that economic reform is motivated by the benefits of international competition, the widespread use of vertical restraints is undesirable, particularly for a country that is attempting to promote private sector development.<sup>51</sup>

Most reforming economies employ a relatively lenient approach to vertical restraints, prohibiting them only as an abuse of a dominant position. This standard has been recommended for Jamaica and is being applied by Argentina, Chile, Hungary, Greece, the Slovak Republic, and Venezuela.<sup>52</sup> Treating vertical restraints as an abuse of dominance requires a complainant to establish that the firm imposing the restraints is dominant. This standard has some notable advantages for the reforming economy: Vertical restraints are prohibited for large, dominant suppliers but are legal for other suppliers. Thus, the restraints are legal where they are most likely to be technically or allocatively efficient.<sup>53</sup> Moreover, as markets and competitive forces develop, it becomes more difficult to show that a particular firm is dominant. Thus, the legal standard automatically weakens in response to the development of competition.

Economic reform is frequently judged by the creation of new suppliers. Where this is so, vertical restraints are best treated as an abuse of dominance, a rule-of-reason standard. Allowing nondominant firms to use vertical restraints can enhance both efficiency and competition. Conversely, preventing dominant firms from using vertical restraints can promote private sector development and support the formation of an independent distribution sector that is available on equal terms to domestic and foreign suppliers.

### **Regulation of market structure**

Most competition laws exert some control over market structure, that is, over private agreements that transfer the property rights or management of productive assets from one party to another. The primary forms of structural regulation are (a) merger control, (b) premerger notification, (c) constraints on corporate cross-ownership (of voting stock), and (d) constraints on interlocking directorates. Of equal importance, reforming economies are increasingly subjecting proposed privatizations to prior competitive review by a competition agency.

Broadly speaking, merger control and other structural regulations are designed to preserve the independence of suppliers and to prevent corporate transactions that would substantially eliminate competition.<sup>54</sup> Absent structural regulation, mergers would allow competing suppliers to coordinate their pricing policies; this conduct would otherwise be illegal.<sup>55</sup> In this sense, structural regulations complement the antitrust regulation of conduct.

In most industrial nations, competition law began with regulations on conduct. Structural regulations followed later. Particularly during the 1980s, structural regulation became a more important and widely used means of competition law enforcement: In that decade, merger control and other structural policies were enacted or strengthened in Australia, Britain, France, Germany, Greece, and New Zealand.<sup>56</sup> Reforming economies have followed this pattern, and recent competition reforms in Korea (1986, 1989), Taiwan (China) (1992), the Philippines (1987), Russia (1991), Ukraine (1991), Kazakhstan (1991), Hungary (1991), Poland (1988, 1990), the Czech and Slovak republics (1991), and Mexico (1992) all have instituted or strengthened merger-control policies (see Pittman 1992, Langenfeld and Blitzer 1992, and Liu 1993). Similar policies have been recommended for ongoing competition reform in Argentina and Jamaica (see White 1990 and Economists Incorporated 1992).

Often, merger-control policies must balance allocative efficiency and consumer welfare against technical efficiency and the competitiveness of suppliers. Mergers and similar corporate transactions (licensing agreements and joint ventures) often are motivated by gains in technical efficiency and seldom harm competition. As a result, in most countries enforcement actions are taken against no more than 1 to 2 percent of all mergers.<sup>57</sup> Moreover, in both large and small economies, merger control tends to focus on markets for which the effects of international competition are either weak or delayed.<sup>58</sup>

*Merger enforcement standards*

Merger enforcement standards in reforming economies tolerate comparatively concentrated markets.<sup>59</sup> Bulgaria, Hungary, Poland, and the Czech and Slovak republics all prohibit mergers that strengthen or create a dominant position (defined as a share of national sales above a given level: 30 percent in Hungary, Poland, and the Czech and Slovak republics, and 50 percent in Bulgaria) [see Bustamante 1992, pp. 9–11, and Antimonopoly Department (Slovak Republic) 1992, p. 5]. These standards are relatively lenient and focus enforcement on mergers involving the leading firm(or firms) in specific national markets. At the same time, whereas enforcers in industrial nations increasingly view nonhorizontal mergers as competitively innocuous and usually efficient, in many reforming economies mergers are judged by whether they create or enhance a dominant firm, an approach that supports active enforcement against nonhorizontal (that is, conglomerate and vertical) mergers.<sup>60</sup>

Merger control influences industrial structure and thereby can affect, for better or for worse, the efficiency of a national economy. Merger control based on dominance, with its focus on potential foreclosure on the part of leading firms, may be an appropriate means of promoting the development of small enterprises. Yet taken too far, it can prevent corporate combinations that would enhance both efficiency and competitiveness. For example, a vertical merger may allow a producer to ensure the availability of its inputs.<sup>61</sup> A conglomerate merger may allow private parties to pool capital and thereby provide a second-best alternative to a well-developed, private capital market.<sup>62</sup> These motives are less important in industrial economies, where rationing, shortages, and capital-market inefficiencies have already been corrected.

Reforming economies use three devices for ensuring that merger control does not impede efficient and pro-competitive mergers. First, some economies restrict the scope of merger control. For example, Korea and Mexico have designed merger control to apply only to markets and to enterprises above a specific minimum size. Second, a country may weaken its standards for judging mergers involving foreign suppliers, as Hungary does to encourage foreign investment. Third, a number of reforming economies offer an efficiencies defense, whereby merging parties can immunize a merger by demonstrating that the merger offers specific cost savings or other technical efficiencies.

An efficiencies defense allows the parties to a transaction to rebut the competitive concerns that a merger may raise; they can do so by showing that the merger would

produce cost savings that would enhance rather than harm competition. This defense can weaken the standards for judging a merger in two ways: first, by showing that the merger would be less likely to harm competition in a specific antitrust market; and second, by showing that the merger offers advantages in other markets that outweigh competitive concerns in a few specific markets. Because the defense is available, private parties are encouraged to undertake transactions for which cost savings and other technical efficiencies can be demonstrated.

A few reforming economies, such as Mexico and Venezuela, offer no formal statutory efficiencies defense.<sup>63</sup> But most do. In Greece, the Competition Counsel has been lenient with respect to mergers involving a failing firm.<sup>64</sup> Korea offers merging parties an efficiencies defense, and the Hungarian enforcement agency must consider the efficiencies occurring in *all* markets [see Pittman 1992, p. 15; see also Fair Trade Commission (Republic of Korea) 1990, pp. 49–50]. Generally, merger analysis treats efficiencies as a factor offsetting the price increases resulting from an anticompetitive merger; thus, only efficiencies affecting the relevant antitrust market (or markets) are pertinent to the analysis. In this regard, Hungary's efficiencies defense is unusually and explicitly broad. In essence, it allows the competition agency to approve a merger notwithstanding anticompetitive effects in a specific market if there are demonstrable efficiencies in other markets (where the merger does not harm competition).<sup>65</sup> The Hungarian efficiencies defense provides for more lenient control of mergers involving a conglomerate (to which the multimarket efficiencies defense best applies).

*Merger notification requirements*

Most nations with active merger control policies require the parties to large corporate transactions to notify the antitrust authorities. Some jurisdictions [France, Germany, the European Union, Mexico, Taiwan (China), and the United States] require premerger notification, whereby the parties to a transaction must notify the antitrust authorities and often must wait a prescribed period of time after notification before consummating the transaction. Others (Greece, New Zealand, Sweden, and the United Kingdom) require only postmerger notification. Premerger notification allows a competition agency to block a merger before consummation. As a result, all notified mergers are slightly delayed as the agency makes its evaluation; only those few that raise a competitive concern are further delayed (by an extended investigation, denial of approval, or litigation).

Numerous economies—including Hungary, Mexico, New Zealand, Poland, Russia, Taiwan (China), and the Slovak and Czech republics—require notification based on the merging parties' shares of national sales (see Pittman 1992, p. 6). For example, Poland and the Czech and Slovak republics require notification of any merger between parties accounting for more than 30 percent of national sales of a product. To avoid the notification requirement, private parties can include even remotely substitutable products in the "market," thereby underestimating their market shares. For this reason, notification standards based on estimated market share are ambiguous and difficult to enforce.<sup>66</sup> Thus, these standards are ineffective in selecting transactions for antitrust review. Recent competition reforms have moved away from this practice. For example, Mexico and Taiwan (China) have adopted premerger notification requirements that depend on the size of the merging parties (measured by sales or assets) and the value of the transaction.<sup>67</sup> Argentina has been advised to do likewise (see Economists Incorporated 1992, pp. 48–49).

Premerger notification enhances the ability of an enforcement agency to review in advance the competitive effects of proposed mergers and similar transactions. Allowing private parties to consummate an anticompetitive merger may create legal (contractual) rights that are extremely costly for an enforcement agency to overturn.<sup>68</sup> Thus, requiring advance notification of large mergers promotes effective merger control and lowers the administrative and legal costs of enforcement actions. However, this administrative efficiency is offset by an economic inefficiency: Premerger notification delays and thereby raises the transactions costs of notifiable transactions, most of which do not harm competition.

To minimize delay, competition laws frequently impose strict time limits on merger review by the competition agency. The new laws of the Eastern European nations grant waiting periods that are considerably longer than those applying to the competition laws of most industrial nations. For example, U.S. notification requirements can delay a merger as long as 20 days after the merging parties have provided information to the reviewing agency. In contrast, Hungary's competition law allows the agency nine months to complete its investigation; The Czech and Slovak republics both allow three months, and Poland allows two months (see Pittman 1992, p. 10).<sup>69</sup> The lengthy investigations allowed in these economies may reflect operational difficulties resulting from the lack of trained personnel and the scarcity of reliable commercial data. These institutional imperfections make it

extremely costly to monitor any but the very largest corporate transactions.

The rules and legal standards of premerger notification derive in part from legal protections granted to privacy and other personal rights. U.S. notification standards are conditioned on the provision of requested information by the parties to a notified transaction. The statutory period allowed for investigation does not begin until the parties have provided the requested information. Given the constitutional protections granted to (private and commercial) privacy in the United States, this conditionality is necessary to ensure that the competition agency has adequate information to reach reasoned decisions in enforcing merger control (otherwise, private parties would have an incentive to withhold information requested by the competition agency). In contrast, the Eastern European nations provide their competition agencies with very broad powers of investigation, powers that would violate the U.S. constitutional protections of privacy (see Pittman 1992, p. 9). Thus, the Eastern European nations need not condition their waiting periods on the provision of information by private parties. If and when the Eastern European nations implement legal protections to personal property and privacy, they must adapt their notification standards accordingly.<sup>70</sup>

#### *Merger-control remedies*

Fundamentally, a merger remedy allows the competition agency to resolve concerns arising from a specific merger. Most often, concerns arise for only a few of the many products of the merging enterprises. For other products, the merger may offer cost savings and other procompetitive efficiencies. If the merger is prohibited, these efficiencies are lost. For this reason, enforcers generally attempt to use remedies that are less intrusive than a blanket prohibition.

The use of merger remedies is closely related to a nation's merger-control standards, specifically to any efficiencies defense. As mentioned, Hungary judges a merger on its (net) efficiencies across multiple markets. Thus, a remedy need only ensure that these net efficiencies are positive. Usually (as in Germany, Korea, Mexico, Venezuela, and the United States), a merger remedy is judged on its capacity to preserve competition in a specific (antitrust) market.

The remedial action most often taken against an anticompetitive merger is divestiture, the sale of productive assets in the relevant antitrust market (that is, the market in which the merger harms competition). Ideally, a divestiture should preserve the premerger, competitive

status quo in the relevant market; it does so by replacing a merged supplier with a new supplier formed from the divested assets of the merging parties.<sup>71</sup> For divestiture to be a realistic and workable alternative to blocking the merger, the new supplier should be commercially viable in the relevant market and independent of the merging parties. Divestiture is commonly used when production and distribution are organized around multiple plants or operating divisions (in which case the new supplier is formed from one of the plants or divisions).

The nature of a remedial divestiture depends on the harm to competition, particularly in view of the impediments to entry into the relevant market. For example, a divestiture of a factory is appropriate when building a factory represents the primary impediment to entry.<sup>72</sup> Alternatively, in consumer goods markets, the expenses, risks, and delays of establishing a branded product represent the primary impediments to entry, in which case trademarks (rather than productive assets) may be included in the package of divested assets.

In some cases, remedial divestiture is not feasible.<sup>73</sup> In these circumstances, contractual remedies such as licensing and supply agreements may achieve the same purpose as divestiture: ensuring the supply of the relevant product by an independent producer. For example, Beer Lowenbrau attempted in 1985 to purchase the beer manufacturing facilities of a competitor in Greece; continued competition between the parties was to be ensured by a long-term supply agreement providing that Lowenbrau would manufacture for both parties, which would then compete in the retail beer market. The merger could be justified as a means of reducing the costs of manufacturing beer. Nevertheless, the Competition Commission ruled that the supply agreement, though eligible for exemption, was insufficient to immunize the merger, which the commission blocked (see Christoforou 1990, p. 71). This decision reflects a generic shortcoming of the application of contractual remedies to mergers: A contract introduces dependency between signatories, whereas merger remedies generally strive to create independent suppliers. Contractual merger remedies are unlikely to work well without reliable contract enforcement.<sup>74</sup>

These comments illustrate the interplay between national commercial institutions, such as contract enforcement, and the design and enforcement of competition law. A nation with effective institutions for enforcing contracts can conduct a relatively restrained form of merger control; in many cases, antitrust concerns can be addressed by specific and limited remedies such as asset divestitures or supply licensing agreements. Conversely,

imperfections in contract enforcement can limit the remedies available to a competition agency. This situation encourages a more invasive, interventionist form of merger control in which anticompetitive mergers are prohibited rather than amended or restructured.

#### *Competitive aspects of privatization*

It is unfortunate that certain countries have pursued rapid privatization irrespective of competitive concerns.<sup>75</sup> Of reforms in the former Soviet republics, former World Bank President Lewis Preston said,

We were too optimistic. . . . I think everybody made some mistakes. Sequencing is important . . . . you really haven't accomplished anything if you privatize in a monopoly situation.<sup>76</sup>

Several reforming economies have instituted administrative review of proposed privatizations, but to date only Poland and Germany—the latter a unique case—have exercised this authority to any significant degree. In many countries privatization has been limited to small firms, and large enterprises have been sold at a much slower rate. Elsewhere, large-scale, anticompetitive privatizations have proceeded with little regard for competitive harm.

Merger control and the competitive review of privatization represent complementary structural policies. Both policies use similar concepts and apply competitive analysis to proposed transactions that would alter the structure and perhaps the performance of a market. Yet there are important differences. Privatization, by replacing bureaucratic control with private control, offers technical efficiencies that probably dwarf the efficiencies of the typical merger. In addition, whereas merger control strives to *preserve* competition, the competitive review of privatization is designed to *create* competition.<sup>77</sup>

The competitive review of privatization can be designed either to promote allocative efficiency (that is, to create an industrial structure providing maximum output at minimum prices) or to inhibit the creation of dominant enterprises. The central question is whether a state-owned firm should be privatized as a whole or as multiple, competing enterprises. Absent competitive review, a privatized monopolist, though likely more efficient than the antecedent state-owned enterprise, may be well positioned to foreclose smaller enterprises and may have no incentive to share cost savings with its customers.

Proposed privatizations sometimes attract only one or a few bids from investors, and the scarcity of bids may limit the alternative means of privatizing an enterprise.<sup>78</sup>

A French investor, for example, submitted a bid for Poland's Hortex, a multiplant food processor, but expressed no interest in purchasing less than the entire company (Langenfeld and Blitzer 1991, p. 384).<sup>79</sup> This investor would not accept a divestiture designed to create competition among the various plants. Other foreign parties have invested in Poland only after negotiating for temporary protections of the domestic market through high tariffs; in most cases, the Antimonopoly Office has limited such requests to the time needed to execute an investment program (see Fornalczyk 1992, p. 6). These examples are not unusual; investors often demand regulatory concessions before investing in a state-owned enterprise.<sup>80</sup> These concessions make the investment more attractive for investors. Note, however, that other potential investments—in the trading partners or smaller rivals of the privatized firm—may then become *less* attractive.

There is a fundamental tension between the demands of rapid privatization and the need for workable competition during transition to a market economy. This tension arises because the value of an enterprise increases with its market power. A financially needy government therefore may be motivated to conduct anticompetitive privatizations or to enhance an enterprise's value through regulatory concessions (for example, high tariffs on competing imports). Frequently, the highest bid for a privatized enterprise is offered by a potential or actual competitor. Such premiums may be supported by higher prices in the market where the bidder and the enterprise would otherwise compete.

Bulgaria, Czechoslovakia, Germany, Poland, and the Slovak Republic, all have enacted or proposed competition statutes calling for review of prospective privatizations. Other nations have pursued aggressive privatization programs without competitive review and have created some notable private (near-) monopolists. The best example is provided by Kazakhstan, which has to date formed 80 state-owned holding companies in basic sectors such as construction, mining, oil and natural gas (exploration, mining, processing, and retailing), and pharmaceuticals.<sup>81</sup> Many of these companies are monopolists in several vertically related markets. In 1990 the government of Argentina sold Aerolinas, an international air carrier that subsequently merged with its only competitor on domestic routes.<sup>82</sup> The sale of Island Cement to Solid Cement Corp. created a firm accounting for 50 percent of cement sales in the Philippines (see U.S. Agency for International Development 1992, p. 88). Argentina, Mexico, and Venezuela all have offered a pro-

ected environment to investors purchasing local telephone enterprises. In Mexico and Venezuela these concessions have been offset by a requirement that the privatized entity perform subsequent divestitures.

Payments (for privatized assets) to the government may partially offset the social losses imposed by private monopoly. Thus, one might expect privatizations to be judged by weaker standards than those applied to mergers among private enterprises. Nevertheless, by sanctioning an anticompetitive privatization, a government mortgages the nation's future by encouraging higher prices and reducing allocative efficiency, neither of which promotes stable growth.<sup>83</sup> For this reason, reforming economies are increasingly being advised to subject proposed privatizations to competitive review and to restructure proposed sales that may reduce competition.

#### *Administrative pricing and the problem of monopoly*

In an antitrust context, administrative pricing refers to a legal or administrative decree ordering a (near-) monopolist to reduce its prices. In Germany and the European Union, administrative pricing orders have been used to reduce the prices of an enterprise determined—in a legal proceeding—to be a (near-) monopolist (Boner and Krueger 1991, pp. 86–89). This form of antitrust policy has been emulated in many reforming economies. The competition laws of Brazil, Kazakhstan, Mexico, the Philippines, Poland, and Russia all allow the government to dictate the pricing of a firm determined to be a monopolist or a dominant firm. Other violations also may be addressed by the imposition of price controls.<sup>84</sup> Kazakhstan, Poland, and Russia have been particularly aggressive in using administrative pricing to remedy allegations of "price gouging" immediately after the removal of broad-based price controls.<sup>85</sup>

Yet administrative pricing has an impact similar to controls, in terms of allocative and technical inefficiencies.<sup>86</sup> Consider the administrative pricing order issued by the Polish Antimonopoly Office against the automakers FSO and FSM, which raised their prices after price controls were removed in Poland (Langenfeld and Blitzer 1991, p. 382). Ideally, this order would result in competitive prices, but there is no good way to calculate competitive prices. In the industrial nations, administrative prices have been based on either cost data or comparable markets.<sup>87</sup> But in Poland this exercise is even more difficult due to a scarcity of reliable cost data.<sup>88</sup> Moreover, what markets are comparable to those of Poland during transition?<sup>89</sup> Due to these difficulties, administrative pricing orders are likely to set prices at ad hoc levels that have lit-

the relation to commercial reality. If the purpose of competition law is to encourage allocative efficiency, then nonstructural remedies such as administrative pricing orders are poor substitutes for structural remedies such as dissolution and divestiture.<sup>90</sup>

### **The structure of enforcement mechanisms**

Understanding the structure of enforcement mechanisms is critical to understanding competition law enforcement:

The differences in financial incentives and other conditions governing litigation exert systematic pressures that influence, in predictable and stable ways, the development of substantive law in different jurisdictions. In simplest terms, therefore, procedure influences substance (Prichard 1988, p. 451).

Scholarly discussions of competition law usually focus on the normative aspects of legal standards. One may ask, for example, whether mergers should be reviewed according to competition (as opposed to other) principles, whether resale price maintenance and other vertical restraints should be legal or illegal, or whether certain legal prohibitions should be enforced to a *per se* standard or to a rule-of-reason standard. Considerably less attention is given to the administrative institutions and legal procedures by which the law is to be enforced. Yet the design of the enforcement mechanism is at least as important as the design of legal standards.

The commercial effect of a competition statute or legal provision depends largely on interpretation and enforcement, that is, on the ability of enforcers to detect violations and impose penalties on violators, and ultimately whether enterprises and other legal persons are deterred from violating the law. For this reason, whether a particular legal standard is efficient—or has any commercial effect—depends in part on the characteristics of the enforcement mechanism.

Most antitrust prohibitions are subject to analytic uncertainties. The ability to monitor compliance with the law is equally uncertain. These uncertainties influence legal standards. For example, even horizontal-pricing agreements, though illegal in many nations, have been found to be efficient in some cases.<sup>91</sup> If detecting horizontal price fixing were error-free and costless, then a *per se* prohibition of price fixing would be unnecessary. Instead, the merits of price fixing could be evaluated on a case-by-case basis; presumably, most but not all cases would be found to be inefficient and therefore illegal.

Monitoring commercial conduct, however, is neither error-free nor costless. Thus, an unconditional *per se* prohibition can yield net benefits because it forbids conduct that is inefficient in almost all cases.

Although national competition laws often address similar forms of conduct and may employ similar legal standards, there are significant differences in the application and interpretation of these standards. These differences result in part from economic incentives built into the enforcement mechanism. An enforcement action occurs only when a complainant—a competition agency or a private party—decides to bear the costs of filing a complaint and proving that the law has been violated. The action occurs only if the complainant has both (a) the legal right (standing) to sue and (b) a financial or political incentive to sue. If the complainant lacks either, then no enforcement action can occur.

The opportunities and financial incentives for bringing an enforcement action depend on the procedures for enforcing the competition law. Moreover, it is through enforcement practices, more than any other factor, that each nation adapts competition law to its specific goals and circumstances.<sup>92</sup> Thus, understanding enforcement mechanisms is central to understanding how competition law can support economic reform.

The enforcement of a competition law in a reforming economy is often weaker than that in the industrial economies because information and enforcement resources are more scarce, and enforcement is often left to a weak administrative agency that is subject to undue political influence. Frequently, the enforcement rights of private parties are limited. The judicial systems of these economies are sometimes plagued by congestion, delay, and corruption. As in industrial nations, competition statutes in reforming economies are often vaguely worded, yet enforcers cannot rely for guidance on a long history of legal precedents.<sup>93</sup> All of these factors undermine the incentives of enforcers or potential complainants to undertake costly law enforcement actions.

A code of commercial conduct does not develop in a vacuum. It develops from a legal statute that describes basic rules of conduct and—in *equal measure*—from the application and refinement of these rules through enforcement. Weak enforcement impedes thereby the development of efficient rules of commercial conduct and contributes to a dearth of jurisprudence. For this reason, recent reform efforts have increasingly focused on enforcement mechanisms so as to ensure greater clarity and stronger enforcement of the competition law.

*Judicial and administrative enforcement*

In nearly all industrial nations, competition law is enforced judicially, that is, by a nonspecialized, independent civil or administrative court system. In contrast, many reforming economies have used administrative enforcement, whereby the law is enforced by administrators or through a specialized, nonindependent, quasi-judicial body (most often organized within another administrative structure such as a ministry). The principal distinction between the two is that judicial enforcement is defined by the availability of review of decisions and orders by an independent judicial or quasi-judicial body. Under administrative enforcement, such review either is not available or is not used.

In some economies a national judiciary addresses almost all legal disputes, including those arising under competition law. Examples include Germany, Hong Kong, Japan, Korea, New Zealand, Poland, Singapore, Taiwan (China), the European Union, and the United States.<sup>94</sup> Others (for example, France, Greece, Italy, Spain) maintain separate court systems: a civil court system for addressing disputes involving only private parties and an administrative system addressing disputes to which the state is a party. Either of these systems can be used to resolve disputes arising under competition law.<sup>95</sup>

Under administrative enforcement of competition law, legal interpretations and enforcement decisions are rendered by administrators, or by a judicial or quasi-judicial body that is (a) separate from that used to enforce the other national laws or (b) dependent on or subordinate to elected officials or their appointees. For example, the Swedish Competition Ombudsman enforces the competition law, subject to review by the Market Court, a specialized administrative court empowered to make legal interpretations of Sweden's competition law. The Argentine National Commission for the Protection of Competition is a quasijudicial body subordinate to the secretary of commerce and the minister of economy (see Cabanellas and Eizrodt 1983, p. 40). The commission is authorized to reach legal findings and make recommendations, but only the secretary of commerce has enforcement authority. Similarly, the superintendent of competition has (nonreviewable) authority to enforce Venezuela's competition law. Administrative enforcement is frequently exercised by ministers who have nonreviewable responsibility for merger control and legal exemptions—as in Great Britain, where the minister of economics has sole authority to enforce merger control laws. Before Mexico amended its competition law in 1993, the president alone had nonreviewable enforce-

ment authority (see Seidman 1989, Mexico section, p. 15).

Judicial and administrative enforcement differ primarily in the use of legal precedent to develop enforcement standards. Broadly speaking, under judicial enforcement, the enforcement standards of competition law derive from statutory language and from precedent. The use of precedent is international in scope, and nations with new competition laws sometimes rely on the legal precedents established in more developed nations. For example, the Competition Commission of Greece has relied on *Continental Can*, an European Union precedent, to develop legal standards regarding merger control, and New Zealand has relied on legal precedents established in Australia and in the United States (see Christoforou 1990, p. 68 and Ahdar 1991, pp. 218, 238). As a rule, international legal precedents are seldom adopted without amendment but are often used to frame the issues arising in an enforcement matter.

Standards of enforcement and legality may be derived from three sources: legislative amendment, judicial interpretation and precedent, and administrative discretion on the part of the enforcement agency. Legislative amendment to the competition law is generally the most costly of these alternatives and therefore is used least frequently. Judicial review is also relatively stable, though judicial precedents change in response to shifting circumstances (see Landes and Posner 1976).<sup>96</sup> The most flexible source of legal standards is administrative discretion, which has often resulted in relatively weak and unstable enforcement of antitrust law.<sup>97</sup>

In the United Kingdom, where competition law serves a broad "public interest" standard, the role of precedent has been enlarged in order to minimize private sector uncertainty regarding enforcement standards and to provide greater stability to U.K. (administrative) merger enforcement (see Lilley 1991, p. 19). This example illustrates that the devices of judicial enforcement can be applied to administrative enforcement. Nevertheless, because stabilizing devices such as precedent are used only at the discretion of the agency, and because the tradeoffs between competition and noncompetition goals involve the exercise of political judgments without review, administrative enforcement is likely to be less stable and more uncertain than judicial enforcement.

An enforcement agency must be able to exercise some discretion in order to apply its commercial expertise to enforcement. But in many ways, the discretionary actions of a competition agency under administrative enforcement can undermine market-based reforms. Broadly

speaking, legal uncertainty inhibits commerce because private parties may be deterred from engaging in transactions whose legality is unclear; that is, legal uncertainty attenuates the property rights on which private commerce is based.<sup>98</sup> Although economic liberalization necessarily involves adjustment by an enforcement agency to a changing economic environment, it is critical for the agency to keep the public informed of changes in enforcement practices and legal standards. That is why a competition agency often spends considerable resources in public education, particularly after a competition law has been enacted or amended or enforcement practices changed.<sup>99</sup> In sum, though administrative discretion has advantages, it often lacks transparency and accountability, raising uncertainty and thereby inhibiting commerce.

The stability—alternatively, the inflexibility—of judicial decisionmaking derives from its transparency, from the gradual development of legal standards, and from its decentralized nature. The parties to judicial enforcement actions often enjoy extensive and stable legal protections provided under the national constitution and other laws. These constrain administrative discretion. Judicial processes (for constructing a record of evidence, rendering a decision based on the record, and publicly disseminating information about the decision) are well suited to establishing public understanding of the rules imposed by a competition law. If the judiciary is independent from political parties, then changes in political administration do not quickly affect the legal standards applied by numerous judges. Similarly, a private right of action represents a decentralized mechanism of selecting cases for applying the law. Where individuals can sue to enforce the law, changes in administration are unlikely to affect the enforcement standards applied by numerous, independent judges.

In some countries administrative enforcement has been motivated by a lack of confidence in the judicial system. The delays experienced by litigants in Brazil and Bolivia would weigh against relying on judicial enforcement of competition in those countries (Crandall, Owen, and Skitol 1991, p. 15). In the Philippines, certain aspects of judicial structure appear to facilitate corruption and inconsistency in legal decisionmaking.<sup>100</sup>

Weaknesses in the judiciary may be addressed by establishing administrative bodies to substitute (at some level) for the national judiciary. For example, Jamaica was advised to form a specialized competition court as a means of resolving disputes arising under its new competition law. It has been proposed that the government of Argentina establish a new enforcement agency

(Prosecutor General of Competition) and a new court (Competition Court) to hear all disputes arising under the new competition law. This court, whose decisions could be appealed to the administrative courts or to the president, would substitute for criminal courts in Argentina (see Economists Incorporated 1992, pp. 34–39).

Imperfections in the judicial system may limit the options for implementing a new competition law. Absent a judicial capacity to enforce commercial contracts (as in Bolivia, Indonesia, and Mongolia), judicial enforcement of a new competition law would likely be ineffective. In some cases, the only realistic means of reform may require a limited competition law enforced solely through an administrative mechanism. In the extreme, constraints on enforcement mechanisms may so limit the benefits of competition reform that it is not worth pursuing.

In countries such as Bolivia and Mongolia, certain critical institutions of a market economy—the private ownership of productive assets, the delineation of private property rights, and the resolution of disputes involving these rights—are quite undeveloped. Without such institutions in place, market mechanisms, through which competition law exerts an effect, cannot function. Developing institutions may require legal (and regulatory) reform. Without such reform, competition reform would be premature. For these reasons, legal reform is a frequently useful and sometimes necessary complement to competition policy.

#### *Public vs. private enforcement*

The private right of action refers to the right of a private party to sue another for violating the competition law. As a rule, the competition laws of the industrial nations nearly always allow private parties to sue for compensatory damages resulting from violations of competition law. In addition, many nations allow private parties to enforce the law by suing to establish legal liability. For example, Germany and the United States allow a private right of action with respect to both legal liability and damages. Similarly, the Slovak Republic grants a private right of action under its new competition law [Antimonopoly Department (Slovak Republic) 1992, p. 9]. In essence, private rights of action *decentralize* the enforcement of the law. Alternatively, enforcement may be *centralized* in that the enforcement of certain provisions of the competition law may be reserved to a national enforcement agency (as in France and the United Kingdom). In Japan, private parties can sue for antitrust damages once liability has been established, but only the Fair Trade Commission may undertake actions to establish liability.

Due to constraints on the private right of action, competition law enforcement is generally far more centralized in the reforming economies than in the industrial nations. For example, private parties in Argentina must pursue complaints through the National Commission; they cannot independently sue in the commercial courts to establish legal liability under Argentine competition law.<sup>101</sup> A similar arrangement applies to the competition law of the Czech and Slovak republics.<sup>102</sup> In Poland, there is no private right of action either to establish liability or to collect damages, although the Antimonopoly Office can choose to demand compensatory damages from law violators (Langenfeld and Blitzer 1991, pp. 377–84).

Private incentives to sue strongly influence the extent to which legal systems, particularly those relying on precedent, develop efficient standards. For conduct that clearly violates the law, a private litigant can be confident of winning a lawsuit, and only the unreimbursed costs of bringing a suit might induce the litigant not to take action. Thus, private lawsuits promote voluntary compliance with the law. Where the success of a suit is less certain, deterrence depends on the likelihood of success and the penalties imposed on the violator. In legal systems that require private litigants to bear litigation costs and allow them to reach out-of-court settlements, private parties will tend to litigate only where the legality of the relevant conduct is uncertain and the economic benefits of a favorable ruling outweigh the costs of litigation.<sup>103</sup>

To the degree that judicial review is used to develop new legal standards, private parties may choose to litigate in order to obtain innovative or precedential legal rulings. This incentive is particularly strong for parties most directly affected by a specific legal standard.<sup>104</sup> Thus, the legal mechanisms of precedent and private bearing of litigation costs promote the selective use of the legal system to develop efficient standards precisely where those standards have the greatest impact on commercial activity (see Rubin 1977, pp. 53–55).

The efficiencies (or inefficiencies) of competition law enforcement are inherited in part from the structure of the judicial system. Some nations require that (a) the court, not the litigants, bear litigation costs or (b) all disputes be resolved through the court, so that out-of-court settlement is not an option for the parties to a dispute. For example, in Bolivia the court conducts the investigation and bears the litigation costs, and both Bolivia and Brazil require that litigated disputes be resolved only by the court, not by the disputants.<sup>105</sup> Absent cost bearing by private litigants, one would expect to see judicial review allocated via nonprice rationing, that is, via delay. Given

the delays encountered in resolving relatively minor commercial disputes, it would be ill-advised to rely on private enforcement of competition law through the judicial system. In contrast, out-of-court settlement provides private parties with a means of defending their legal rights without contributing to or bearing the costs of congestion in the judicial system.<sup>106</sup>

Policymakers in reforming economies often are reluctant to enforce a new competition law through a private right of action. This reluctance may reflect a desire to retain political control over competition law enforcement. Alternatively, the uncertainties regarding the production of legal precedents through private litigation may support fears that the law will be “abused,” that is, that inefficient or counterproductive legal precedents will result from private suits.

The concern that private parties may abuse judicial enforcement, or that judges may make mistakes, is valid. But “bad” rulings are not limited to reforming economies; nor are they limited to judges.<sup>107</sup> Clearly, enforcement practices must address the lack of public familiarity with any new competition law. Yet national courts seldom impose high penalties for violations of embryonic competition laws. Moreover, restraining the private right of action represents a very costly way of addressing this problem, because the enforcement mechanism can be designed to correct both legal uncertainty and the occasional bad ruling.

In a reforming economy with a private right of action, one would expect the uncertainties of economic reform to result in a high rate of private litigation and, therefore, in the rapid formation of legal precedents. The formation of legal precedents, by reducing uncertainty, reduces the private incentives to litigate. Thus, the private right of action is responsive and self-correcting.<sup>108</sup> In addition, to the degree that judges make mistakes and formulate inefficient legal standards, private parties injured by an inefficient standard have a natural incentive to litigate to vacate the precedent and change the standard.

If the stock of legal precedents is regarded as informational capital, then an enforcement mechanism should at the very least encourage the formation of such capital.<sup>109</sup> Restraining or withholding private rights of action retards the formation of legal precedents and chills public information regarding the legality of commercial conduct. This aggravates public ignorance regarding a new competition law and is accountable in part for the weak and erratic enforcement seen in many reforming economies.

*Centralized administrative enforcement*

From an efficiency perspective, what is the role of centralized enforcement agencies? Some economies [for example, Greece, Hong Kong, Singapore, and Taiwan (China)] have allowed private enforcement against unfair methods of competition (a commercial tort) without resorting to enforcement by an administrative agency.<sup>110</sup> Moreover, these economies have, for the most part, grown rapidly. One may ask: Can private enforcement alone be efficient? Must competition law be enforced by a public agency?

The advantages of centralized enforcement derive in part from natural constraints on the private incentives to litigate. Many forms of prohibited conduct directly injure only a few specific parties. This is often true of abuses of dominance such as predation or foreclosure, for which an enterprise causes significant injury to a specific competitor or trading partner. Where inefficient or abusive conduct injures a few specific parties, these parties have strong incentives to litigate on their own behalf. Consequently, these problems could likely be solved without resort to a central enforcement agency.<sup>111</sup>

Yet many forms of prohibited conduct impose damages that, though large in the aggregate, are small for each of many injured parties. In this case, private enforcement yields benefits that have “public good” properties; that is, litigants cannot exclude nonlitigants from benefiting from a successful enforcement action. Because some injured parties can “free-ride” on the enforcement actions of other injured parties, and because litigation is costly, private incentives support an inefficiently low level of litigation.<sup>112</sup>

Collective private actions, such as class action suits for which the injured parties finance a single lawsuit, represent a means of internalizing the social costs and benefits of litigation, and in this way efficiently strengthen private incentives to litigate. But collective actions may be unstable and costly for private parties to administer.<sup>113</sup> Thus, the primary role of a central enforcement agency should be to pursue precisely these actions.<sup>114</sup> Greece, after repeated attempts to treat anticompetitive conduct as a commercial tort under its civil, contract, and intellectual property laws, finally instituted centralized enforcement by an administrative agency in 1977 (Christoforou 1990, p. 49). This example mirrors that of numerous other nations, including France, Germany, the United Kingdom, and the United States.

*Administrative structure of an enforcement agency*

The efforts to implement new competition laws have faced recurring questions regarding the administrative

structure and location of a new competition agency: Should the agency be independent, or should it be located within a larger agency or ministry? Should the agency be responsible for enforcing noncompetition laws such as those pertaining to consumer protection, antidumping, or price controls? Should competition law enforcement be delegated to a single agency or to multiple agencies?

Nations enforce competition laws through a variety of administrative arrangements. Moreover, enforcement is often fairly effective, which suggests that there are few generic constraints on the administrative structure of a competition agency. As a rule, an agency operates with considerable independence, whether it is structurally independent (for example, Japan’s Fair Trade Commission and Australia’s Trade Practices Commission) or located within a ministry of economics or commerce. The latter arrangement is common practice among both industrial and reforming economies.<sup>115</sup>

A competition agency is rarely authorized to pursue policies that fundamentally conflict with competition policy. For example, competition agencies seldom can implement industrial policies or enforce antidumping or countervailing duty laws—policies that conflict directly with competition policy.<sup>116</sup> An agency with conflicting missions is invited to choose one mission to the exclusion of the other.<sup>117</sup> The weak enforcement of competition law in Britain and Sweden has resulted in part from administrative conflict, and Jamaica has been advised against assigning competition and antidumping enforcement to a single agency.<sup>118</sup> Finally, although pricing remedies have been used to enforce competition law, ongoing price regulation is not consistent with other market-oriented reforms.<sup>119</sup> For this reason, Jamaica was advised against reconfiguring its Price Commission, which designs and enforces price controls, to form its new competition agency (Crandall, Owen, and Skitol 1991, p. 31).<sup>120</sup>

Some scholars recommend that competition law be administered separately from consumer protection laws (see White 1990, p. 4). Yet competition agencies often enforce consumer protection laws regarding commercial fraud and truth in advertising (for example, the U.S. Federal Trade Commission and the Australian Trade Practices Commission). No fundamental conflict arises in this situation, because consumer protection laws are designed to promote an institution—namely, accurate and reliable advertising—that supports the competitive process.<sup>121</sup>

Poor administrative design promotes poor performance, and some scholars warn of the “regulatory cap-

ture" of a competition agency (see Schuck and Litan 1986, p. 53). By regulatory capture, they refer to undue influence exerted over an administrative agency by a specific industry or small group of private parties. In principle, regulatory capture is plausible for an administrative agency that deals with a specific industry in an ongoing fashion. But regulatory capture is not a realistic concern for a competition agency. Most private enterprises interact with a competition agency very infrequently. Thus, most enterprises have very weak incentives to engage in capture. The sole exception is the private (legal) bar, which might support fluctuating and uncertain legal standards as a means of encouraging costly litigation (by which private attorneys earn their living).<sup>122</sup> But there is no empirical evidence that a competition agency has ever been subject to regulatory capture, particularly in the presence of judicial review (see Posner 1972, p. 316).<sup>123</sup> Moreover, many reforming economies require that a competition agency select its chief administrators from commercial professions (namely, accounting, business, finance, and economics),<sup>124</sup> a requirement that may also reduce the potential for capture by the private antitrust bar.

#### *Penalty structure*

Competition law relies on voluntary compliance by private parties. Enforcement devices such as compensatory and punitive damages establish a financial incentive to comply. Compensatory damages are payments to parties injured through competition law violations; for example, a supplier found guilty of price fixing would be ordered to return illegal price premiums to its customers. Punitive damages, which are imposed in addition to compensatory damages, are designed to ensure that commercial parties have a financial incentive to comply with the competition law.

Competition laws allow for punitive damages because there are uncertainties in enforcement. Law violators may—with some probability—be able to engage in illegal conduct without being detected and successfully prosecuted. Under these circumstances, allowing only compensatory damages would result in a financial incentive to violate the law. Consider, for example, the decision by incumbent suppliers regarding the potential profitability of price fixing. Would subsequent prosecution remove suppliers' illegal profits (or more), leaving the suppliers no better off (or worse off) than if they had not fixed prices? If the price fixing is undetected or the prosecution unsuccessful, suppliers can keep the profits gained from price fixing. Absent punitive damages, suppliers would be no worse off from having fixed prices. The expected prof-

its of price fixing would be positive, and suppliers would therefore have a financial incentive to violate the law.<sup>125</sup>

Properly designed punitive damages eliminate this incentive and strengthen voluntary compliance. For example, (only) private claimants in the United States can recover treble damages for price fixing and related violations of the Sherman Act, and Germany allows the recovery of three times the illegal profits gained through violations of its competition law. The Philippines allows successful private complainants to recover treble damages plus litigation costs from violators of competition law, and Jamaica was advised to set financial penalties at some multiple of illegal gains (see Seidman 1989, Philippines section, pp. 21–22).<sup>126</sup>

Strengthening voluntary compliance through the use of punitive damages is a double-edged sword; taken too far, it can be counterproductive. If, on the one hand, a business practice has an ambiguous effect on competition and efficiency, then rule-of-reason treatment allows the law to prohibit the practice only where it is shown to be harmful. But, on the other hand, for potentially efficient conduct, punitive damages encourage litigation and deter private parties from engaging in the conduct, irrespective of potential efficiencies.<sup>127</sup> Such suits can impose considerable costs and risks on private parties that may be attempting to conduct themselves in a legal manner. For this reason, significant punitive damages should attach only to per se prohibitions for which legality is clearly defined, and not to rule-of-reason prohibitions for which the standard of legality is inherently ambiguous.<sup>128</sup>

In addition, treble damages and similarly severe penalties should be reserved for conduct that is both illegal and difficult to detect or prosecute. Conduct that injures a specific party (for example, resale price maintenance, exclusive dealing, tying, and many other vertical restraints) is usually easily detected and, if per se illegal, is easily judged to violate the law.<sup>129</sup> For such conduct, high penalties are not justified in light of near-certain detection. Where high penalties do attach to such violations, private litigation is likely to result even where competition is not threatened.<sup>130</sup>

Low penalties weaken the financial incentives to comply with a law and may undermine the ability to enforce a competition law. Both Mexico and Brazil under earlier law allowed private parties to initiate enforcement actions, but neither country allowed significant punitive damages.<sup>131</sup>

Some competition laws impose, in addition, nonfinancial penalties, the major forms of which are criminal penalties and public apology. Criminal penalties, that is,

jail sentences for those who violate the law, are generally reserved for only the most egregious and injurious violations. These are appropriate where ownership and control of an enterprise are separated.<sup>132</sup> In Japan and in the Republic of Korea, law violators may also be required to issue public apologies. While seldom incorporated into the antitrust laws of Western nations, public apology does appear to be costly for violators (for cultural if not economic reasons) and may provide an effective means of ensuring compliance (Boner and Krueger 1991, p. 45).

Some scholars have recommended that new competition laws be designed to address only the most egregious violations (for example, horizontal price fixing) through per se prohibitions and criminal penalties (see Willig 1991, p. 191).<sup>133</sup> This advice is unsound because criminal penalties often require criminal standards of proof, which are usually quite high. Typically, the commercial and economic uncertainties intrinsic to enforcement of competition law do not allow a plaintiff to meet these higher standards of proof. Argentina, Canada (before 1986), and the Philippines originally developed competition laws relying almost entirely on criminality. In Argentina and Canada, subsequent enforcement was extremely weak owing to the high standards of proof by which plaintiffs' claims were judged.<sup>134</sup> To correct this problem, recent reforms (Argentina in 1980 and 1992, Canada in 1986) have made greater use of civil process and civil penalties. The Philippines now may be following these examples.<sup>135</sup>

In summary, because competition law relies on voluntary compliance by private parties, the vitality of enforcement mechanisms strongly affects the commercial impact of the law. Obviously, no law can be effective without active enforcement. In addition, the structure of enforcement mechanisms greatly influences the nature and development of legal and enforcement standards, as well as the likelihood that a nation will develop efficient standards appropriate to its unique goals and circumstances. The institutional and procedural aspects of competition reform are at least as important as the design of the statute. Thus, to be effective, competition reform must provide a process for determining and, if necessary, adjusting the rules that govern commercial conduct. It is not enough to simply prescribe a fixed set of such rules. This is particularly true for reforming economies with rapidly changing commercial environments.

### **Directions for future research**

The adaptation of competition law to reforming economies raises a number of important questions. For some, good answers can be extracted from the history and

analytic framework of antitrust enforcement in the industrial economies; for others, answers can be found only through further research. Much is now known about the short-run commercial effects of specific antitrust provisions. There is widespread—though by no means unanimous—agreement regarding the types of provisions that should be written into a competition law. Governments know how to write “sensible” competition laws. But how should the law be implemented? What institutional framework is best suited to supporting market-oriented reform? This chapter has provided a conceptual framework, adapted from law and economics, in which these questions can be examined. Nevertheless, many questions remain.

To be of value to a reforming economy, competition law must contribute to efficiency, growth, and private sector development. The microeconomic basis of competition law suggests that it can do precisely this. Yet neither theory nor observation would suggest that developing an effective “code of competition” is easy, rapid, or stable. In nearly all nations, the early enforcement of a competition law is left entirely to an administrative agency and is often weak or perverse. As in Argentina, Japan, and the United States, it may take years for a new competition law to be enforced. Alternatively, inactive or ineffective enforcement may persist, as in Canada and the Philippines. Elsewhere, as in Kazakhstan, Poland, and Russia, strong early enforcement may be counterproductive. In most nations, enforcement improves over time. One may ask, what factors support viable, sensible enforcement?<sup>136</sup> If sensible enforcement cannot be provided—and, for a reforming economy, relatively quickly—one may wonder whether competition law can be a useful component of market-oriented reform.

Private rights to enforce competition law have a long history in industrial nations such as the United States and most EU member states. Many reforming economies do not offer a private right of action. Consequently, one would expect a different pattern of enforcement. For example, “plaintiff’s antitrust”—involving predation and other abuses of dominance—might disappear without a private right of action.<sup>137</sup> Do we observe this occurring in reforming economies? That is, does enforcement respond to economic incentives? More fundamentally, where market forces have long been suppressed, do economic incentives influence the performance of enforcement institutions?

Broadly speaking, a competition law prevents rent seeking by private enterprises; that is, it prohibits specific commercial tactics—for example, price fixing—that raise

private profits while reducing social welfare. Might these tactics then be replaced by others, such as rent seeking, through the political or regulatory process? To address this concern, Rodriguez and Williams (1994, pp. 209–32), promote competition advocacy as a means of inhibiting regulatory and political rent seeking. Unfortunately, advisory advocacy may have limited effect.<sup>138</sup> Stronger forms of advocacy—such as the compulsory forms used in Kazakhstan and the Czech Republic—might be desirable.<sup>139</sup> But in that event, there is a bit of a dilemma: Private enterprises, in lobbying for commercial advantage, are exercising valuable political rights; in conducting a strong form of advocacy, a competition agency may be reducing these rights. It is far from obvious how this tension—between efficiency-enhancing advocacy and a democratic institution, private lobbying—should be resolved.

The above questions, and many others, derive from a single theme best expressed by Douglas North in his 1993 Nobel Prize speech. For industrial economies, much is known about the relation between economic performance and the formal (that is, institutional) and informal rules of commercial activity. Thus, much is understood about the contribution of institutions to economic growth. Far less is known of the reforming economies, whose institutions and informal rules have been very different,<sup>140</sup> and without a better understanding, the economic improvement of reforming economies may lag.

### Conclusions

In recent years, numerous countries have engaged in market-oriented reforms. Competition law and policy reforms have been an important component, and the recent development of new competition laws is unprecedented.

A new competition law can be judged only by its commercial effects. To implement competition reform effectively, then, one must understand the institutional aspects of law enforcement and advocacy—in particular, how statutory language exerts an effect, through enforcement, on the commercial conduct of economic entities.

The commercial effect of competition law flows not only from a specific set of legal standards but also from the characteristics of enforcement mechanisms. The discussion in this chapter has emphasized the role of a variety of legal (and administrative) mechanisms—notably private rights of action, informal settlement, cost sharing, civil penalties, and precedent—in influencing the commercial results of competition law. The discussion of enforcement illustrates the range of concerns that must be addressed in implementing a new competition law and policy.

Designing new competition law is achieved not by following a simple recipe but rather by adapting new institutions to the present requirements of a national economy.

### Notes

This piece complements and expands on an earlier survey of international antitrust; see Boner and Krueger 1991. The author thanks Claudio Frischtag, Douglas Webb, F. M. Scherer, Steve Nelson, and Armando Rodriguez for many helpful comments and suggestions. The opinions expressed herein are those of the author alone and do not necessarily reflect the views of the U.S. Federal Trade Commission, the U.S. Agency for International Development, or the World Bank.

1. Competition policy is herein defined narrowly, as a national policy intended to preserve or enhance competition between private enterprises serving the national marketplace. This definition excludes policies (for example, taxation) motivated by concerns other than competition. At this writing, the following economies and country groups are considering or have recently enacted domestic competition reforms: Argentina, Australia, Britain, Bolivia, Brazil, Bulgaria, Canada, Chile, Colombia, Costa Rica, the European Union, Ecuador, the EFTA, Germany, Hungary, Italy, Japan, Kazakhstan, the Republic of Korea, Mexico, New Zealand, Poland, Russia, Slovak Republic, Taiwan (China), Uruguay, and Venezuela. See Azcuenaga 1992.
2. A variety of national goals—such as equity, fairness, the promotion of trade, or the promotion of small enterprises—may motivate a national competition policy. The discussion here will focus on competition policy as a means of enhancing economic efficiency and economic growth.
3. In March 1992, Brazilian and Argentine steel producers announced an agreement to limit crossborder shipments of steel in the post-MERCOSUR environment. (MERCOSUR is an agreement designed to expand trade between Brazil and Argentina.) After Brazil lifted price controls on dairy products in 1990, Brazilian dairy producers met publicly to set prices at significantly higher levels. In many nations actions such as these would be viewed as horizontal price fixing and would attract severe criminal and civil penalties. See Willig 1991, pp. 187–88.
4. In Japan the natural conflict between industrial policy and competition policy has resulted in binding constraints on the former. Since the 1950s, Japan's MITI has conducted industrial policy by encouraging and sometimes requiring independent suppliers to form cartels. In the last twenty years, Japan's Fair Trade Commission has been authorized to deny cartel applications on competitive grounds. Thus, what began in Japan as competition *advocacy* ultimately became competition *law*. See Yoshikawa 1983, pp. 489–504.
5. Competition advocacy is compulsory when a law applies to state employees and agencies. For example, Article 6 of Kazakhstan's Antimonopoly Law prohibits bodies of state authority and gover-

nance from adopting acts that restrict the independence of economic entities or discriminate against specific commercial entities.

6. See Langenfeld and Blitzer 1991, p. 50; Christoforou 1990, p. 50; and Coate, Bustamante, and Rodriguez 1992, pp. 52–55 and 67.

7. See Coate, Bustamante, and Rodriguez 1992, pp. 62–79.

8. For example, almost all industrial nations enforce a *per se* prohibition of horizontal price fixing. See Boner and Krueger 1991, p. 51.

9. The competition laws of reforming economies are frequently based on dominance. Such a law promotes private sector development by granting specific legal protections to small enterprises. In mature economies with a well-developed private sector, the dominance concept is less suitable and less frequently applied. Dominance, focusing on the large (near-) monopolist, has been developed primarily by the European Union and its member states, particularly Germany. Yet as these economies have grown, the primary concern of competition law enforcement has shifted from dominance to multilateral market power, an approach that does not focus solely on large or monopolistic firms. See Boner and Krueger 1991, p. 78; and Coate, Bustamante, and Rodriguez 1992, p. 70.

10. To illustrate, an enforcement action against a merger generally requires a finding that entry into a relevant market is impeded. Implicitly, this finding means that the market mechanism would not automatically correct distortions caused by noncompetitive conduct. See Boner and Krueger 1991, section V.

11. This statement, although obvious, is frequently overlooked in discussions of the relative efficiencies of market and nonmarket mechanisms. Some scholars attempt to characterize market efficiency in absolute terms (for example, marginal cost pricing) and recommend a regulatory solution whenever this condition does not hold. See Murrell 1991, pp. 59–76. So defined, a market may “fail” even where a regulatory solution could not improve matters. For further details see Hahn 1990 pp. 211–18.

12. See the debate between Godek 1992 and Boner and Langenfeld 1992.

13. The dissemination of legal systems and statutes across international borders has a long history. For example, Japan adopted the Prussian civil code in the late 1800s; see Ramseyer 1989, p. 51–77. For a discussion of U.S. antitrust principles in New Zealand competition law, see Ahdar 1991, pp. 217–47.

14. The economic grounds for competition law derive largely from a single result, namely, that the pricing and output of a pure monopolist are *allocatively inefficient*. That is, a monopolist’s price is too high, and its output too low, so that aggregate welfare would rise if additional resources were invested in the market served by the monopolist. In contrast, resource allocation with perfectly competitive or contestable markets is, under general conditions, allocatively efficient. Broadly speaking, the inefficient conduct of a monopolist results from an absence of competition, which is why competition law (based on market power) is often suspicious of conduct or transactions that prevent or inhibit competition.

15. One might imagine that a dominant firm must have a large market share. Yet the German cartel office once found Rossignol, whose market share was 12 percent, to be dominant. See Boner and Krueger 1991, p. 78. Although U.S. competition law is not based on dominance, the recent Kodak case makes sense only as a dominance action. A lawsuit resulted from Kodak’s refusal to supply its parts to independent repair and service enterprises. Could this harm consumers? Both litigants stipulated that Kodak possessed no significant market power with respect to office copiers and other equipment. The court established that Kodak had the ability to overcharge certain “locked in” customers. Yet, without market power, Kodak would have no *incentive* to do this. Nevertheless, the U.S. Supreme Court did not dismiss this case. Thus, any merits of the case must derive from legal rights granted to enterprises rather than consumers. See Caulkins 1993, pp. 285–310.

16. At worst, these actions may reduce both short-run and long-run economic growth and welfare.

17. Removing an impediment to short-run efficiency generates only a one-time increase in economic output; it does not necessarily increase the rate of growth. See Lucas 1986.

18. In industrial economies, the firm-size distribution is generally close to the Pareto distribution, a unimodal distribution for which the rate of (firm) growth is statistically independent of the size of the firm. Size can be measured by assets, sales, or number of employees. In contrast, in the Philippines, the firm-size distribution is strongly bimodal: There are very many small firms, many large firms, but relatively few medium-size firms. In this case the growth rate of the firm is highly dependent on its size, and at some point small firms do not continue to grow. This pattern is consistent with a dual economy supporting many small, informal enterprises and a few very large, formal enterprises. The constraint on growth may result from access to organized capital markets being related to size. See the 1988 *Census of Establishments*, National Census and Statistics Office, Manila.

19. This effect is particularly plausible where reforms encourage the rapid formation of new private enterprises. In Hungary and Poland, for example, new private enterprises account for more than 20 percent of gross national product in spite of the slow privatizing of large enterprises. See Svejnar 1991, p. 127. The Republic of Korea, whose growth has been propelled primarily by large enterprises, found that by 1979 the productivity of small and medium-size enterprise rivaled that of large enterprises. See World Bank 1987, p. 33.

20. For example, vertical restraints can reduce *intra-brand* competition among distributors. But this reduction is harmless if distributors are subject to sufficient *inter-brand* competition. Similarly, the “deep pocket” allowing a large predator to engage in persistent below-cost (predatory) pricing can be offset by the equally deep pocket of a private capital market to which the prey, usually a small

rival or entrant, may have access. See Boner and Krueger 1991 p. 57 and pp. 63–66.

21. Reforming economies often have a poorly developed capital market, highly concentrated goods markets, and a high concentration of wealth. Imperfect access to capital generally impedes resource mobility and thereby reduces the importance of entry and exit as a market force restoring efficiency in poorly performing markets. Frequently, a few large domestic enterprises exert a strong influence on national commerce and politics. The early environments of Japan (with its *keiretsu*) and Korea (with the *chaebol*) are notable examples, and Argentina, the Philippines, and many other small economies exhibit similar structures.

22. A variety of institutions are necessary for the market mechanism to function efficiently. These include well-defined and -enforced private property rights, commercial legal codes describing contractual rights, privatization of state-owned enterprises, stability and convertibility of currency, removal of trade barriers, regulatory reform and the removal of price controls, and the establishment of private capital and credit markets. See Willig 1991, p. 187.

23. See Boner and Krueger 1991, section VI and Langenfeld and Blitzer 1991, p. 382.

24. For example, in *Fisher v. Paykel*, the New Zealand high court dismissed a complaint regarding distribution restraints. The court found that impending entry by Australian manufacturers (previously barred from the market) would limit competitive harm to the short run. See Ahdar 1991, p. 242. In Kazakhstan regional anti-monopoly committees originally considered each oblast to be an isolated economic entity. Thus, for a regional market, the committee could rely on regional sources of information. In the past two years, the committees have discovered the need for inter-regional information to evaluate growing inter-regional competition.

25. An enforcement concept, though theoretically sound, is useless if lack of information prevents the proper application of the concept. For example, price predation does not make economic sense if the predator is charging prices above its marginal costs. But even in developed economies, it is difficult to estimate marginal costs, and enforcers frequently use average cost as a surrogate for marginal cost. Unfortunately, average cost is the wrong measure by which to define predation, and aggressive price competition is sometimes alleged to be predatory. This example illustrates that an institutional shortcoming—in this case, a scarcity of cost information—can result in an *inefficiently high* level of enforcement, even in industrial economies. Reliable cost information is equally scarce, if not more so, in the reforming economy. Thus, constraints on information constrain enforcement. See Boner and Krueger 1991, pp. 64–66.

26. The “Price Act” of the Philippines includes a few antitrust provisions. But the act is primarily designed to (a) authorize administrative price controls by several agencies, none of which is a genuine

antitrust enforcement agency, and (b) prevent hoarding. Unfortunately, since price controls commonly cause shortages, to which hoarding is the natural response, the act would tend to distort market incentives and impede the development of markets. See Congress of the Philippines, *Republic Act No. 7581*, Metro Manila, July 22, 1991.

27. Thus, Godek 1992 questions using competition policy as a means of promoting economic reform. He argues instead that trade liberalization is the best reform measure for a small economy. Therefore, any new competition law should be limited, with weak enforcement against predation, vertical restraints, price discrimination, mergers, and other forms of conduct. For a contrary view, see Boner and Langenfeld 1992.

Constraints on the discretion of the antitrust agency respond to the concern that the agency would run amok. Constraints can be imposed in two ways: first, through the design of legal standards that require the agency to bear a specific burden of proof (see sections 3 and 4, *infra*); and second, through the design of administrative procedures controlling the development of agency investigations, decisions, and orders (see section 5, *infra*).

28. In some reforming economies, regulatory and judicial institutions may be slow, corrupt, or highly politicized. Other countries have encountered these problems and addressed them through systems of administrative law (encompassing civil service, the budget process, and the delegation of authority to public agencies), devices that are frequently less developed in the reforming economies. See World Bank 1993, section 5.C.

29. See Boner and Krueger (1991), sections 2 and 7. In Latin America the developmental strategy of “import substitution” has created a variety of institutional and regulatory impediments to competition; these provide good targets for amendment through competition advocacy. See Coate, Bustamante, and Rodriguez (1992), pp. 40–44 and 56–57.

30. The regulatory changes affect petroleum refining and marketing, liquor, financial services (banking, insurance, and securities), trucking, maritime shipping, automobile maintenance, engineering, and telecommunications. Other reforms are designed to encourage entry into food, construction, chartered aer\_shipping, and industrial gases. See Fair Trade Commission (Republic of Korea) 1992, pp. 30–31. The Korea of 1975 is reminiscent of certain former Soviet republics. Until 1975, Korea’s economic strategy relied greatly on the formation of large conglomerates and on price controls (just as a number of former Soviet republics are now encouraging the formation of large holding companies subject to price controls). Since that time, Korea’s economic strategy has taken a sudden and surprising turn toward liberalization and relatively unfettered competition. See R. Boner (1994a).

31. Elsewhere, advocacy has strengthened certain forms of regulation. For example, the early efforts of the U.S. Federal Trade Commission helped establish regulation in radio broadcasting

(Radio Act of 1927, Federal Communications Act of 1934), energy (Public Utilities Holding Company Act of 1935), and securities (Securities Act of 1933). See Scherer 1990, pp. 465–70.

32. Some scholars fear that the enforcement efforts of nascent antitrust agencies will be biased in favor of state bodies and enterprises and against foreign suppliers and embryonic private enterprises. One does not see this in Kazakhstan. To date, the largest fine for a law violation was imposed on the Almaty City railroad system, a state-owned enterprise (interview with Almaty City Antimonopoly Committee, May 1992).

33. For example, the Council of Ministers of Kazakhstan has actively monopolized numerous basic industrial sectors by forming sector-wide holding companies. Although originally intended as transitory vehicles for protecting state property and then privatizing state enterprises, these companies have become permanent. Privatization has proceeded only very slowly. The Antimonopoly Committee and its chairman, a council member, have repeatedly and unsuccessfully objected to the formation of many of the holding companies.

34. This is particularly so regarding agreements directly restricting price, output, investment, or marketing territories. See Pengilly 1983, p. 890.

35. A counterexample is *Broadcast Music, Inc., v. Columbia Broadcasting Systems, Inc.* [441 U.S. 1 (1979)]. In this case, BMI administered a collective agreement among thousands of recording artists and hundreds of radio stations fixing royalty payments for music broadcasts. The agreement was found to be legal for two reasons: first, it greatly reduced the transactions costs that would have resulted from bilateral negotiations between numerous artists and radio stations; and second, artists and stations were free to reach bilateral agreements if they wished. In short, the agreement was structured so that it would most likely raise the quantity and variety of music broadcasts.

36. Managers accustomed to central planning and price controls often use a variety of informal price-fixing agreements after privatization and the removal of price controls. To combat such agreements, Poland enforces strongly against price fixing and carefully scrutinizes the pricing of the independent units formed from dissolving a state-owned enterprise. See Fornalczyk 1992, p. 7. In countries of the former Soviet Union (for example, Kazakhstan, Russia), enforcement has focused on the implementation of price controls rather than the prevention of horizontal price fixing.

37. Argentina, Hungary, Poland, and Taiwan (China) all treat horizontal restraints under a rule-of-reason analysis in which complainants must demonstrate the harmful effects of the restraints on a case-by-case basis. See Langenfeld and Blitzer 1991, p. 388; Liu 1993, p. 154; and Economists Incorporated 1992, pp. 4–5. In the Czech and Slovak republics, horizontal agreements (including agreements to fix price) are exempt from the general prohibition against cartel contracts if the participants to the agreement account

for less than 5 percent of sales in the republican market or less than 30 percent of sales in a local market (see Competition Protection Act, Article 3.3.D). The standard for exemption requires that private parties exercise considerable judgment in estimating their sales share of a market. To gain exemption, participating enterprises would have an incentive to include the broadest collection of substitutes in the market. Such claims are not always easily verified. Consequently, the use of restrictive agreements by small enterprises is likely to be more widespread than contemplated by the new law.

38. See U.S. Agency for International Development 1992, p. 87. In the Philippines, cement is subject to the antitrust provisions of the Price Act, the stated purpose of which is “to ensure the availability of basic necessities and prime commodities at reasonable prices. . . .” See Republic Act No. 7581 (Price Act), Congress of the Philippines, Metro Manila, July 22, 1991.

39. This proposal, Senate bill no. 845, represents a serious overreaction to a problem that can be addressed by other means. Notwithstanding the competitive harm that almost always accompanies price fixing, vacating the rule of reason for horizontal restraints represents an extremely costly means of strengthening enforcement. Certain nonprice restraints exert an ambiguous effect on competition and efficiency, and others—for example, industrial product standards—almost always enhance both competition and technical efficiency. These advantages would be lost through a per se prohibition. See Senate bill no. 845, Manila, October 15, 1992.

40. The exemptions include cartels addressing standardization, research and marketing, specialization (of productive facilities), export and import, recession, and small businesses. See Liu 1993, pp. 155–56.

41. One potential problem with the new Taiwanese law is that horizontal agreements are judged under a “public interest” standard, an ambiguous and ineffective approach tried in the United Kingdom and since amended. See Boner and Krueger 1991, pp. 51–52.

42. See the earlier section on enforcement standards in reforming economies for a discussion of the comparative performance of administrative enforcement (through an agency) and judicial enforcement (through the courts).

43. For example, in countries of the former Soviet Union, proposals to form holding companies are intended to combine strong and weak enterprises, thereby propping up the latter at the expense of the former. See Joskow, Schmalensee, and Tsukanova 1994, pp. 346–47.

44. To wit, Mexico’s new competition law is explicitly not intended as a device for promoting the interests of consumers.

45. Whether vertical restraints should be prohibited has been hotly debated in recent years. Many economists believe that vertical restraints offer technical efficiencies and therefore should be legalized. See Winter 1993 and Klein and Murphy 1988. For a contrary view, see Rey and Tirole 1986.

46. In the United States certain forms of resale price maintenance are per se illegal, and nonprice vertical restraints are generally judged under a rule of reason that addresses the likely competitive effect of the restraints. To support a common market, the European Union applies a per se prohibition against territorial restraints, as does the German competition law to vertical restraints, exercised by dominant firms. See Boner and Krueger 1991, section IV.

47. One could debate whether a 34 percent market share is sufficient for a manufacturer to threaten the viability of smaller trading partners (here, diaper retailers). Nevertheless, this case poses a bit of a paradox: To the degree that vertical restraints allow a supplier to operate efficiently, they will ultimately be found illegal (because the supplier, through growth, will ultimately be found to be dominant). Actually, the paradox is more apparent than real; in Pigi's case, the efficiencies of vertical restraints would decline as the Pigi brand name became better established with buyers.

48. Vertical restraints were immune under early Japanese enforcement. In 1953 the Tokyo High Court held that actions against vertical restraints required proof that the restraints resulted from an agreement among competing manufacturers or competing retailers. This enforcement posture was slightly amended in 1962, when vertical restraints began to be judged under a weak rule of reason. See Boner and Krueger 1991, p. 61.

49. The Japanese Fair Trade Commission issued new enforcement guidelines on vertical restraints in 1991. Under these guidelines, a restraint will be found illegal if it lessens competition, a rule-of-reason standard. Where the restraint is imposed by a large manufacturer—one of the three largest or one whose market share is at least 10 percent—the restraint is illegal if it reduces the business opportunities of competitors; this is a stronger, more easily enforced standard applying only to large manufacturers. See Rill 1992, pp. 641–46.

50. Chile, the only Latin American nation with strong antitrust enforcement, has enforced strongly against the use of vertical restraints. There, Firestone was fined for (minimum) resale price maintenance in tires and car batteries; Cantollo y Cia, distributors of SONY electronic products, was sued and fined for refusing to service other brands of electronic products; and Andina de Cosméticos, jointly owned by Revlon and Martini & Rossi, was sued and fined for geographic restraints on distribution. See Bustamante 1992.

51. The debate regarding vertical restraints in the developed economies may suffer from a fallacy of composition. Numerous studies conclude that specific vertical restraints have usually been competitively innocuous, particularly when used by suppliers with little or no market power. The same can be said of horizontal price fixing among suppliers with no market power. Nevertheless, a serious competitive concern may arise when vertical restraints are widely used, as in Japan. This concern arises not from the effect of individual restraints, but rather from the cumulative effect of many restraints.

52. See Crandall, Owen, and Skitol 1991, p. 36; Antimonopoly Department (Slovak Republic) 1992, p. 2; Christoforou 1990 p. 57–58; and Coate, Bustamante, and Rodriguez 1992, p. 13. Only Hungary, where exclusive dealing and refusal to deal are per se illegal, has taken a strong posture against vertical restraints. Moreover, Hungary's competition law further proposes that private parties be allowed to abrogate contracts whenever they believe that the contracts are disadvantageous. Although these legal provisions discourage potentially efficient contracts between manufacturer and distributor, they also create commercial opportunities for the numerous small retailers formed in Hungary during liberalization. See Langenfeld and Blitzer 1991, p. 396.

53. Broadly speaking, vertical restraints offset product uncertainty (on the part of potential buyers), the presence of which can impede the marketing of a product. This impediment is most plausible for a new product or a new supplier. In most cases, a new supplier entering a market is unlikely to be found dominant in that market; similarly, a large firm extending its product line by introducing a new product is unlikely to be found dominant in the market for the new product.

54. Structural regulations focus on several forms of competitive harm. Most often, a horizontal merger reduces competition by eliminating competition between the merging firms, thereby enhancing market power. Alternatively, a vertical merger may create incentives for the merged firm to foreclose former trading partners, a potential abuse of a dominant position. See Boner and Krueger 1991, section V.

55. Similarly, an interlocking director—a high-level executive shared by two independent enterprises—facilitates coordinated conduct between the two enterprises.

56. For example, Greece instituted conduct regulations in 1913. After repeated attempts to incorporate competitive provisions in commercial, trademark, and intellectual property codes, Greece instituted administrative enforcement of merger control provisions in 1977. See Christoforou (1990), p. 49, and Ahdar (1991), pp. 223–24.

57. For example, between 1981 and 1990, the Korean Fair Trade Commission reviewed 2,003 mergers and similar corporate transactions. The commission filed only a single complaint and formally imposed conditions on two other mergers. In another 302 cases, the commission issued informal warnings. See Fair Trade Commission (Republic of Korea) 1992, p. 14. These figures understate the effect of merger control. Private parties have little incentive to undertake mergers that would clearly be illegal and in most cases voluntarily notify the commission of any planned mergers.

58. As a rule, the presence of effective international competitors ensures that a domestic merger will not substantially harm competition. Conversely, high transportation costs can sustain local or regional markets, and trade barriers (tariffs) or structural impediments (national product quality standards, onerous customs inspec-

tion requirements, professional licensing requirements, domestic context requirements, and limits on the foreign ownership of enterprises) can isolate a national market from international competition.

59. Permissive merger control policies may allow domestic producers to achieve the scale and other economies available to large enterprises, thereby enhancing international competitiveness. In addition, small domestic markets may be efficiently served by only a few suppliers, which would recommend that merger control apply only to highly concentrated markets.

60. In the industrial economies, nonhorizontal mergers may raise two competitive concerns: (a) a vertical merger may facilitate the exercise of market power by allowing one of the merging parties to evade cost-plus regulation; and (b) a conglomerate merger may prevent potential competition, even though the merging parties do not presently compete. These concerns arise only rarely. See Boner and Krueger 1991, pp. 83–84.

61. Vertical integration may reduce the transactions costs of an enterprise attempting to secure inputs in the face of inflation and price controls. See Litwack 1991, p. 130.

62. Conglomeration may have played precisely this role in the development of Japan and Korea, through the formation of the *keiretsu* and *chaebol* industrial conglomerates. See Jones and Sakong 1980.

63. See the Mexican Federal Law of Economic Competition, section III; see also White 1990, p. 10. Note that U.S. law offers no formal efficiencies defense, yet enforcers have developed guidelines that do take account of merger-specific efficiencies. See U.S. Department of Justice and Federal Trade Commission 1992.

64. Nearly half of the merger cases before the counsel have involved a failing firm. See Christoforou 1990, p. 72.

65. This treatment of efficiencies is similar to that originally used in merger control and cartelization in Germany, Japan, and the United Kingdom. Over time, competition agencies in these countries have increasingly required evidence that efficiencies would be passed on to consumers or would otherwise encourage lower, rather than higher, prices. In other words, the agencies have been less willing to accept higher prices in one market for potentially lower prices in other markets. See Boner and Krueger 1991, pp. 28–29, 31–32, and 78.

66. Even if a nation constitutes an economic or antitrust market, a private party (to a merger) often will be uncertain about the sales of its competitors. This is particularly so where, as in a reforming economy, commercial information is scarce.

67. The new law of Taiwan (China) uses both methods: It requires premerger notice if (a) the merged firm will have one-third of any market; (b) a merging party accounts for one-fourth of any market; or (c) the annual sales of any merging party exceed the minimum sales amount (presently NT\$2 billion). This last criterion represents an unambiguous requirement that large firms notify the competition agency of any planned mergers. See Liu 1993, pp. 152–54.

Korea provides another variation that avoids ambiguity. Every year, the Fair Trade Commission must designate and notify large enterprises (“market-dominating business concerns” and “large enterprise groups”). This designation and notice imposes a stricter legal responsibility on the notified parties to report corporate transactions and agreements. See Art. 2(4) of Fair Trade Commission (Republic of Korea) 1990. See also Mexico’s Federal Law of Economic Competition, chapter III, Article 20.

68. In this regard, the Slovak requirement—that unapproved mergers are null and void—is adopted from EU merger control. As the European Union has discovered, attempting to reverse, through legal decree, a consummated merger is similar to unscrambling an omelet. It is extremely complex and costly to reverse the many contracts (stock purchases, transfers of assets and property rights, transfers of personnel) that result from a consummated merger. That is one reason for the premerger notification recently required of large transactions in the European Union. See Boner and Krueger 1991, p. 38.

69. Under U.S. law, within 20 to 30 days of being notified of a proposed merger, the enforcement agency must initiate an investigation by requesting additional information from the merging parties. Once the parties have provided this information, the agency has 10 to 20 days to contest the merger. See Boner and Krueger 1991, section 5.

70. This is no casual observation. A nation cannot establish markets and private enterprise without protecting private-property rights. See Litwack 1991, pp. 77–90.

71. More precisely, an anticompetitive merger is likely to result in a reduction of output; any divested supplier should be able to offset this reduction. Similarly, under a dominance analysis, the divested supplier should preserve the premerger commercial opportunities of small trading partners and rivals of the (dominant) merging parties.

72. The expense and corresponding financial risks, delays in construction, and government permits are all commonly cited as entry impediments related to plant construction.

73. Many firms—for example, commercial aircraft, industrial fasteners, and chemicals—produce multiple products from a common, proprietary technology. In this case, it may be impossible to establish a new supplier without divesting the entire technology. Such a divestiture is fairly drastic if competitive concerns are confined to a specific product.

74. If contract enforcement were costless, then an appropriate supply contract might enhance competition. But in *Lowenbrau*, the competitive concern was that a supply agreement could be easily breached, allowing Lowenbrau to remove a competing supplier. Even if contract law were to penalize breach, these penalties would not necessarily restore competition.

75. Rapid and unchecked privatization is usually defended on political grounds. In Russia privatizing without demonopolizing has been seen as a way to accelerate the withdrawal of the state from

commercial matters. This tactic is puzzling, since opponents of reform have used the fear of monopolies to delay market reforms. In my view, this fear is valid, particularly since monopoly is seldom a necessary feature of a market mechanism. See Joskow, Schmalensee, and Tsukanova 1994, pp. 303 and 352–53.

76. "World Bank's Preston Says Hopes Were Too High on Soviet Bloc," *Wall Street Journal*, October 14, 1994, p. A4-D.

77. On these grounds, the German cartel office has opposed the sale of the eastern German airline (Interflug) to Lufthansa, preferring instead a bid from British Airways. See Sinn 1991, p. 32. Similarly, in the Slovak Republic, the law will apply more lenient standards to privatizations involving foreign firms. See Antimonopoly Department (Slovak Republic) 1992, p. 2.

78. A similar problem frequently arises when the approval of a merger is conditioned on a divestiture; there may be relatively few bidders for the divested assets.

79. Similarly, the two plants of Skoda were sold to Volkswagen rather than to two independent entities. Thus, this privatization did not take advantage of the potential competition between the two domestic plants.

80. Bergeron (1991) describes the difficulties encountered by the government of Togo in privatizing the steel mill at Lome. The mill, opened in 1979, suffered from very high manufacturing costs, which prevented its competing (with low-cost European producers) in nearby countries. Ultimately, the mill was leased to Ibson S.A. and operated successfully. Yet this success was achieved only through a statutory monopoly in Togo and a 41 percent tariff on imported steel. Notwithstanding the political advantages, this transaction, in terms of allocative efficiency, was a disaster. Consumer welfare would have been better served if Togo had liquidated the plant (as recommended by the World Bank) and imported its steel products.

81. In Russia proposals to form holding companies through privatization are subject to prior competitive review. See Joskow, Schmalensee, and Tsukanova (1994), pp. 346–47. In Kazakhstan the holding companies, in which the state typically retains a majority share, were formed to prevent the theft of state-owned property. Theft may have resulted from a recent law allowing state-owned enterprises (and their managers) to establish private enterprises. See "On the Protection and Support of Private Entrepreneurship," *Review of Central and East European Law*, pp. 179–94.

Sector-wide holding companies are sometimes touted as a desirable imitation of Japanese and Korean corporate practice. Yet even Japan and Korea did not allow horizontal integration on the level permitted in Kazakhstan. Moreover, Korea recently took steps to limit cross-ownership and shrink its holding companies through compulsory divestiture. See R. Boner 1994a and Fair Trade Commission 1992. In Kazakhstan the creation of administrative and management formations—including the state holding companies—requires the approval of the Antimonopoly Committee.

Unfortunately for Kazakhstan, this provision has not been enforced, and most of the largest holding companies were registered without the committee's approval.

82. Recently, to prevent and correct anticompetitive privatizations, the government of Argentina was advised to authorize a (new) Competition Court to review proposed privatizations and to dissolve already privatized enterprises. See Economists Incorporated 1992, p. 48.

83. In addition, highly concentrated industries may be more successful in resisting reforms and inhibiting economic transition. For example, the problems of inflation and oversubsidization in Russia may have been exacerbated by the formation of large, regional agro-industrial groups. See Shleifer 1994, pp. 375–76.

84. See Seidman 1989, Brazil section, p. 31, and Philippines section, p. 15; Article 7 of the (Mexican) Federal Law of Economic Competition; and Langenfeld and Blitzer 1991, p. 382.

85. Former Soviet economies often use price controls; see Langenfeld and Blitzer 1991, p. 382. The competition law of Kazakhstan authorizes the Antimonopoly Committee to identify dominant firms—those whose market share is no less than 35 percent—and to impose legal limits on profits and (implicitly) on prices. Violations result in fines that are paid into the republican budget. In addition, where privatizing an enterprise raises competitive concerns that cannot, for technological reasons, be solved through divestiture, the law calls for price controls after privatization. See "On the Development of Competition and the Restriction of Monopolistic Activity," *Almaty*, June 1991.

86. Among other shortcomings, price controls undermine the capacity of prices to reflect the (comparative) need for investment across sectors. That is, controls undermine the informational and allocative role of prices.

87. In the former procedure, prices are calculated so as to allow an enterprise to realize a competitive rate of return, similar to public utility rate regulation. In the latter, prices are set to be similar to those in a market with a comparable structure. For details, see Schmidt 1983.

88. Even accurate accounting data does not allow one to calculate the economic rate of return to which entry and exit respond for a specific industry. This exercise was applied and rejected in *U.S. v. IBM*.

89. Other characteristics of the reforming economy may make price controls unworkable. Some reforming economies—Kazakhstan, Russia, and Ukraine—suffer from high inflation or even hyperinflation, making it difficult to implement price controls. In addition, price controls can be evaded. Firms on the monopoly registers of Russia and Kazakhstan, by virtue of a 35 percent market share, can be subjected to price controls. Yet many of them are vertically integrated and can therefore evade price regulation by buying their inputs at unregulated and inflated prices from affiliates (similar to the tactics that led to the dissolution of AT&T). See R. Boner 1994b.

90. To encourage allocative efficiency, the use of price regulation should be limited to natural monopolies—that is, to markets whose size is approximately equal to the minimum efficient scale of a supplier. Assuming that price regulation is even moderately effective, it can then be used to correct the resulting allocative distortion. Conversely, if regulation is imperfect, it may be better to allow competition between multiple suppliers that, being small, are technically inefficient but, due to competition, offer lower prices and higher output than would an imperfectly regulated monopolist. See L. Boner 1993.

In many cases, price controls are more a product of nostalgia than of efficiency. For example, in Kazakhstan the sole vodka producer of Almaty oblast is subject to price controls even though one can buy imported vodka from Bulgaria, Greece, Russia, and other countries. Obviously, the local vodka plant is neither dominant nor a pure monopolist.

91. *BMI* is the classic American example; also see Jorde and Teece 1989.

92. For example, Russian courts demand documentary proof of price fixing and other forms of cartelization. But price fixing does not require explicit contracts. Thus, this procedure emasculates enforcement against price fixing. See Joskow, Schmalensee, and Tsukanova 1994, p. 362.

93. Consider Article 6(2) of the Ukrainian Law on Containing Monopolism and Preventing Unfair Competition 1992: “Exceptions to . . . this article may be instituted by other legislative acts of Ukraine for the purpose of safeguarding national security and defense or social interests.” What, one may wonder, are the relevant “social interests”?

94. I have omitted the United Kingdom from this list. Even though the United Kingdom provides for judicial oversight of administrative agencies, British courts tend to respect the decisions of regulators. See Prichard 1988, pp. 462–63 and 469–70. Hong Kong, Singapore, and (prior to 1992) Taiwan (China) have no national competition law but do allow private enforcement against commercial torts, including several forms of “unfair competition.” See Seidman 1989, appropriate national sections.

95. Common law is defined by its explicit reliance on precedent; in this system, statutory language is generally vague, and judges develop more detailed legal standards through repeated application of a statute. In contrast, civil code systems are characterized by detailed statutes that leave far less discretion to judges. Most developing nations employ a civil code system; some (present or former members of the British Commonwealth) employ common law. In practice, most legal systems are hybrids, and even civil code systems develop commercial law through the use of precedent. As a rule, competition statutes are amended more frequently in civil code systems than in common law systems. For example, the major U.S. competition statutes have seldom been amended by the Congress, whereas Germany’s Act Against Restraints on

Competition was amended six times during the 1980s. See Boner and Krueger 1991, sections II and III.

96. Landes and Posner use legal citation to estimate the useful life of a legal precedent. They find that the legal precedents of U.S. commercial law (that is, labor, contracts, antitrust, regulatory, and others) depreciate at an annual rate of approximately 4 to 5 percent.

97. The flexibility of administrative agencies may be reduced by legal constraints on administrative conduct. In the United States, constitutional protections of due process significantly restrain administrative agencies, and the Administrative Procedures Act enhances these protections by ensuring, in a variety of ways, public access to agency decisionmaking. See the Administrative Procedures Act, 5 U.S.C.A. ss. 551–706.

98. For a good expression of these concerns, see Godek 1991, which proposes that economic development may be impeded by an antitrust agency “run amok.” For this and other reasons, Godek recommends against antitrust policy for the nations of Eastern Europe.

99. This occurred in Australia, France, Korea, New Zealand, and after their competition laws were enacted or amended. See Ahdar 1991, p. 220; Fair Trade Commission (Republic of Korea) 1992, section IV 4.8.; and Trade Practices Commission (Australia) 1990, chapter 4.

100. The Supreme Court of the Philippines, consisting of 15 judges, does not generally rule *en banc* (as a single body). Instead, decisions are issued by five-member “divisions.” There are two difficulties with this structure. First, to establish a voting majority, a litigant need influence only three judges instead of eight; thus, this structure may facilitate corruption. Second, nothing prevents the various divisions from issuing conflicting rulings, and the constitution requires that the court sit *en banc* to reverse itself. This structure impedes the court’s primary role, which is to reconcile the conflicting rulings of the lower courts. See the Constitution of the Philippines, Article VIII, section 4(3).

A similar structure has been recommended for Argentina’s new Competition Court, whose decisions would be made by three-member panels. See Economists Incorporated 1992, pp. 36–37. Because decisions of that court could be appealed to Argentina’s administrative courts, inconsistent decisions would likely be corrected. This cannot be said of Supreme Court decisions in the Philippines.

101. Private parties can appeal negative findings by the commission or the secretary of commerce to the civil courts and, where the commission has established legal liability, can also sue in the civil courts for compensatory damages. See Cabanellas and Etzrodt 1983, p. 40.

102. See Competition Protection Act No. 63/1991, Part IV (Proceedings of the Authority) and Part VI (Disputes Arising from Unfair Competition).

103. Out-of-court settlements are available in Argentina, France, the Philippines, Taiwan (China), the United Kingdom, and the United States.

104. For example, the rule-of-reason standard reflects the complex efficiency effects of vertical restraints. Where vertical restraints would enhance the technical efficiency of an enterprise, the enterprise has an incentive to litigate (to defend this efficient practice) if the efficiencies outweigh the unreimbursed costs of litigation. Thus, one would expect that Hungary's (unusual) per se prohibition of vertical restraints is likely to be weakened as private parties litigate to defend efficient uses of these restraints. This has occurred in the United States with respect to the per se prohibition of resale price maintenance. See Langenfeld and Blitzer 1991, p. 396, and Coate, Bustamante, and Rodriguez 1992, p. 76.

105. Collateral repossession can take as long as five years under Bolivia's *proceso ordinario*, the legal process that typically applies to civil contracts. See Fleisig and others 1991, p. 21; see also Seidman 1989, Brazil section, p. 33.

106. Legal uncertainty results in part from informational asymmetry between plaintiff and defendant. This uncertainty, and therefore the private motive to litigate, can disappear when the litigants exchange information prior to trial. For example, the defendant to a predation (low pricing) allegation can show to the plaintiff that low prices are justified by low costs. Under these circumstances, a judicial investigation and finding would serve no purpose.

107. For example, the Competition Commission of Greece denied an exemption to selective distribution by Toyota. This ruling, based on the notion that Toyota spare parts constituted a valid antitrust market, disadvantaged a firm with a long history of aggressive competition. In addition, the ruling represents an unusual antitrust principle (to say the least) and is soundly criticized by Christoforou 1990, p. 62. Yet this finding parallels an EU precedent, *Hugin*, in which the European Commission found that spare parts to Hugin typewriters constituted a valid product market. See Commission of European Communities, "Decision on Hugin v. Lipton," *Official Journal*, L22, 1977, pp. 23–35. Note also the similarity to the recent U.S. Supreme Court case involving Kodak.

108. Private rights also influence the mix of cases. In Russia, which offers no private right of action, enforcement has focused on the regulation of "monopolists" (defined by a 35 percent market share) rather than the creation and preservation of competition. In 1993 the antimonopoly committees litigated 1,067 cases, 70 percent of which involved monopoly registration. Less than 1 percent of the cases involved cartels and price fixing, and none involved the dissolution of monopolists. It is doubtful that private actions would have generated this caseload. Monopoly registration was discontinued at the end of 1993. See Joskow, Schmalensee, and Tsukanova 1994, p. 359.

109. See Landes and Posner 1976, pp. 249–307. They treat legal precedents as a form of informational capital, the formation and depreciation of which (they find) follow rules derived from the economic theory of investment and capital formation.

110. Commercial torts have been addressed under Greek law since 1913, whereas enforcement by an administrative agency dates only to 1977. See Christoforou 1990, p. 49. Taiwan (China), of course, recently instituted administrative enforcement by a new Fair Trade Commission. See Liu 1993.

111. Administrative agencies do not necessarily have the same incentives as private complainants. For example, a recent lengthy investigation by the U.S. Department of Justice—and earlier by the Federal Trade Commission—involved alleged exclusion (of competitors) by Microsoft. Because the alleged conduct would directly affect a relatively small number of software suppliers, the matter could efficiently have been left to private litigation. Instead, it was addressed through an administrative investigation that ultimately resulted in a consent order. Although the order was widely condemned as ineffective by the software industry, software suppliers were unwilling to invest their own resources to sue Microsoft. Given the administrative and private incentives to undertake a costly litigation, this outcome is perhaps not surprising. See "Microsoft: Not Guilty by Reason of Reality," *The Economist*, July 23, 1994, pp. 62–63.

112. Suppose that overt (and easily proved) price fixing results in a price premium of \$5 for each of 10,000 private parties. If the unreimbursed cost of a private suit to enjoin the conduct and collect individual damages is \$100, then no individual has an incentive to bring a lawsuit, even though such a suit would raise social welfare (an expenditure of \$100 prevents antitrust damages of \$50,000). Absent a collective enforcement action (involving an injured class of no fewer than 20 parties), there is no legal deterrent to the conduct. See Posner 1977.

113. The cost and fee rules applying to litigation may significantly chill the incentives of private parties to engage in class action suits. For example, whereas class actions are relatively common under U.S. law, the cost and fee-sharing rules of the United Kingdom ensure that private parties almost never have an incentive to participate in a class action suit. See Prichard 1988, pp. 457–59.

114. The U.S. Federal Trade Commission has been criticized at times for enforcing dominance-style laws such as the Robinson–Patman Act, which is designed to protect *competitors*. See Bork 1978, who proposes that enforcement of this act chills efficient price competition. Notice also that the enforcement of competition laws protecting *competitors* (not the competitive process) may often be left to private parties. These criticisms may explain why the Federal Trade Commission has undertaken few Robinson–Patman enforcement actions since the 1970s.

115. Argentina, Britain, France, Germany, the Republic of Korea, and the United States (Department of Justice, Antitrust Division) all operate competition agencies within a larger, cabinet- or ministry-level agency.

116. In the Philippines, the Board of Investments has primary authority to enforce competition law. In practice, the board has

emphasized regulatory initiatives—for example, capacity licensing and industrial consolidation—that have harmed competition. For a list of the industrial policies conducted by the board, see U.S. Agency for International Development 1992, table 4.1.

Brazil's 1988 constitution formally recognizes the domestic market as a national asset, access to which may be restricted on the grounds of socioeconomic development, public welfare, or technological autonomy. Naturally, this provision tends to undermine competition in Brazil. See Seidman 1989, Brazil section, pp. 1, 31–33. 117. Enforcement of competition law may be weakened due to conflicts with other agency mandates. For example, the competition provisions of the (Philippine) Price Act are enforced by the (four) Departments of Health, Agriculture, Trade and Industry, and Environmental and Natural Resources, none of which is structured as an enforcement agency. See Republic Act No. 7581, Congress of the Philippines, Manila, July 22, 1991.

In Brazil, the Administrative Council of Economic Defense (CADE) is responsible for enforcing competition law, yet other agencies impose competition-style prohibitions (presumably on noncompetition grounds); for example, the Institute of Industrial Property enforces a *per se* prohibition against territorial restrictions on franchises (a vertical restraint). See Seidman 1989, Brazil section, pp. 1, 31–33.

118. By design, antidumping and countervailing duty laws protect the interests of domestic producers by *weakening* competition from foreign producers. See Boner and Krueger 1991, pp. 100–06.

Weak enforcement is a likely result of competition laws designed to pursue a variety of “public interest” criteria (for example, regional employment). At times, this has plagued enforcement in Britain and Sweden. See Boner and Krueger 1991, pp. 35, 51–53, and 77, and Crandall, Owen, and Skitol 1991, p. 9. Nevertheless, competition and trade policies can be coordinated, with trade liberalization aimed at concentrated sectors (as in Australia, Canada, Korea, and Poland). See Langenfeld and Blitzer 1991, p. 25.

119. Fundamentally, market-oriented reform rests on a conviction that markets usually outperform regulatory mechanisms such as price controls. See the earlier section in this chapter on dominance, growth, and efficiency.

120. Note that Kazakhstan recently combined its Antimonopoly Committee and its Pricing Committee.

121. The efficiency of a market depends greatly on the information available to buyers in that market. If information is absent or distorted, a market process will break down or operate inefficiently. Thus, consumer protection policy, by ensuring that buyers are adequately informed, increases the efficiency of the market mechanism (in the same way as does antitrust policy). See Boner and Krueger 1991, p. 57; R. Boner 1994c; and Akerlof 1970.

122. To illustrate, Anne Bingaman, the chief antitrust official at the U.S. Department of Justice, recently teased private antitrust attor-

neys, “My job is to put your kids through college.” See “Happy Warriors: NASDAQ Investigation Showcases New Moxie at Justice Department,” *Wall Street Journal*, October 20, 1994, p. A1.

Only a select few private attorneys can benefit if a competition agency focuses its efforts and resources on a small number of litigated cases. Such enforcement is thus inconsistent with regulatory capture. Alternatively, diffuse enforcement—over numerous investigations that are seldom litigated—conveys equally diffuse benefits on the members of the private antitrust bar. This style of enforcement is more consistent with the capture hypothesis.

A draft law for Venezuela contained the requirement that the professional staff of its competition agency be limited to attorneys. See White 1990, p. 15. Given the concerns regarding regulatory capture, and because economic and commercial expertise is highly relevant to competition matters, this requirement seems very ill-advised.

123. Where agency decisions are subject to judicial review, a captured agency would be less successful in litigation, particularly at the appellate level. Examining evidence on litigation by the U.S. Federal Trade Commission, Posner rejects the capture hypothesis. Historically, the commission has been highly successful in litigating on appeal; this success would not be expected were the agency captured by a private interest group.

124. For example, three of the five members, including the chairman, of Argentina's Comisión Nacional are economists, as is the superintendent of Venezuela's new competition agency. Jamaica was advised to draw its competition agency executives from law, accounting, and economics. See Crandall, Owen, and Skitol 1991, p. 29. In contrast, a draft law for the Philippines would require that the agency executive be a lawyer. This requirement seems ill-advised since regulatory capture has been an ongoing problem in the Philippines. See Bill No. 7011, House of Representatives, Quezon City, February 2, 1993.

125. The incentive to violate the law may be weakened, at least for clearly egregious violations, where, as in the United Kingdom, the loser is required to bear the litigation expenses of the winner. Competition litigation in Hong Kong (based on contract and commercial tort law) has adopted this feature, with losers required to bear between one-third and one-half of the litigation expenses of winners. See Prichard 1988, p. 455. See also Seidman 1989, Hong Kong section, p. 8.

126. In the Philippines criminal penalties apply to conduct restricting the availability of any food substance, motor fuel, or any “article of prime necessity” (Seidman 1989, p. 19). Civil suits, but not criminal suits, may be settled out of court. In addition, the state may expropriate any property possessed under any contract or combination found to restrain trade. This extremely severe penalty would not be feasible in the United States, where the Fifth Amendment to the Constitution prohibits uncompensated takings of private property by the state. See also Crandall, Owen, and Skitol 1991, p. 47.

127. It is sometimes recommended that repeat violations attract severe (or criminal) penalties. See, for example, Economists Incorporated 1992, p. 5. This recommendation is intuitively appealing and may be necessary as a matter of law. Nevertheless, repeat violations may indicate simply that the benefits to the violator exceed the damages to the injured party, notwithstanding any earlier findings to the contrary. This consideration is frequently encountered in discussions of legal liability. See Rubin 1977, pp. 52–53. Thus, even the “obvious” candidates for severe penalties must be examined for potential efficiencies.

128. One can often find legal prohibitions that, though defined in a statute under a rule of reason, are essentially enforced under a per se standard (and conversely). The important question is whether the effective standard of legality is well understood by private parties. High punitive damages are appropriate only if legal standards are well understood by private parties.

129. Detection is more difficult if these forms of conduct are treated as an abuse of dominance, since evaluating dominance may require an in-depth analysis.

130. See Godek's (1991) criticisms of U.S. enforcement against resale price maintenance, which is both per se illegal and subject to treble damages.

131. Under the old Mexican law, punitive damages were limited to \$250 (U.S. dollars). The new law substantially raises penalties. For example, some maximum penalties are as follows: for submitting misleading information to an investigation, \$34,000; for engaging in an absolute (that is, per se illegal) monopolistic practice, \$1.7 million; and for engaging in a relative (that is, rule of reason) monopolistic practice, \$1.02 million. The new Mexican law defines financial penalties as multiples of the general minimum wage prevailing in the Federal District of Mexico City. So defined, penalties are automatically indexed for inflation. See the Federal Law of Economic Competition, chapter VI.

132. If professional managers were exempt from criminal prosecution, then the financial penalties assessed an enterprise for, say, price fixing would be paid only by the owners. Of course, if the manager is the owner, properly designed financial penalties alone can be sufficient to deter illegal conduct.

133. An emerging market economy is commonly advised that prosecution against horizontal price fixing should be the highest priority of the competition agency. This advice makes sense because horizontal price fixing is so seldom conducive to either efficiency or development. However, if the primary role of the agency is to develop a commercial code of conduct, the agency should describe not only what is illegal but also what is legal. The Fair Trade Commission of Taiwan (China) discovered this quickly, as statutory prohibitions had to be clarified to allow (common and competitively innocuous) contracts between prime contractors and subcontractors in the construction trades. See Liu 1993.

134. See Hawk 1979, p. 217; see also Cabanellas and Etzrodt 1983, p. 34.

135. In the Philippines amendments to treat antitrust violations under civil process date to the 1970s. These amendments have not been enacted. See U.S. Agency for International Development 1992, note 6, p. 79.

136. In some cases, aggressive enforcement may result from a learning curve, as over time an agency accumulates commercial information and an experienced, well-informed staff able to utilize that information. See Scherer 1990.

137. In the industrial nations, private lawsuits address price fixing, and (compensatory and punitive) damages are paid to victims. In contrast, in Kazakhstan “price gouging” can be addressed only through administrative enforcement, and penalties are paid not to victims but to the state budget. Given the incentives, one might wonder whether such actions are motivated by budgetary rather than competitive concerns.

138. For a number of years, the U.S. Federal Trade Commission routinely issued public comments on the economic harm caused by antidumping measures implemented by the U.S. International Trade Commission; these studies were frequently ignored.

139. The competition statutes of Kazakhstan and the Czech Republic allow the competition agency to vacate the actions of other state bodies.

140. For example, the de facto interpretation of legal statutes may vary across nations. In Western economies, conduct not expressly prohibited by some statute is assumed to be legal. But in Kazakhstan interviews revealed that conduct not expressly permitted under a statute may be assumed by private parties to be illegal. This interpretation, puzzling to a Westerner, makes perfect sense if the private party recognizes that only those rights that have been expressly permitted can easily be defended against intrusion by the state.

## References

- Ahdar, R. J. 1991. “American Antitrust in New Zealand.” *Antitrust Bulletin*. Spring:217–47.
- Akerlof, G. 1970. “The Market for ‘Lemons’: Quality, Uncertainty and the Market Mechanism.” *Quarterly Journal of Economics* 84: 488–500.
- Antimonopoly Department (Slovak Republic). 1992. “Fundamental Issues, Organizational Structure: Competition Protection Act.”
- Azcuenaga, M. L. 1992. “The Evolution of International Competition Policy: A Federal Trade Commission Perspective.” Remarks before the 19th Annual Fordham Corporate Law Institute, Fordham University, Bronx, New York, October.
- Bergeron, I. 1991. “Privatization through Leasing: The Togo Steel Case.” In Ravi Ramamurti and Raymond Vernon, eds., *Privatization and Control of State-Owned Enterprises*. Washington, D.C.: World Bank.

- Boner, L. H. 1993. "Privatizing the Power Sector: Regulatory Alternatives." Remarks before the Eastern Economic Association, Washington, D.C., March.
- Boner, R. A. 1994a. "Industrial Structure and Private Sector Development." Paper presented at the Conference on Competition Policy and Market Reform, Almaty, Kazakhstan, May 31–June 1.
- . 1994b. "Demonopolization through Divestiture." Paper presented at the Conference on Competition Policy and Market Reform, Almaty, Kazakhstan, May 31–June 1.
- . 1994c. "Consumer Protection and Advertising Fraud." Paper presented at the Conference on Competition Policy and Market Reform, Almaty, Kazakhstan, May 31–June 1.
- Boner, Roger Alan, and Reinald Krueger. 1991. *The Basics of Antitrust Policy: A Review of Ten Nations and the European Communities*. World Bank Technical Paper 160. Washington, D.C.
- Boner, R. A. and J. Langenfeld. 1992. "Liberal Trade and Antitrust in Developing Nations." *Regulation* 15: 5–6.
- Bork, R. 1978. *The Antitrust Paradox: A Policy at War with Itself*. New York: Basic Books.
- Bustamante, R. 1992. "Structural Reforms in Third World Countries." U.S. Federal Trade Commission, Washington, D.C.
- Cabanelas, G., and W. Etzrodt. 1983. "The New Argentine Antitrust Law: Competition as an Economic Policy Instrument." *Journal of World Trade Law* 17: 34–53.
- Caulkins, Stephen. 1993. "Supreme Court Antitrust 1991–1992: The Revenge of the Amici." *Antitrust Law Journal* 61: 269–311.
- Christoforou, T. 1990. "Greek Law on Competition: An Analysis of Twelve Years' Case Law." *World Competition* 14: 49–77.
- Coate, M. B., R. Bustamante, and A. E. Rodriguez. 1992. "Antitrust in Latin America: Regulating Government and Business." Paper presented at the annual meeting of the American Agricultural Economics Association, August.
- Crandall, R. W., B. M. Owen, and R. A. Skitol. 1991. "Report of the Advisory Team: Competition Policy in Jamaica." U.S. Agency for International Development, Washington, D.C.
- Economists Incorporated. 1992. "Report of the Advisory Team on Competition Policy and Consumer Protection in Argentina." Washington, D.C.
- Evans, D. S. 1987. "Tests of Alternative Theories of Firm Growth." *Journal of Political Economy* 95: 657–74.
- Fair Trade Commission (Republic of Korea) 1990. *Monopoly Regulation and Fair Trade: Acts and Enforcement Decrees*. Seoul: Republic of Korea.
- . 1992. *Competition Policy in Korea: The First Ten Years*. Seoul: Republic of Korea.
- Fleisig, H., and others. 1991. "Law, Legal Procedure, and the Economic Value of Collateral: The Case of Bolivia." World Bank, Washington, D.C.
- Fornalczyk, A. 1992. "Competition Law and Policy in Poland: A Welcome and a Warning for International Business." Remarks before the International Bar Association, Third Eastern European Regional Conference, Budapest, June 21.
- Godek, P. 1991. "Antitrust Will Stifle, Not Spur, Eastern Growth." *Wall Street Journal*, European edition, July 26–27, p. 8.
- . 1992. "One U.S. Export Eastern Europe Does Not Need." *Regulation* 20:19–22.
- Hahn, F., ed. 1990. *The Economics of Missing Markets, Information, and Games*. New York: Clarendon Press.
- Hawk, B. 1979. "International Antitrust." Remarks before the Fifth Annual Fordham Corporate Law Institute, Fordham University, Bronx, New York.
- Jones, L. P., and I. Sakong. 1980. "Government, Business, and Entrepreneurship in Economic Development: The Korean Case." In *Studies in the Modernization of the Republic of Korea, 1945–1975*. Cambridge, Mass.: Harvard University Press.
- Jorde, T. M., and D. J. Teece. 1989. "Acceptable Cooperation Among Competitors in the Face of Growing International Competition." *Antitrust Law Journal* 58:529–56.
- Joskow, P. L., R. Schmalensee, and N. Tsukanova. 1994. "Competition Policy in Russia during and after Privatization." *Brookings Papers on Economic Activity: Microeconomics 1994*. Washington, D.C.: Brookings Institution.
- Klein, B., and K. M. Murphy. 1988. "Vertical Restraints as Contract Enforcement Mechanisms." *Journal of Law and Economics* 31: 265–97.
- Kovacic, W. E. 1992. "Competition Policy, Economic Development, and the Transition to Free Markets in the Third World: The Case of Zimbabwe." *Antitrust Law Journal* 61:253–65.
- Landes, W. M., and R. A. Posner. 1976. "Legal Precedent: A Theoretical and Empirical Study." *Journal of Law and Economics* 19: 249–307.
- Langenfeld, James, and Marsha W. Blitzer. 1991. "Is Competition Policy the Last Thing Central and Eastern Europe Need?" *American University Journal of International Law and Policy* 6: 347–98.
- Lilley, P. 1991. "Peter Lilley Speaks on UK Merger Policy." *International Merger Law: Events and Commentary*. No. 13. Washington, D.C.: Washington Regulatory Reporting Associates.
- Litwack, J. M. 1991. "Legality and Market Reform in Soviet-type Economics." *Journal of Economic Perspectives* 5: 77–89.
- Liu, L. S. 1993. "In the Name of Fair Trade: A Commentary on the New Competition Law and Policy of Taiwan, The Republic of China." *The International Lawyer* 27: 145–67.
- Lucas, Jr., R. E. 1986. "On the Mechanics of Economic Development." Chung-Hua Lectures, No. 12, Republic of China.

- Murrell, P. 1991. "Can Neoclassical Economics Underpin the Reform of Centrally Planned Economies?" *Journal of Economic Perspectives* 5: 59-76.
- North, D. C. 1994. "Economic Performance through Time." *American Economic Review* 84: 359-68.
- Pengilly, W. 1983. "Comparative Approaches to the Enforcement of Antitrust Laws against Price-Fixing Arrangements (with special emphasis on the lessons to be learned from antitrust law and enforcement in Australia)." *Antitrust Bulletin* 28: 883-939.
- Pittman, Russell. 1992. "Merger Law in Central and Eastern Europe." Working Paper EAG 92-2. U.S. Department of Justice, Economic Analysis Group, Washington D.C.
- Posner, R. A. 1977. *Economic Analysis of Law*. 2d ed. New York: Little, Brown & Co.
- Prichard, J. R. S. 1988. "A Systematic Approach to Comparative Law: The Effect of Cost, Fee, and Financing Rules on the Development of Substantive Law." *Journal of Legal Studies* 17: 451-75.
- Ramseyer, J. M. 1989. "Water Law in Imperial Japan: Public Goods, Private Claims, and Legal Convergence." *Journal of Legal Studies* 18: 51-77.
- Rey, P. and J. Tirole. 1986. "The Logic of Vertical Restraints." *American Economic Review* 76: 921-39.
- Rill, J. F. 1992. "Competition Policy: A Force for Open Markets." *Antitrust Law Journal* 61: 637-50.
- Rodriguez, A. E., and M. D. Williams. 1994. "The Effectiveness of Proposed Antitrust Programs for Developing Countries." *North Carolina Journal of International Law* 19: 209-32.
- Rubin, P. H. 1977. "Why Is the Common Law Efficient?" *Journal of Legal Studies* 6: 51-63.
- Scherer, F. M. 1990. "Sunrise and Sunset at the Federal Trade Commission." *Administrative Law Review* 42: 461-87.
- Schmidt, Ingo. 1983. "Different Approaches and Problems of Dealing with Market Power: A Comparison of German, European, and U.S. Policy toward Market-Dominating Enterprises." *Antitrust Bulletin* 28: 417-60.
- Schuck, P. H., and R. E. Litan. 1986. "Regulatory Reform in the Third World: The Case of Peru." *Yale Journal on Regulation* 4: 51-78.
- Seidman, P. Z., ed. 1989. *Survey of Foreign Laws and Regulations Affecting International Franchising*. 2d ed. Chicago: American Bar Association.
- Shleifer, A. 1994. "Comments and Discussion." *Brookings Papers on Economic Activity: Microeconomics 1994*. Washington, D.C.: Brookings Institution.
- Sinn, H. W. 1991. "Macroeconomic Aspects of German Unification." Working Paper No. 3596. National Bureau of Economic Research, Cambridge, Mass.
- Svejnar, Jan. 1991. "Microeconomic Issues in the Transition to a Market Economy." *Journal of Economic Perspectives* 5: 123-38.
- Trade Practices Commission (Australia). 1990. *Annual Report, 1989-90*. Canberra: Australian Government Printing Service.
- U.S. Agency for International Development. 1992. "Barriers to Entry Study." Consultant's report. SGV Consulting, Washington, D.C.
- U.S. Department of Justice and Federal Trade Commission. 1992. "1992 Horizontal Merger Guidelines." U.S. Government Printing Office, Washington, D.C.
- Willig, R. 1991. "Antimonopoly Policies and Institutions." In C. Clague and G. C. Rausser, eds., *The Emergence of Market Economies in Eastern Europe* London: Blackwell.
- Winter, R. A. 1993. "Vertical Control and Price versus Non-price Competition." *Quarterly Journal of Economics* 108: 61-76.
- World Bank. 1987. *Korea: Managing the Industrial Transition*. vol. 2. Country Study. Washington, D.C.
- . 1993. "Philippines Private Sector Assessment." Report No. 11853-PH. Washington, D.C.
- Yoshikawa, S. 1983. "Fair Trade vs. MITI: History of the Conflicts between the Antimonopoly Policy and the Industrial Policy in the Post-War Period of Japan." *Case Western Reserve Journal of International Law* 15: 489-504.

# Competition issues beyond trade liberalization: distribution and domestic market access

**Mark A. Dutz and Sethaput Suthiwart-Narueput**

*Several years after substantial liberalization of tariff and nontariff barriers to imports, retail prices of various tradable goods had not decreased significantly, as had been expected at the onset of trade liberalization. Policymakers were concerned that certain large manufacturers with market power had integrated forward into domestic distribution and were capturing the differential between international and domestic prices that had been created by the earlier, protectionist trade policies. Government officials feared that consumers, angered that the benefits they had expected to reap from international trade liberalization were instead being captured by powerful business interests, would demand that the government impose price controls. The national competition agency, eager to avoid price controls, believed it should take some action.*

—From conversations with Venezuelan officials<sup>1</sup>

This scenario, which describes a conundrum faced by Venezuelan policymakers in 1993, points to a major issue at the interface between international trade and domestic competition policy: the extent to which barriers to distribution and market access may prevent domestic and international prices from converging. Such barriers to price convergence may be of two types: barriers to within-country, internal trade and barriers to cross-country, international trade. It appears that barriers to internal trade are becoming relatively more important in explaining price differences after international trade liberalization. This has strong policy relevance in reforming economies, where competition problems related to distribution and market access may be particularly severe, thus limiting the potential efficiency gains from liberalization.

The law of one price (LOP) is a basic proposition in economics. It states that within a single market, identical goods must sell at identical prices. Underlying this proposition is the compelling assumption of arbitrage: Market participants have a profit incentive to buy goods where

prices are low and resell where prices are higher, with the arbitrage process continuing until the profit incentive disappears. In environments with no trade frictions (that is, no barriers to entry), incentives for resale will be exhausted only when prices are equalized. International and interregional price linkages are thought to be strongest for standardized, traded goods, for which one firm's product is virtually indistinguishable from that of others (for example, certain raw materials and some intermediate goods such as specific types of steel and chemicals). If goods are differentiated, that is, if the products of competing firms are viewed as distinct by purchasers, LOP is still thought to be a useful guide to the extent that goods are relatively close substitutes. Since international prices are the appropriate shadow prices for traded goods in perfectly competitive markets, deviations from LOP will result in efficiency losses.

In practice, however, markets are not perfectly competitive. For price discrimination to be a viable solution to an enterprise's pricing problem, the enterprise must have some market power, with the ability to sort customers and to prevent resale. In such situations, price discrimination, or LOP violations, not only will be likely but may even be welfare-enhancing. As a simple example, consider the case where one of two distinct markets would not be served if a monopolist were prohibited from charging different prices in the two markets (because the uniform monopoly price over both markets exceeds the willingness to pay of consumers in the second market). In all situations where a new market is opened up because of price discrimination, there typically will be not only a welfare gain but a Pareto-improving welfare enhancement.<sup>2</sup>

For the policy analyst, there are two critical questions: What types of price discrimination, or LOP violations, reduce economic welfare and thus should be discouraged? Is there a useful role for corrective government intervention? Since different pricing outcomes lead to different

transfers of surplus between countries, these answers will depend on which perspective on welfare is taken—that of the world as a whole or that of the importing country. This chapter takes the perspective of national policymakers in a developing country who, like the Venezuelan policymakers in the opening scenario of this chapter, face import prices that may be higher than warranted.

A prior question is how common LOP violations are. Empirical evidence indicates that LOP violations are both widespread and significant in industrial economies. The European Union provides an informative benchmark since economic integration has been actively sought over the past years. But, despite the steps taken to ensure an integrated market, significant price dispersion persists in the European Union in a large number of products. Based on a 1985 Eurostat survey of 20 equipment product headings and 93 consumption product headings (67 goods and 26 services), Emerson and others (1988) calculated that overall dispersion of final end-user prices net of tax (measured by standard deviation) was over 15 percent for consumer goods and over 12 percent for equipment.<sup>3</sup> It should be stressed that this overall dispersion masks some significant absolute price differences for the same product across specific country pairs. For instance, whereas overall dispersion for cars, bicycles, and motorcycles was 14 percent, the absolute difference between Denmark and the United Kingdom was 55 percent. For refrigerators and washing machines, the dispersion of prices net of tax across countries was 10 percent, whereas the largest absolute difference was 39 percent, between France and Italy.

Japan presents perhaps the best-known case of distortions associated with domestic market access, due in large part to publicity associated with U.S. government efforts to increase the access of U.S. companies to Japanese markets. Despite low Japanese formal tariffs and nontariff barriers to imports, consumer prices in Japan for many comparable manufactured goods appear to be much higher than those in the United States. Data collected in 1989 and 1991 (during a period of exchange rate stability) by the U.S. Department of Commerce and Japan's MITI as part of the Structural Impediments Initiative talks revealed that consumer prices of 42 products across broad product classes produced in the United States were on average 65 percent higher in Japan than in the United States (Yager 1993). Conversely, consumer prices of products produced in Japan were on average only 2 percent higher in Japan than in the United States. Finally, prices of products that U.S. manufacturers produced in Japan were on average 24 percent higher in Japan than in the

United States in 1989 and 30 percent higher in 1991. The existence of parallel imports suggests that these price differences are due to some market failure in arbitrage, rather than tariffs or inherently higher transportation and distribution costs.

Few studies have examined price differentials in developing countries. With Argentina's 1991 trade reform, all nontariff barriers to trade were removed (except for the automotive sector), and nominal tariffs for most iron and steel products were reduced to an average of 11 percent. Yet imports did not appear to be providing a sufficient competitive threat to local producers. The prevailing domestic price of hot rolled coils at the end of 1991, for example, was about 35 percent higher than f.o.b. prices in Europe. Reportedly, intimate, nontransparent, and difficult to quantify relations between public enterprises and their suppliers constituted a powerful nontariff barrier.

In Chile, the difference between local retail prices and tariff-inclusive import prices for goods such as butter, cooking oil, and refrigerators ranged from 33 percent to 139 percent. Among the attempts to explain such deviations from LOP, one useful microeconomic model was developed for the case in which final goods are produced by combining tradable goods and commercial intermediation services. Chilean data were shown to be consistent with the hypothesis that variations in retail and wholesale prices of selected imported goods are explained by changes in the cost of domestic distribution (Morande 1983, 1986). While this study showed that price variations were linked to changes in the cost of domestic distribution, the precise nature and causes of these cost differences was not explored.

A recent study by Batista and Mesquita (1994) on the behavior of domestic prices in Brazil suggests that price convergence to international levels is much more likely for some types of products than others. The study examined the inflation-adjusted movement of real domestic wholesale prices from January 1990 to December 1993, a period during which international trade barriers and domestic price restrictions were substantially liberalized. The study revealed two distinct groups of products. For the first group—"readily commercializable" products such as powdered milk, wheat flour, tractors, bicycles, and electric appliances—prices fell 27 percent on average. Prices for the second group—"difficult-to-commercialize" products such as cement, beer, margarine, sausage, ham, powdered detergent, and chlorine—rose 10 percent.

China provides one of the starkest examples of persistent within-country LOP violations. A World Bank study

(1994a) revealed that the price of steel at six different provincial locations in 1987 ranged from ¥1,080 (in Shanxi) to ¥2,000 (Zhejiang) and in 1991 ranged from ¥1,857 (Hebei) to ¥2,435 (Chongqing City). The retail price of 45-centimeter color televisions in 29 provincial capitals in 1986 ranged from ¥1,330 (Beijing, Tianjin, Guiyang) to ¥1,621 (Guangzhou) and in 1991 ranged from ¥1,836 (Guiyang) to ¥3,185 (Harbin). The ex-factory price of urea in 1992 varied from \$89 per ton to \$200 per ton. The within-country price variation between Chinese provinces is often far greater than the cross-country price variation between Shenzhen and Hong Kong, for example, or Shanghai and Tokyo.<sup>4</sup>

These examples should not lead to the inference that international trade liberalization has not had a significant impact on domestic prices. Corbo and McNelis (1989) show how manufacturing pricing rules changed during the trade liberalization process in Chile, Israel, and the Republic of Korea, with external prices becoming more important than domestic cost variables. Still, as illustrated above, important barriers to price equalization remain even after international trade liberalization.

There are several reasons for domestic and international LOP violations, ranging from the presence of multiple effective exchange rates, government-mandated monopoly trading companies, and differential product standards across regions to informational asymmetries and large differences in transportation costs. The price for a good sold to different consumers also may vary due to practices such as price discrimination by firms with market power and vertical restraints imposed by either manufacturing or distribution enterprises. Alternatively, apparent LOP violations may be due only to inappropriately segmented product categories that result in the comparison of goods of different quality. Additional government intervention beyond international trade liberalization is justified in only some of these cases. While a substantial amount of conceptual work has been conducted in several specific areas, such as the treatment of vertical restraints in competition law,<sup>5</sup> it appears that a comprehensive assessment of the causes and policy remedies of LOP violations in reforming economies has not been undertaken.

This chapter develops a framework for evaluating the case for intervention by providing a methodology for identifying the cause of LOP violations and setting out appropriate corrective measures. To what extent are LOP violations a manifestation of serious competition problems, caused by either public policy-related or market power-induced barriers to competition in reforming

economies? It is argued that there is no facile relationship between LOP violations and competition problems related to distribution and market access. Evaluating the case for intervention therefore requires more detailed micro-level studies in order to document the existence and impact of barriers to distribution and market access.

The chapter addresses two basic questions:

- *To what extent are LOP violations due largely to policy, market power, or other causes?* The first section provides some methodological tools for identifying the relevant cause of LOP violations.

- *Is there a case for corrective government intervention?* Whether there is a case for public intervention may depend crucially on the cause for the LOP violation. The second section offers a framework for evaluating whether such intervention is warranted, while the concluding section suggests areas for further research.

### **Identifying causes of LOP violations**

Although many industrializing countries have undergone international trade liberalization over the past years, there is little evidence regarding the extent to which prices faced by industrial and end-use consumers have converged to international levels. There are two dimensions of the problem: within-country differentials (divergence between the price of the same or similar goods—domestic, foreign locally produced, or imported—in different regions within the same national boundaries) and cross-country differentials (divergence between the domestic and out-of-country international price).

Conceptually, any LOP violation can be broken down into these two components, that is, the within-country deviation around the average domestic price, plus the deviation between the average domestic and international price. The distinction is relevant because the policy implications differ. If price differences are due largely to international price discrimination by foreign manufacturers, then domestic competition authorities may have little recourse other than ensuring that undue (for example, legal) restrictions against parallel imports do not exist. For domestic, within-country price differentials, however, the authorities may have access to a much wider range of policy measures to promote competition. The significance of such violations for domestic welfare also remains an unresolved issue.

The fundamental reasons underlying these LOP violations must be assessed carefully if informed policy judgments are to be made. It would be a costly error indeed if, for instance, an overzealous competition agency recommended breaking up a large, efficient manufacturing-

distribution enterprise, if lack of price convergence was caused mainly by multiple exchange rates that prevented effective import competition. Since the case for policy intervention depends on whether lack of price convergence is caused by anticompetitive behavior or simply by higher costs in a competitive market, it is critical to identify the original source of LOP violations.

LOP violations can stem from three distinct and mutually exclusive causes:

- *Public policy.* Even after liberalization of tariffs and quotas, remaining international trade policy measures such as antidumping duties and multiple effective exchange rates allow price divergences to persist. Distribution-specific restraints include government-mandated trading companies or distribution agencies, regulations limiting the number or product range of local distributors, and differential regional taxation.
- *Market power.* Incumbent manufacturers and distributors with market power can maintain higher prices in specific markets by foreclosing entry through such arrangements as sole distributorships and exclusive dealing. Informational asymmetries leading to price differences also may be related to market power, since enterprises in imperfectly competitive markets will seek to manipulate rivals' and customers' information.<sup>6</sup>
- *Other causes.* LOP violations can occur and persist because of large differences in transportation costs, the perishable nature of certain products, or can be caused indirectly due to capital market imperfections, among other factors.<sup>7</sup>

#### *Market power versus other causes*

As a first cut toward establishing a case for further intervention, cases of LOP violation that are due to causes other than market power must be identified. Since manufacturer or distributor markups are typically not directly observable, large positive differences between domestic and tariff-inclusive world prices are *not* sufficient to indicate the presence of market power in distribution. Such differences could indicate, for example, high distribution or transportation costs endemic to the economy (for example, due to a geographically fragmented market) rather than the presence of market power.

To illustrate this point more formally and develop a methodology for identifying the presence of market power, we construct the following decomposition of prices.<sup>8</sup> The domestic price of tradable product  $j$  at time  $t$ —denoted  $P_{jt}$ —can be decomposed into its specific elements as follows,

$$(4.1) \quad P_{jt} = P_{jt}^* T_{jt} E_{jt} G_{jt}$$

where  $P_{jt}^*$  is the world price in foreign currency,  $T_{jt}$  is one plus the tariff or tariff-equivalent rate, and  $E_{jt}$  is the appropriate effective exchange rate.  $G_{jt}$  is a residual that includes “everything else.” We assume that this residual takes the form

$$(4.2) \quad G_{jt} = M_{jt} D_{jt} C_{jt}$$

where  $M_{jt}$  and  $D_{jt}$  are the additional manufacturer and distributor markups, respectively, and  $C_{jt}$  is the transportation and distribution cost defined in markup terms.  $M_{jt}$  could differ from one if the manufacturer practices international price discrimination. In general, the terms comprising  $G_{jt}$  will be unobservable.

Using lower-case letters to denote (natural) logarithms, equation (4.1) becomes

$$(4.1)' \quad p_{jt} = p_{jt}^* + t_{jt} + e_t + g_{jt}$$

where we have assumed that all sectors face the same exchange rate,  $e_{jt} = e_t$ . If the international price  $p_{jt}^*$  is known, then the residual  $g_{jt}$  can be computed directly. However, large values of  $g_{jt}$  could be due to large values for the transportation and distribution costs  $c_{jt}$  rather than large markups.

Suppose it is possible to segment products into certain distinct categories  $J$  in which it is likely that  $c_{jt} \approx c_t^j$  (for example, different types of apparel are likely to have similar transportation and distribution costs). Since price differences also stem from differences in quality, we must be careful about segmenting only similar products since differences in prices due to variations in product quality are not considered an LOP violation per se, but are symptomatic of inappropriately segmented product categories. Then for products  $j$  and  $k$  in category  $J$ , we can compute the difference in residuals,  $g_{jt} - g_{kt}$ , which from (4.1)' is exactly equal to

$$(4.1)'' \quad (p_{jt} - p_{kt}) - (p_{jt}^* - p_{kt}^*) - (t_{jt} - t_{kt}) \approx (m_{jt} - m_{kt}) + (d_{jt} - d_{kt})$$

where the approximate equality follows from our product segmentation. Products with high  $g_{jt} - g_{kt}$  in each category have high relative markups. For these products, the LOP violation may indeed be due to market power.

Alternatively, if the international price  $p_{jt}^*$  is not exactly known but data over time are available, we could com-

pute the change in the price residual over time,  $\Delta g_{jt} \equiv g_{jt} - g_{jt-1}$ , which is equal to

$$(4.3) \quad \Delta p_{jt} - \Delta p_{jt}^* - \Delta c_{jt} = \Delta m_{jt} + \Delta d_{jt} + \Delta c_{jt}$$

where  $\Delta x_{jt} = x_{jt} - x_{jt-1}$ . Since percentage changes approximate log differences, we can substitute changes in the foreign producer price index (PPI) for  $\Delta p_{jt}^*$ . If transportation and distribution costs do not change much over the chosen time period ( $\Delta c_{jt}$  is small), then large values of  $\Delta g_{jt}$  indicate large changes in markups, and again an LOP violation is possibly induced by market power.<sup>9</sup>

The above analyses, however, are not without problems. A consistently high  $g_j$  will not appear in the preceding time-series analysis based on log differences. Yet this same analysis would pick up  $g_j$  which is almost always negligible but sometimes large due to, for example, adjustment lags to exogenous shocks in foreign prices. A large  $\Delta p_{jt}^*$  will generate, other things being equal, a large  $\Delta g_{jt}$  if it takes some time for this change in foreign prices to become known in the country under consideration; those large changes in markups that come primarily from deficiencies in information rather than from market power will be captured in this analysis. Sectoral analysis, such as the one based on segmenting related products into distinct categories, may therefore be more appropriate for identifying violations induced by market power or (sectoral) public policies, whereas analysis in terms of log differences may be more likely to stress the role of informational imperfections. Another empirical problem common to both analyses arises to the extent that price differences capture variability in the quality of the same good across regions and countries.

In addition to looking at prices, it is also useful to consider nonprice signals. The existence of parallel imports, for example, indicates that large  $g_{jt}$  values are likely to be due to high markups and market power rather than high transportation and distribution costs.<sup>10</sup> However, the absence of significant parallel imports does not necessarily indicate an absence of high markups. Manufacturers often attempt to diminish the demand for parallel imports by refusing to provide warranties or after-sales service for products sold outside the authorized system.

#### *Sources of market power*

Although a careful examination of price and nonprice data can provide some indication of LOP violations that are induced by market power in distribution, it does not allow us to determine whether the source of market failure is market power at the manufacturer or the distribu-

tor level. This distinction is important because the appropriate policy response will differ in the two cases. Suppose that the price variations are largely cross-country differences due to international price discrimination by manufacturers. In the absence of an international authority, there is little that the national competition agency could directly accomplish other than to ensure that no barriers to parallel imports (for example, in the form of contractual restrictions) are placed on local distributors.<sup>11</sup> Conversely, if the price variations are largely within-country differences due to market power of local distributors, then the competition authority can use appropriate measures to reduce such market power.

Some additional, simple price analyses may provide some indication of the source of market power. Problems in the distribution network may show up in the form of diverging sectoral (PPI) and consumer price index (CPI) series.<sup>12</sup> This divergence can be identified by first denoting the ratio of the two indices in sector  $j$  at time  $t$  as  $r_{jt} = \text{CPI}_{jt} / \text{PPI}_{jt}$ . Industries with  $r_{jt+1} / r_{jt}$  greater than one could indicate an industry with possible market power at the distributor level, since there would be direct evidence of consumer prices increasing at a faster rate than producer prices. However, there are several problems with this approach. Market power in distribution implies markups that are large but that do not necessarily increase over time. Also, industry definitions could vary across the two indices, which could make comparisons problematic.

Alternatively, if data on cost, insurance, and freight (c.i.f.), wholesale, and retail prices are available, one can directly determine whether differences in selling prices occur largely at the manufacturer or the distributor level (or at what level of distributor in a multitiered system). If differences in selling prices are largely due to c.i.f. price differences, then international price discrimination and manufacturer market power is a more likely cause of LOP violations. However, such an approach is also problematic. Since distributors can purchase goods directly through overseas subsidiaries, it is not clear what the c.i.f. prices would then reflect.

While additional price analyses may be helpful in distinguishing whether LOP violations are induced by market power, inferences regarding market power that are drawn from the simple observation of price differentials can be misleading and should be supplemented with micro-level studies. In practice, it may be difficult to determine what the (counterfactual) level of prices would be in an environment with a different degree of competition.

Consider the case of the beer industry in Venezuela. The leading domestic manufacturer, Polar, distributes beer

through 8 hierarchically controlled distribution companies, 80 warehouses, and 2,000 owner-drivers whose trucks were financed by the company and who work under exclusivity contracts (they cannot sell competitors' products), with assigned sales volumes, routes, and prices. Since 1983, 95 percent of the beer produced in Venezuela reaches the consumer through their distribution channels, with 42 percent sold through some 35,000 family houses (where owners offer service to surrounding homes via Polar-financed refrigerators).<sup>13</sup> Although tariffs were reduced to 30 from 80 percent, imports reportedly still were being sold in 1990 at prices 3 to 5 times higher than domestic production (Venezuela, 1991). Polar, who at an intuitive level would appear to be exercising market power with 83 percent of domestic sales and controlling almost all beer distribution, had in fact much lower prices than those charged by foreign manufacturers. It is the foreign manufacturers whose prices seem abnormally high, much higher than warranted by international prices and transport costs. The main point here is that it is not necessarily the firm(s) charging the higher price (the foreign manufacturers in this case) that are the original source of market power. While it may be argued that the relevant domestic comparison is between the price charged by Polar and the price Polar would have charged if there had been more competition in the mass market segment, it is difficult to use the abstract notion of a counterfactual, competitive price level as a practical guide for policymakers in cases such as these.

Consider the prices of certain foreign-produced goods in Japan, which have remained significantly higher than prices of comparable locally produced goods, even during periods of exchange rate stability. The higher prices of imports could be due to either (foreign) manufacturer or (domestic) distributor market power.<sup>14</sup> There is extensive but anecdotal evidence in favor of the latter, in the form of complaints by foreign manufacturers that the final selling price of their goods in Japan is far higher than they would like (Batzer and Laumer 1989). There are sound theoretical reasons for expecting domestic distributor market power to be a cause. The standard double marginalization vertical externality is exacerbated in Japan by the often-observed practice of having a (domestic) distributor act as sole distributor for several competing (foreign) products.<sup>15</sup> Since the sole distributor internalizes the cross-price effects, it faces a more inelastic effective demand curve and consequently charges a higher markup than would a distributor that acted as sole distributor for only one product. This reasoning is supported by an example. Since BMW vertically integrated into distribution in 1981, it has twice reduced its selling prices.<sup>16</sup>

The above discussion suggests the need for industry case studies to supplement the suggested price analyses. Industry practices that take the form of vertical restraints and other nonprice factors, such as supply foreclosure, are critical in determining whether the relevant market imperfections occur at the manufacturer or the distributor level. The higher the likelihood and impact of supply foreclosure, the more probable it is that manufacturers are the relevant source of market power. This may explain, for example, why dealers and distributors may be unable to take advantage of c.i.f. price differentials and import directly from lower-priced sources. The threat of foreclosure would be indicated by:

- *Product characteristics.* If products are undifferentiated, alternative sources of supply should be easier to obtain and foreclosure should be less damaging to the distributor than would be the case for differentiated products. Distributors of products with a significant after-sales service component are likely to be more dependent on manufacturers (for example, for technical information, and spare parts) than distributors of products with no service after the sale would be. In general, the existence or necessity of long-term contracts (for example, for after-sales service) may help one party to exert market power.
- *Likelihood of manufacturer collusion.* If manufacturers collude, foreclosure is a problem (even for manufacturers of undifferentiated products). Manufacturer collusion is in turn more likely in a protected market with trade barriers, or where the relationship between manufacturers is relatively symmetric.<sup>17</sup> Symmetry in capacity as well as multimarket interaction may also make collusion more likely (see Bernheim and Whinston 1986).
- *Degree of asset-specific, sunk investment.* Even in the case of differentiated products, distributors could switch to an alternative brand or an alternative product. If distributors have made significant asset-specific, sunk investments, however, then switching costs are much higher and foreclosure more damaging. Asset-specific, sunk investments may include product advertisement as well as product-specific training of sales and service staff.<sup>18</sup> Note that even if there is some ex ante competition between competing manufacturers before a distributor chooses a particular product, there later may be little ex post competition once a particular manufacturer has been chosen because of "putty-clay" types of relationships due to after-sales service or other sunk, manufacturer-specific investments.
- *Range of products handled by the distributor.* If the distributor handles multiple products from multiple manufacturers, then the threat of foreclosure by any one manufacturer carries less weight.<sup>19</sup>

- *Ease of entry into distribution.* Strong market power of distributors relative to manufacturers must be predicated on difficult independent entry by the latter into distribution.<sup>20</sup> Entry could be difficult for a variety of reasons, including legal restrictions, high fixed cost relative to market size, economies of scale and scope in production and distribution, and specialization economies in production versus diversification (multiproduct) economies in distribution for full-line retailers.

### Remedies

This section discusses the case for intervention against violations induced by policy and the exercise of market power. Persistent price differentials also may be explained by large differences in transportation and distribution costs, generated either by natural geography or by non-strategic sunk cost investments in transportation and distribution facilities. Enterprise-specific sunk cost investments that are strategic in nature and motivated by a desire to preserve or enhance market power, as well as competition restraints in the transportation sector, are addressed in the latter part of this section. Discussion of capital market imperfections as a possible source of trade friction is beyond the scope of this chapter.

#### *Policy-induced LOP violations*

Apart from international trade policy distortions, domestic policy and regulatory distortions—including domestic price controls, differential regional taxation, government-granted legal monopolies in distribution, and even weak delineation and enforcement of property rights—can be important determinants of persistent price differentials. By preventing resources from being reallocated in response to market forces of demand and supply, government-generated barriers to trade create distortions that lower national income. Although removal of the policy or regulation that created the barrier could reduce the cost in terms of LOP violations, this option must be counterbalanced with the other benefits engendered by the existing regulation.<sup>21</sup> In an environment with other remaining market failures and imperfections, policymakers need to carefully analyze the impact that removing existing public restraints will have on competition, to ensure that their removal does not make other distortions even more pronounced. In addition, how public restraints are removed, both in terms of the approach and sequencing, will have an impact on how market power-based barriers to arbitrage subsequently manifest themselves.

The China fertilizer industry provides a clear illustration of LOP violations that appear to be caused exclu-

sively by public restraints on competition, both international and domestic. In urea, for example, whereas the international price in 1992 was \$150 per ton, the domestic retail price varied from \$98 for heavily subsidized product linked to mandatory procurement of specific crops, to \$201 for allocation by sown areas of crops that local governments wanted to promote, to \$236 for negotiated “out-of-plan” quantities (for negotiated prices, maximum upper-bound levels reportedly are set).<sup>22</sup> Price differentials persist for a number of reasons. First, access to imports remains constrained by quantitative restrictions in the form of import licenses and limited foreign exchange, in addition to the continued dominance of SINOCHEM (a large, state-owned company that was granted monopoly rights to import fertilizer until 1990). However, even if international trade were fully liberalized, there remains a complex web of domestic price controls and cross-subsidies that prevents international and local prices from converging. The pricing and subsidy regime in the fertilizer industry stems from a basic national social contract of low wages in return for low food prices for the urban sector. Farmers, who receive low remuneration for their crops, in turn can pay only low prices for inputs such as fertilizer, the price of which therefore is controlled. In addition, considerable quantities of fertilizer are still being distributed and marketed within China under complex quota and allocation systems run by the three distinct levels of government, with many government-run trading companies enjoying legal monopolies over specific categories of fertilizer.

The recommended means of intervention in this example—the removal by central, provincial, and county-level government of public restraints on competition—should be implemented carefully. Fertilizer pricing is one component of the complex, integrated national pricing policy that supports low food prices and low wages for the urban sector. Price liberalization should be synchronized with import and domestic distribution and marketing liberalization to prevent shortages of fertilizer during the transition period. The lessons learned from the failure of the fertilizer liberalization reforms of the 1980s must be incorporated—including key links between reforms in fertilizer and agriculture, energy, trade, and transportation—to prevent social unrest and ensure enhanced performance.

In Russia, the prices of many potentially tradable goods differ from international prices, in large measure due to international trade policy distortions (including the sizable export controls needed to support remaining domestic price controls on raw materials). However, even

for certain goods whose border price is identical to the international price, insecurity of property rights in general and access to land in particular are critical public restraints on competition. Absence of well-functioning property rights laws discourages long-term investments in facilities, equipment, and training necessary for adequate, well-functioning, and competitive domestic distribution networks. Many warehouses remain tightly controlled by the former state companies, and lack of access to land prevents potential new entry by competing warehouse facilities. Wholesale and retail premises that are being privatized separately usually are being sold without the land on which they are built, with leases that either are short or narrowly circumscribe tenants' rights. Contract enforcement is weak, and distribution entrepreneurs are preyed on by racketeers and corrupt officials. Government pledges are not trusted at least in part due to past renegeing and political instability. The remedy in the case of land access goes beyond the establishment of clear laws to include the development of institutions and a legal culture that will enable those laws to work well in practice.

In Indonesia, domestic prices of some products remain higher than international prices, despite tariff and quota reductions. Elaborate government regulations limit the number and product range of distributors by conferring monopoly rights on designated local "registered agents." The importation of finished products into Indonesia must be handled by these registered agents. Domestic manufacturers, by contrast, are free to distribute their products without going through an agent. Foreign producers, foreign distributors, and foreign-domestic joint ventures are effectively prohibited from setting up their own distribution networks and must operate through the Indonesian distributors to whom the government has assigned their products. In general, one registered agent exists for all the foreign manufacturers of a particular product; in cases where more than one agent is selected for a particular product group, distribution rights are granted only for specific market segments. For example, one agent distributes exclusively to state-owned enterprises and another to privately owned companies. Foreign companies do not have the option of firing their mandated agent and working with another one.

These Indonesian regulations were intended to protect small, local distributors. There apparently was a concern that local distributors, after investing and developing the market for their products, would be vulnerable to being cut off by large, multinational manufacturers.

Other instruments, however, might better address such concerns. Removal of these public restraints on competition appears straightforward, but the government will meet resistance from the powerful vested interest groups that are benefiting from the current policy regime. As it seeks new and better instruments, the government also will have to ensure that alternative commercial dispute resolution mechanisms are in place and can be effectively implemented.

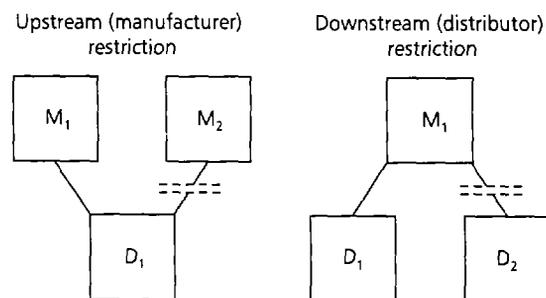
#### *Market power-induced LOP violations*

Even absent policy distortions, LOP violations may occur as a result of the exercise of market power in distribution. Anticompetitive behavior will, in fact, likely manifest itself more starkly when binding public restraints are removed. Conversely, weaker implementation of competition policies—such as a more lenient merger policy or attitude toward vertical restraints—may confer more market power to established firms.

Distribution channels can be foreclosed by vertical integration, vertical restraints, or other restrictive commercial practices. That these practices have anticompetitive effects and lead to possible LOP violations is not, however, sufficient to establish the case for intervention, which must be made based on net (national) welfare considerations. At a minimum, evaluating the case for intervention requires assessing whether the practice is anticompetitive, and if so, whether it has efficiency or other beneficial effects, whether such effects outweigh the anticompetitive effects, and whether the likely benefits of the preferred means of intervention outweigh any costs associated with active policy intervention. Each of these points will be discussed in turn by way of creating a checklist for determining whether policy intervention against LOP violations induced by market power is advisable.<sup>23</sup>

It is helpful to begin with a classification scheme that considers in greater detail the relationship between an upstream firm producing a good—called the manufac-

**Figure 4.1 Target of restrictive practice**



turer (M)—and a user of that good that resells it—a downstream firm called the distributor (D). D could be a wholesaler, a retailer, another firm in a longer distribution chain, or an industrial user of the intermediate good.<sup>24</sup> The range of possible restrictive practices between M and D varies from vertical integration (where M absorbs D or vice versa and all assets are commonly owned), to vertical restraints (explicit contractual arrangements between M and D restricting the market choices otherwise available to each), and finally to restrictions based not on contracts but on actual actions or credible threats of market foreclosure (including commercial practices that reduce D's access to M or limit M's access to D).

Traditionally, restrictive practices have been distinguished on the basis of whether they aim at reducing competition among established firms or at foreclosing markets and deterring entry.<sup>25</sup> For the purposes of this chapter, it is more helpful to classify restrictive practices in distribution according to their likely target, that is, whether the practice under examination has the potential to restrict entry or expansion of a manufacturer or of a distributor (figure 4.1). Upstream restrictions block access of one or more manufacturers to a given distribution channel. In figure 4.1,  $D_1$  denies access to  $M_2$ . Restrictions on interbrand competition fall within this class, since they prevent a distributor from selling goods that are close substitutes and otherwise would compete directly with the manufacturer's product. Through a clause in a sole distributorship contract, for example, M can restrict D to handle only its products and not those of its local and foreign competitors. Conversely, D can agree to market one or more manufacturers' products but not others.

Downstream restrictions by manufacturers prevent one or more distributors from entering an existing market or force an incumbent firm to be squeezed out (see figure 4.1). Restrictions on intrabrand competition fall within this second class, since they prevent other distributors from carrying the same brand. Through a clause in a sole distributorship contract, for example, M can restrict distribution of its product to a single D, denying access to that product to other distributors; an exclusive-territory clause also allows M to restrict intrabrand competition among distributors.<sup>26</sup> Finally, D can prevent other distributors from being supplied by a given manufacturer, as illustrated by the case of a multiproduct retailer in the United States.<sup>27</sup>

*Anticompetitive effects.* Restrictive practices in distribution have anticompetitive effects when they allow a firm, unilaterally or in collusion with others, to reduce output,

profitably raise price, and maintain the higher price over a significant period of time without competitive response by other (existing or potential) firms. Vertical restraints can be used to enhance market power, either by facilitating collusion (strengthening horizontal agreements or weakening competition at upstream or downstream stages) or by enabling market foreclosure (excluding competitors and thereby maintaining or enhancing the market power of dominant firms). However, as discussed earlier, a simple observation of prices is not sufficient to make a determination of restrictive practices in distribution due to market power.

A more practical line of inquiry is to focus directly on the nature of restrictive practices in distribution and the possible underlying source of market power. A good initial step is to ask whether the firm imposing the restrictive practice has market power or whether a group of competing firms imposing similar restrictive practices might collectively have market power. Restrictive practices by firms that possess little or no market power are unlikely to harm competition. Two questions are useful in assessing whether there are opportunities for the exercise of market power: Are barriers to entry sufficiently high to permit the unilateral exercise of market power by a dominant firm? In the case of collective agreements, does the agreement affect a large percentage of the market, are markets highly concentrated to facilitate collusion, and if so, what allows markets to remain concentrated?

Second, the firm or firms in question must enjoy some bargaining power in their relationship with enterprises at the other stage. Here, a useful question is, what makes such control feasible in practice? Third, the restrictive practice must allow the firm or firms with market power to maintain or enhance their market power. If the restrictive practice has no impact on current and expected future sales and retail prices, but only redistributes the division of profits between manufacturer and distributor, it likely will not have a negative effect on welfare. In this context, it is important to assess whether the practice reduces sales and inflates the costs facing potential entrants beyond the cost of doing business absent the practice.

A recent competition law ruling in the United States illustrates the importance of keeping the first point in mind, that is, a careful assessment of whether entry is difficult.<sup>28</sup> As was decided in that case, a manufacturer's threat to terminate a distributor for continuing to sell a competing product did not in and of itself violate U.S. antitrust law. To prove an antitrust violation, the court ruled, the plaintiff was required to show harm to the com-

petitive process, not injury solely to the plaintiff. That an alternative new product had, within a short time, successfully entered the market despite the heavy-handed competitive tactics of the defendant manufacturer was sufficient evidence, in the court's view, of a lack of injury to competition.

*Efficiency effects.* That certain anticompetitive practices have efficiency effects and are therefore tolerated is widely known. Patents, for example, have clear anticompetitive effects but are tolerated because they encourage investment. Similarly, vertical restraints can create a means and incentive for both parties in a vertical relationship to enhance intrabrand coordination. For instance, vertical restraints can help solve problems of double marginalization, or overcome market failures that arise when explicit contracts cannot be written and enforced—because of the unobservability of variables such as promotional effort or services. Vertical restraints imposed for efficiency purposes have a positive effect—increased sales of the product—that is procompetitive.

Vertical restraints in distribution may, for example, encourage investments that improve distributive efficiency and increase sales. Consider the example of beer distribution in Venezuela. Polar, the leading domestic manufacturer, might not have made investments in trucks and refrigerators if it had not been granted exclusivity of use. If distributors were allowed to use trucks and refrigerators financed by Polar to transport and sell competitors' beer brands, Polar would not have provided the financing. Lacking collateral and therefore access to alternate sources of financing, distributors would have been unable to supply end-use customers with desired quantities of beer, at the desired temperature. As a consequence, sales would have been substantially lower.

Restrictive practices in distribution may also encourage the provision of complementary services. For example, sole distributorships and exclusive territories—by preventing other distributors or dealers from free-riding on demand-promoting measures such as sales services or product advertisement—may encourage those distributors or dealers to provide such measures. Polar's requirement that each owner-driver have a separate, nonoverlapping, assigned delivery route thus could be viewed as a measure that increases promotional effort. If one distributor is unable to reap the full benefit of his advertising efforts because the consequent increased demand for the product benefits other distributors, it will reduce advertising expenditures and less product will be sold. Such demand-promoting services are typically underprovided by distributors because they fail to inter-

nalize the markup that manufacturers earn from additional sales.<sup>29</sup>

A less obvious point is that anticompetitive, entry-detering behavior may have positive welfare implications. In models of outlet and product location, for example, the private-market outcome can generate excessive entry relative to the social optimum. Such excessive entry occurs because the private incentive to enter depends only on positive profits for the entrant. Entrants do not take into account the trade-diversion effect of stealing business from incumbents. In such a context, anticompetitive practices that deter entry by artificially inflating the costs of potential entrants could be welfare-improving by preventing unnecessary duplication of fixed-entry costs in distribution.

If some of the enterprises are foreign, issues of national welfare and the transfer of foreign rents must also enter into the policymaker's assessment. Few clear prescriptive guidelines can be offered in this regard since the analysis here is predicated on the relative weights assigned to consumer and producer welfare. Consider the example of a relationship between multiple manufacturers and a sole distributor that results in higher end prices for consumers. A policymaker who places a high weight on consumer welfare is more likely to intervene to correct any anticompetitive practices—regardless of the national origin of the players. Conversely, a policymaker who places a lower weight on consumer welfare may be less inclined to intervene in a situation involving foreign manufacturers and a domestic distributor.

*Anticompetitive versus efficiency effects.* An assessment that the anticompetitive effects of a particular restrictive practice in distribution outweigh any efficiency effects is a prerequisite for an initial presumption in favor of intervention. As the above discussion suggests, it is difficult to offer any categorical rules for policymakers. Nonetheless, it is useful to draw some rules of thumb and heuristics for weighing the anticompetitive and efficiency effects of restrictive practices in distribution. Given the lack of robustness of theory on this front, however, the following guidelines are intended only to be indicative.<sup>30</sup>

In cases of *downstream restrictions*, anticompetitive effects are likely to outweigh efficiency effects:

- *When the good being sold is not bundled and has a low service component.* Since free-riding on services by other dealers is not significant, sole distributorships are likely to have mostly anticompetitive effects.
- *When the exclusionary practice is instigated by the distributor rather than the manufacturer.*<sup>31</sup> Since the positive service externality is felt mostly by the manufacturer (due to

its additional upstream markup) and not by the distributor, sole distributorships sought by distributors are likely to be motivated largely by anticompetitive considerations.

- *When interbrand competition is low.*

In cases of *upstream restrictions*, anticompetitive effects are likely to outweigh efficiency effects:

- *The smaller the fixed costs of entry into distribution relative to the size of the market.* This is the case because potential entrants are likely to be able to earn non-negative profits, and there are fewer economies of scale or scope. Unnecessary duplication of fixed costs in distribution is less relevant.

- *The smaller the size of the investments potentially subject to hold-up problems and which would otherwise be precluded in the absence of exclusionary, contractual restrictions.*<sup>32</sup>

In general, intervention is more likely to be justified against exclusionary practices in the form of upstream restrictions, which are less likely to have an efficiency rationale, than against practices that take the form of downstream restrictions, such as preventing free-riding on dealer services.<sup>33</sup> Intervention also is more likely to be warranted when the exclusionary practices are enforced by implicit threats of foreclosure rather than by explicit restrictive contractual arrangements. If the exclusionary practice is motivated largely on efficiency grounds with clear benefits to both parties, then there is little reason not to explicitly spell out the terms of the relationship in a contract.<sup>34</sup> By contrast, contracts for purely anticompetitive arrangements such as collusion, for example, are not likely to be entered into since they will not be upheld by courts. Of course, the mere existence of a contractual arrangement by no means suggests that intervention is not justified for at least two reasons: First, seemingly innocent contractual terms can have an anticompetitive intent and effect and second, parties can be coerced into a contractual agreement through noncontractual means.

While exclusionary practices by contractual rather than noncontractual means are preferable because of their greater transparency, the guidelines for determining appropriate contractual terms—for example, with regard to length—are less clear. Shorter-term contracts will clearly be less restrictive and less anticompetitive, but there may well be efficiency reasons for negotiating longer-term contracts.<sup>35</sup> A rule-of-reason approach toward contracts therefore is suggested. One possibility is to adopt criteria akin to those used in awarding patents: On efficiency grounds, exclusionary contractual arrangements with anticompetitive effects are allowed for time periods that depend on the size of the investment subject to potential hold-up problems. A broad range of efficiency

problems also might be solved through instruments and policies that are less harmful to competition.<sup>36</sup> Hold-up problems might be resolved by contract law that protects the fruits of asset-specific investments. Free-rider problems in the provision of distributor services could be resolved by contracts specifying what these services are.

The case for intervention against restrictive practices should be evaluated on the basis of the effects of the practices and independently of the mode of asset ownership. Practices that warrant intervention do so regardless of whether they are induced by vertical integration, contractual arrangements, or threats of foreclosure. Since it is usually possible for independently-owned upstream and downstream enterprises to find a set of vertical restraints that replicate the integrated outcome, the mode of ownership becomes less relevant.

*Net benefits of intervention.* Even if the anticompetitive effects of a particular restrictive practice in distribution outweigh any efficiency effects, the final judgment regarding the appropriateness of intervention depends on a comparison of the net benefits of the preferred means of active intervention versus no intervention. A crucial benchmark in evaluating the case for intervention is to ask, What can the intervention realistically accomplish, and at what cost? Existing pricing and access arrangements in distribution are endogenous and reflect the current market structure. Since some forms of intervention change the existing market structure (for example, by breaking up previously integrated firms), the pricing and access arrangements will change as well, and the resulting outcome may differ from that hoped for by participants.

The cost of the intervention itself also must be taken into account. What institution within government will perform the required investigation and analysis, enforcement, and adjudication functions? What are the incentives of individuals working within these institutions, and how likely are they to face pressure from entrenched rent-seekers in society? And to what extent does the potential of intervention (versus a clear nonintervention posture) adversely affect business certainty and therefore new business investment? If there is a carefully crafted competition law supported by well-functioning investigation, enforcement, and adjudication institutions, the cost of intervention will be substantially lower than if these laws and institutions are either weak or nonexistent.<sup>37</sup>

### Questions for further research

While far from complete, the available evidence indicates that LOP violations are fairly widespread. Since there are numerous possible reasons for LOP violations, deter-

mining whether policy intervention is warranted at a minimum requires determining the relevant cause for the LOP violation. This chapter provides a scheme for classifying LOP violations as well as some preliminary tools for identifying their cause. Given proper identification of the relevant reason for LOP violations, the paper also provides a checklist for evaluating the case for policy intervention.

The policy guidelines set out in the preceding section are rather preliminary. To endow these guidelines with greater operational significance requires further work around the following key research questions:

- *How significant are LOP violations caused by public policy, market power, and natural causes?* In the case of LOP violations induced by market power, particularly by anti-competitive practices in distribution, their significance will depend on (a) how often such practices occur (frequency); (b) how long their effects last (duration); and (c) their impact on welfare (magnitude). A first-cut assessment of the frequency of such practices can be obtained by examining the type of price decompositions discussed in the first section to identify possible areas where market failure might occur. However, as noted earlier, there is no necessary direct link between observable price differences and anti-competitive practices. Although price differentials may reflect the existence of market power and competition restraints, and may call for intervention, the firm charging the higher price may not necessarily be the source of monopoly power. Consider, for example, the case of beer in Venezuela. Heineken beer was priced much higher than warranted by international prices and transport costs, but it was Polar, the domestic manufacturer, that had the tightly controlled distribution network. In fact, a higher price could be an endogenous response to an incumbent's tight control over distribution: An entrant blocked from the mass market could find it more profitable to pursue a high-price strategy aimed only at specific market segments. In the Venezuela case, the higher price of imported beers seems to have been generated by the market power and anti-competitive behavior of the domestic manufacturer. This case could be analyzed in a vertical product differentiation model with endogenous product positioning and pricing choices,<sup>38</sup> which suggests that additional research on identifying signs of market failure in distribution may be warranted.<sup>39</sup>

Turning to the likely duration of competitive problems in distribution, one should ask whether these problems are bound to correct themselves as the size of the market increases sufficiently to justify independent or additional entry into distribution. In the Japanese case example, the

absence of independent, sole distributors was cited as an obstacle by foreign manufacturers. But in principle, there appeared to be few formal obstacles to distributorships. The difficulties faced by foreigners in acquiring existing Japanese companies and in finding Japanese employees willing to work for them have been cited as obstacles. However, these reasons are likely to be subordinate to the issue of whether the market is large enough to justify entry into distribution. Recall that BMW in Japan eventually set up its own (proprietary) distributor and subsequently charged less for its autos than its (independent) distributor had charged. To the extent that a market can grow over time (for example, as consumers become aware of new products), competitive problems in distribution may resolve themselves.

Assessing the duration of competitive problems in distribution therefore requires understanding the root cause of the anti-competitive effects. One issue that cuts across most of the case examples is why import arbitrage (for example, parallel imports) and entry into distribution do not occur sufficiently to counteract the effects of anti-competitive practices. Again, if the reason is that the fixed costs of entry are large relative to market size, the problem is likely to correct itself as the market size increases.<sup>40</sup>

At issue here is the speed or pace of price convergence. Even though price convergence following trade liberalization may occur over time, intervention may be called for if convergence is slow. There appears to be little systematic data and analysis on price convergence following trade liberalization in developing economies. Cross-country comparisons in well-functioning industrial economies where trade is open suggest that convergence is slow (Mueller 1990). It is reasonable to postulate that convergence is slower in developing economies.

Finally, how important is the welfare impact of LOP violations induced by public policy, market power, and other causes? Here it would be useful to develop a tool kit for roughly assessing the quantitative welfare impact in particular cases, especially the case of anti-competitive practices in distribution. Such an effort could include guidelines for benchmarking the usual (static) welfare loss triangles and determining whether there are significant additional dynamic losses.<sup>41</sup> Other rules to help policymakers set priorities for intervention also would be helpful. For instance, it would be useful to know whether welfare losses from anti-competitive practices in distribution generally are larger for producer goods than for consumer goods.

- *How should the rules for intervention used in industrial countries against LOP deviations, especially those induced by market power, be amended to take into account developing-*

*country characteristics?* To illustrate this point concretely, we return to the evaluation of the case for intervention and assess how particular features of the developing-country context affect the earlier discussion of policy remedies. The small market size of many developing-country markets means that interbrand competition is likely to be low, and the fixed costs of entry are likely to be high (relative to market size). From the checklist for intervention, this would tend to argue for greater intervention, other things being equal. Developing-country markets, however, are characterized by information imperfections. Many products and services are likely to be new and unfamiliar and therefore to require a significant information and service component. There also is likely to be greater inherent risk and uncertainty in a developing-country environment, making investment less attractive. Both of these features suggest that less intervention against LOP violations induced by market power is warranted in developing countries than in industrial countries.<sup>42</sup>

The following related questions also must be addressed: Which features of the developing-country context (for example, market size, information imperfections, risk) are most important from the standpoint of restrictive distribution practices that induce LOP violations? How does the appropriate balance between anti-competitive versus efficiency effects of particular practices in distribution differ in developing countries from the balance in industrial countries?

## Appendix

### A note on parallel imports

Parallel imports, or gray-market goods, are genuine branded goods that are sold outside authorized distribution channels.<sup>43</sup> Differences in international selling prices—typically due to manufacturer price discrimination—provide an incentive for goods to be diverted from lower- to higher-price countries. Middlemen of the authorized channel are typically the major source of gray goods.

Parallel imports imply an increase in intrabrand competition, whereas sole distributorship arrangements restrict it. There may therefore be a presumption that if sole distributorships are desirable, then parallel imports are undesirable. However, this presumption is not justified. There may be sound service promotion and specific investment reasons for the existence of sole distributorships. Parallel imports occur in response to higher retail prices in one country relative to another.<sup>44</sup> However, high retail prices are likely to be due to manufacturer price dis-

crimination rather than (dealer) service promotion reasons, since there are other vertical restraints that can encourage service provision in lieu of high retail prices. The loss of consumer surplus that such prices engender may justify encouraging parallel imports, especially if the manufacturer is foreign and there is no compensating increase in domestic producer surplus. Parallel imports—and the limits on price discrimination that they imply—should not be sufficient to override the other reasons for maintaining a sole distributorship network. Consider, for example, the Ford AG case (note 11) in which Ford agreed to limit retail (list) price differences in various countries in return for being allowed to maintain its dealer network.

Parallel imports are a way of limiting manufacturer market power with respect to intrabrand restrictions. Since markets in many developing countries are small, with limited interbrand competition, parallel imports are likely to be a more expedient way of limiting manufacturer market power in developing countries than in industrial countries.

Parallel imports may not suffice to eliminate LOP violations. First, the volume of parallel imports can be limited and variable since it depends largely on middlemen in the authorized channel who are ultimately answerable to the manufacturer. Second, product attributes are not identical from country to country. While some of these are inherent—for example, left hand-drive versus right hand-drive cars, different voltage requirements—others are a result of manufacturer policy—for example, foreclosing or price discriminating on after-sales services.<sup>45</sup>

## Notes

The authors thank Claudio Frischtak, Patrick Rey, and Randi Ryterman for their comments on an early draft of this chapter.

1. Procompetencia (Superintendencia para la Promocion y Proteccion de la Libre Competencia, the Venezuelan competition office), Caracas, June 1993.
2. For a recent survey of the economics of price discrimination, see Varian (1989).
3. In Emerson and others (1988), see especially section 7.1, Effects of competition on costs and prices, pp. 145–57.
4. See World Bank (1994a and 1994b). In the former study, the section entitled “The Impact of Price Reform” concludes by noting: “On balance it is clear that regional price variations were still significant in 1991 and have not noticeably diminished since 1986. The analysis . . . reinforces the findings on regionally segmented markets and suggests that this issue should now be the area of focus” (p. 26).

5. For example, a recent review of the treatment of franchising agreements in industrial countries' competition policy and law is provided by the Organization for Economic Cooperation and Development (1994). The scope of this OECD report goes beyond franchising agreements, providing an evaluation of the competitive consequences of a fuller set of vertical restraints. For a recent discussion of the view that both vertical restraints and the enforcement of competition policy toward vertical restraints can have serious implications for the flow of international trade, see Comanor and Rey (1994a).

6. Tirole (1988, chapter 9) provides an overview of issues related to information and strategic behavior.

7. Perishability is one dimension of possible variability in quality. Although two goods may be similar in use and even in brand, their price may differ to the extent that they are not identical, for instance, sold at different times (fresh versus spoiled) or produced in different locations (and therefore possibly viewed differently by consumers).

8. Adapted from Berry and Levinsohn (1993).

9. This would capture any exchange-rate pass through or "marking to market" effects. These effects recently have received greater attention and analytical justification due to hysteresis considerations.

10. Parallel (gray-market) imports can be significant. In the United States parallel imports are estimated to have accounted for approximately 10 percent of leading ski brand imports, 20 percent of Seiko watches, and 33 percent of camera imports. Such significant shares are due to large price differentials: K-Mart reported that the prices it obtains from gray-market wholesalers are 20 to 40 percent below those of authorized wholesalers. The appendix discusses the case for parallel imports.

11. There are cases in which such an international authority exists, especially given the increasing emergence of regional trading arrangements. The European Commission, for example, took Ford to court for impending parallel imports among its dealers. See *Ford Werke AG and Ford of Europe, Inc. v. E.C. Commission*, cases 25/84 and 26/84 before the European Commission Court of Justice; [1985] 3 C.M.L.R. 528, as cited in Adams (1989).

12. As noted in Berry and Levinsohn (1993).

13. For an overview of the nature of vertical control in the Venezuelan beer industry, see Jatar (1991), chapter 6—Vertical Control in the Beer Industry.

14. The high prices of foreign cars could also come from the market power and anticompetitive practices of domestic car manufacturers that use strategies to raise rivals' costs.

15. Double marginalization occurs when you have two monopolists in a chain. The downstream monopolist charges a markup on a price which has already been marked up by the upstream monopolist.

16. "The difference in pricing between selling through the company's own organization and through a Japanese sole importer is made plain by the following example: whereas both the BMW 745i and the Mercedes 500 SE sell to the consumer at about the same price on the German market, the 745i sells at 9.5 million yen in

Japan but the Mercedes 500 SE at 13 million yen" (Batzer and Laumer 1989, p. 194)

17. For example, in the Venezuela beer industry case, Polar would be less likely to collude with Heineken than with the Colombian manufacturer due to the asymmetric relationship between Heineken and Polar: Polar cannot credibly retaliate significantly against Heineken in the Dutch company's markets but could do so against its Colombian neighbor.

18. Expenditures at automobile dealerships, for example, share these characteristics.

19. As appears to be the case in Japan, for example, where domestic car dealers handle several foreign brands.

20. Porter (1976, p. 21) and Steiner (1985) offer more discussion on the relative bargaining power of large distributors.

21. There also may be other regulations that could accomplish the same benefit as the existing ones but at lower costs in terms of LOP violations.

22. See World Bank (1994b), table 5, p. 14.

23. For a more detailed discussion of the conceptual issues raised here, see, for example, Organization for Economic Cooperation and Development (1994), especially chapter 2 on the economics of franchising and vertical restraints and the references contained therein; Katz (1989); and chapter 4 on vertical control in Tirole (1988).

24. Such relationships between clusters of assets at different stages of the production or distribution of goods and services are referred to more generally as vertical relations. For an analysis of the anti-competitive effects of vertical relations, it is customary to focus on one vertical link in the channel, since the existence of one bottleneck is sufficient for the exercise of market power.

25. For examples of restrictive practices that reduce competition, see Bonanno and Vickers (1988) and Rey and Stiglitz (1985, 1988). For examples of entry-detering practices, see Aghion and Bolton (1987) and Comanor and Rey (1994a).

26. Brands need not be identified solely with manufacturers. In several countries large supermarket chains, for example, now carry their own brand names.

27. In response to a discounting distributor's practice of selling at substantial discounts below the prices of full-line distributors, a second, larger (full-line) distributor complained to two swimwear manufacturers. The larger distributor threatened to reduce or discontinue its business with these firms if the two manufacturers continued to sell to the discounter. In response to this threat, both manufacturers refused to sell their primary line of swimwear to the discounting distributor. See *Toys 'R' Us, Inc. v. R. H. Macy & Co. Inc.*, 1990-1 Trade Cases 68, 890 (S.D.N.Y. 1990) as cited in Comanor and Rey (1994b). The authors also cite *Business Electronics Corporation v. Sharp Electronic Corporation*, 485 U.S. 717 (1988), in which a large distributor (Hartwell) and a smaller one (Business Electronics) repeatedly cut retail prices below the levels recommended by the manufacturer (Sharp).

## REGULATORY POLICIES AND REFORM: A COMPARATIVE PERSPECTIVE

- Comanor, W., and P. Rey. 1994a. "Competition Policy Toward Vertical Foreclosure in a Global Economy." University of California, Department of Economics, Santa Barbara.
- . 1994b. "Vertical Restraints and the Market Power of Large Distributors." University of California, Department of Economics, Santa Barbara.
- Corbo, V., and P. McNelis. 1989. "The Pricing of Manufactured Goods During Trade Liberalization: Evidence from Chile, Israel, and Korea." *The Review of Economics and Statistics* 71: 491–99.
- Dutz, M. A., and R. S. Khemani. 1994. "The Instruments of Competition Policy and Their Relevance for Economic Development." World Bank, Private Sector Development Department, Washington, D.C.
- Emerson, M., M. Aujean, M. Catinat, P. Goybet, and A. Jacquemin. 1988. *The Economics of 1992: The E.C. Commission's Assessment of the Economic Effects of Completing the Internal Market*. London: Oxford University Press.
- Jatar, A.J. 1991. "Determinants of Vertical Integration and Control in Distribution Channels," Unpublished Ph.D. dissertation, University of Warwick.
- Katz, M. 1989. "Vertical Contractual Relations." In Richard Schmalensee and R. D. Willig, eds., *Handbook of Industrial Organization*. London: Elsevier Science Publishers, B.V.
- Morande, F. G. 1983. "Retail and Wholesale Prices of Tradable Goods in Chile, International Commodity Arbitrage and Commerce Intermediation Services." Ph.D. diss., University of Minnesota, Department of Economics.
- . 1986. "Domestic Prices of Importable Goods in Chile and the Law of One Price 1975–1982." *Journal of Development Economics*. 21: 131–47.
- Mueller, D. C., ed. 1990. *The Dynamics of Company Profits: An International Comparison*. Cambridge: Cambridge University Press.
- OECD. 1994. *Competition Policy and Vertical Restraints: Franchising Agreements* Paris.
- Porter, M. E. 1976. *Interbrand Choice, Strategy and Bilateral Market Power*. Cambridge, Mass.: Harvard University Press.
- Rey, P., and J. Stiglitz. 1985. "The Role of Exclusive Territories in Producers' Competition." *Mimeo* (revised 1987,1991).
- . 1988. "Vertical Restraints and Producers' Competition." *European Economic Review* 32: 561–68.
- Steiner, R. 1985. "The Nature of Vertical Restraints." *Antitrust Bulletin* 30: 143–97.
- Stoll, N. R. and S. Goldfein. 1994. "Heavy-Handed Competitive Tactics." *New York Law Journal* 211.
- Tirole, J. 1988. *The Theory of Industrial Organization*. Cambridge, Mass.: MIT Press.
- Varian, H. R. 1989. "Price Discrimination." In Richard Schmalensee and R. D. Willig, eds., *Handbook of Industrial Organization*. London: Elsevier Science Publishers, B.V.
- Venezuela, Ministry of Investment Promotion. 1991. *Estudio del Mercado "Cerveza"*. Caracas, March.
- World Bank. 1994a. "China: Internal Market Development and Regulation." Report No. 12291-CHA. Washington, D.C.
- World Bank 1994b. "China Fertilizer Industry Report: Policies to Improve Production and Use."
- Yager, L. 1993. "Price Comparisons between the Japanese and U.S. Markets." Working Paper N-3337-CUSJR. Rand, Palo Alto.

28. The defendant, Welch Foods, Inc., a producer of fruit juices and related products, decided to discontinue its informal relationship with the plaintiff, R. W. International Corp. (RW), a Puerto Rican distributor, on the ground that RW's handling of a competing line of juice products represented an irreconcilable conflict of interest between RW and Welch. See Stoll and Goldfein (1994).

29. Several other vertical restraints also could be used to overcome such a problem, for example, a combination of low wholesale price and franchise fee.

30. At the very least, it must be stressed that these guidelines hold only if all other things are held constant.

31. In practice, it may be difficult to determine the instigator of the exclusionary practice in the absence of specific evidence. In the Sharp case cited by Comanor and Rey (1994b) (see note 25), the large distributor, Hartwell, threatened to terminate its own dealership unless the manufacturer, Sharp, ended its relationship with Business Electronics, the rival distributor, within 30 days (the rival distributor's retail prices were generally below Hartwell's own retail prices). Sharp complied and terminated Business Electronics' dealership. Even in this case, the majority of the court took the view that these restraints would be present only if they served the interests of the manufacturer. Only a dissenting opinion stressed the presence of market power at the distribution stage.

32. The "small" size of investment must be appropriately defined with respect to the players involved.

33. Note that free-riding could take place at either the manufacturer or distributor stage.

34. Provided the transactions costs of contracting are not excessively high. However, if the efficiency grounds for resorting to exclusionary practices are large to begin with (for example, large size of investment), these should outweigh the transactions costs of contracting.

35. There are few grounds for the general presumption that contracts in developing countries should be either longer or shorter than those in industrial countries. One can argue that since the environment changes more rapidly in developing countries, (repeated) shorter contracts would allow such information to be incorporated more readily. Conversely, however, one of the benefits of longer-term contracts discussed in the principal-agent literature is the ability of the principal to offer intertemporal risk sharing and of the agent to commit to future payoffs and punishments. These advantages of longer-term contracts are typically undone by access of the agent to credit markets. Since such access is less likely in the imperfect credit markets of developing countries, one can argue that the efficiency of longer-term contracts is therefore greater in developing countries than in industrial countries.

36. In France, for example, Sony does not use selective or exclusive distribution arrangements but instead offers lower wholesale prices to dealers providing a variety of customer services.

37. For a more detailed discussion of competition law issues within the broader context of competition policy, see Boner (1995) and Dutz and Khemani (1994).

38. However, such models often run into problems of the existence of equilibria with nonuniform distributions of consumers, problems that are particularly relevant in the dual or bimodal markets of developing countries. More appropriate assumptions capable of generating an equilibrium with the relevant features are needed.

39. Rey and Stiglitz (1985) refer to the possible use of exclusive territories as an entry deterrence device. By delegating pricing decisions to independent retailers with limited territories, a manufacturer could commit itself to reacting more aggressively to geographically limited entry, since each retailer does not take into account the losses inflicted on neighboring retailers of the same manufacturer when considering a reduction in its price.

40. The likelihood that the problem will correct itself in the long run does not imply that intervention is not warranted.

41. Above and beyond a simple repetition of the static losses.

42. The two are related. An investment may be deterred by high levels of up-front uncertainty, even though it could then be easily replicated if successful due to informational externalities.

43. As distinguished from counterfeit goods.

44. They also may be used to free-ride on promotional efforts.

45. In principle, the criteria used in evaluating whether such foreclosure of after-sales services is anticompetitive should be the same as that used to evaluate anticompetitive practices in distribution in general.

## References

- Adams, W. J. 1989. "Exposure of French Manufacturing to International Competition." In D. Audretsch, L. Sleuwagen, and H. Yamawaki, eds., *The Convergence of International and Domestic Markets*. Amsterdam: North-Holland.
- Aghion, P., and P. Bolton. 1987. "Contracts as Barriers to Entry." *American Economic Review* 77: 388-401.
- Baústa, J. C., and M. M. Mesquita, 1994. "Oligopolio: Em Defesa Da Razao." *Folla De São Paulo*, March 31, 1994.
- Batzer, E., and H. Laumer. 1989. *Marketing Strategies and Distribution Channels for Foreign Companies in Japan*. Boulder: Westview Press.
- Bernheim, Douglas, and M. Whinston. 1986. "Multimarket Contact and Collusive Behavior." Harvard University, Department of Economics, Cambridge, Mass.
- Berry, S., and J. Levinsohn. 1993. "Notes for an Empirical Study of Price Changes in Venezuelan Industry." University of Michigan, Department of Economics, Ann Arbor.
- Bonanno, G., and J. Vickers. 1988. "Vertical Separation." *Journal of Industrial Economics* 36: 257-65.
- Boner, Roger Alan. 1995. "Competition Policy and Institutions in Reforming Economies." World Bank, Private Sector Development Department, Washington, D.C.

# Antidumping policy and competition

Sadao Nagaoka

In recent years antidumping policy has become a major trade policy instrument in industrial countries, and increasingly in developing countries as well. Other instruments, such as tariffs, quotas, and voluntary export restraints (VERs), used to dominate antidumping barriers, but antidumping measures are being employed to a growing extent for protectionist purposes under the rhetoric of fair trade (Boltuck and Litan 1991; Finger 1992a). Although the Uruguay Round made substantial progress in streamlining other trade restrictions, including VERs, it did little to reverse the strong protectionist bias of antidumping regulations. Antidumping regulation thus remains one of the most restrictive trade barriers in industrial countries.

In the 1980s, as many developing countries took unilateral steps to liberalize their trade regimes, they also enacted antidumping laws to protect their domestic industries from “unfair” foreign competition in the new, more liberal trade environment. In recent years some of these countries have become such active users of antidumping legislation that both competition and their national economic welfare may be significantly harmed.

Developing countries must design and manage their trade policy instruments intelligently. They must avoid the mistakes made by industrial countries and safeguard their past liberalization achievements. At the same time, both multilateral and unilateral efforts to reform antidumping policy should be intensified. The reform of antidumping regulations may well be a high-priority issue in the next round of trade negotiations.

This chapter raises issues that should be considered in any effort to reform antidumping policy. Its objectives are to:

- Review some basic definition issues concerning the General Agreement on Tariffs and Trade (GATT) and specific antidumping regulations and their implementation.
- Examine antidumping measures from both a global and a national welfare perspective.

- Discuss the issues that have emerged during debates on antidumping policy and its effect on competition.
- Derive policy recommendations and identify priority research issues.

The first section of this chapter starts by noting the sharp increase in antidumping investigations, as well as the more recent use of antidumping measures by some industrializing countries, such as Mexico. A discussion of the determination of dumping and material injury is followed by a review of antidumping measures taken by the four major users—Australia, Canada, the United States, and the European Union. The section highlights the strong bias of antidumping policy in favor of domestic industry, the absence of clear rules and criteria by the GATT (and the World Trade Organization) concerning issues such as material injury, and the major differences of antidumping regulations across jurisdictions.

The second section presents an economic analysis of antidumping measures in an attempt to answer the following questions: Why does dumping occur? Do antidumping measures affect the export price more than the home-market price of the exporters? Can the imposition of antidumping duties improve the terms of trade of the importing country? Is there a stable relationship between the extent of injury to the domestic industry and the welfare of the importing country? What are the global and national welfare effects of antidumping measures?

The third section focuses on the issues that have emerged in recent antidumping policy debates: the use of antidumping regulations for anticompetitive purposes such as collusion and predation; the most efficient approach to prevent international predation; the challenges to antidumping policy posed by the globalization of industry; and whether antidumping policy can contribute to the removal of distortions of global competition.

The concluding section presents policy recommendations and suggests some priority research issues.

### Implementation of antidumping policy

Antidumping investigations are undertaken by the governments of importing countries in response to petitions by domestic industries. The number of antidumping investigations has increased significantly over the last 25 years (table 5.1). Whereas in the late 1960s and early 1970s about 40 cases were brought each year, by the late 1980s that average had reached 140 cases a year, more than a threefold increase. In the early 1990s, the number of antidumping investigations increased further still, to around 200 cases a year.<sup>1</sup>

Antidumping investigations have been undertaken most frequently by Australia, Canada, the United States, and the European Union. In 1904 Canada enacted the first antidumping laws, followed by Australia in 1906 and the United States in 1916 and 1921. Between 1969 and 1993, these three countries and the European Union were responsible for almost 90 percent of the 2,770 antidumping investigations (the United States accounted for 29 percent; Australia, 27 percent; Canada, 17 percent; and the European Union, 15 percent).

The number of countries that have enacted antidumping laws has also increased markedly, according to the GATT secretariat: from 24 countries in 1990 to more than 40 by 1993. At the same time, several industrializing countries, notably Brazil, Mexico, and the Republic of South Korea, have become very active in using antidumping

measures (table 5.2). These three countries accounted for approximately 15 percent of the total antidumping investigations between 1990 and 1993. Mexico was the third most frequent user between July 1991 and June 1992.<sup>2</sup>

### Determination of dumping

Dumping has two definitions: export sales below home-market price and export sales below cost. GATT Article VI defines dumping as sales below "normal value," which in turn is defined as the comparable price, in the ordinary course of trade, for the like product destined for domestic consumption. Thus normal value is home-market price when home-market sales are in the ordinary course of trade. As explained later, home-market sales below cost are not considered to be in the ordinary course of trade.

Export sales below home-market price are generally understood to indicate that exporters are engaged in international price discrimination. However, discrimination exists only if export and home-market prices are compared in a symmetric manner. It is now well established that the current antidumping practices of the four major user jurisdictions are biased toward a finding of artificially high dumping margins—and consequently international price discrimination—even when no discrimination exists.

The "dumping margin" is the maximum level of the duty that the importing country can impose on dumped

TABLE 5.1  
Antidumping investigations initiated by signatories to the GATT Antidumping Code, 1969–93

Initiator	1969–74	1975–79	1980–84	1985–89	1990–93 <sup>a</sup>	Total
Australia	0	120	242	180	204	746
Canada	42	74	176	115	66	473
European Union	19	55	138	101	90	403
United States	125	140	146	219	183	813
Other	39	64	10	74	148 <sup>b</sup>	335
Total	225	453	712	689	691	2,770
Average cases per year	38	91	142	138	197	

a. Through May 1993.

b. Two-thirds is accounted for by Brazil, Mexico, and the Republic of Korea.

Source: GATT documents as reported by the Industrial Structure Council of Japan (1994).

TABLE 5.2  
Antidumping investigations initiated by industrializing countries, 1988–93

Country	1988	1989	1990	1991	1992	1993 <sup>a</sup>
Brazil	1	1	3	2	13	..
India	0	0	0	0	8	0
Korea, Republic of	0	1	6	0	5	2
Mexico	12	7	12	10	25	21

.. Not available.

a. Through May 1993.

Source: GATT documents as reported by the Industrial Structure Council of Japan (1994).

imports. The GATT Antidumping Code recommends that the duty be less than the dumping margin (the “lesser duty” rule), if this amount is adequate to remove the injury to the domestic industry. Australia and the European Union have adopted this rule.<sup>3</sup> To calculate the adequate duty level to remove injury, the export price is compared either with the price of the domestic product

of the importing country or, if such a price is depressed, with the full cost of production plus a “reasonable profit” for domestic producers.

Table 5.3 summarizes the three major sources of bias in the methods used by the United States and the European Union to establish price discrimination. First, in calculating the dumping margin, each individual export

TABLE 5.3  
**Systematic biases in calculating dumping margins**

<i>Current practice</i>	<i>Biases for high dumping margins in US and EU practices</i>	<i>Provision of 1994 Antidumping Code</i>
<b>Asymmetric price comparison</b>		
Averaging and zeroing	Each individual export price is compared with the average home-market price, with negative dumping margins in such comparisons being treated as zero margins in calculating the overall dumping margin.	The comparison generally is to be made either on a transaction-to-transaction basis or on an average-to-average basis.
Asymmetric adjustment of sales cost	All of the sales cost, including the profit of the related distributor in the case of the European Union, is subtracted from the export price. There are restrictions on the subtraction of sales cost from the home-market price (only direct sales cost can be deducted in the case of the European Union) in order to derive prices on the ex-factory basis.	No substantive changes. Calls for a fair comparison and for due allowances to be made for the differences affecting price comparability, as does the old code.
Disregarding below-cost home-market sales in the calculation of the dumping margin	If more than 10 percent of home-market sales are below cost during the investigation period, all below-cost home-market sales are typically disregarded for the calculation of the average home-market price in the United States. (This threshold is 20 percent in the European Union.)	Conditions that permit the treatment of below-cost sales as not in the ordinary course of trade are specified.
<b>Frequent use of inflated constructed value</b>		
Automatic presumption that home-market sales below cost are not in the ordinary course of trade	Home-market sales below average total cost during the investigation period (six months to one year) are automatically presumed to justify the use of constructed value, typically when 90 percent or more sales are below cost in the case of the United States.	No substantive change.
Short investigation period for calculating constructed value	A normal value is calculated from the production and cost data of the short investigation period (six months to one year). Typically, no adjustments are made for either business cycles or developments during product life span, such as learning curve effects.	Introduces a special provision for the start-up period.
Artificially high overhead cost and profit margin used for calculating constructed value	In the United States there are artificial minimum floors for overhead cost (a minimum 10 percent of production cost) as well as for profit (a minimum 8 percent of the total cost).	Introduces a provision requiring the use of cost and profit standards based on actual data, when feasible
Asymmetric adjustment of sales cost	The same biases shown for asymmetric price comparison, above, are created.	No substantive change.

Source: Jackson and Vermulst (1990), Boltuck and Litan (1991), and the Final Act of the Uruguay Round.

price is compared with the average home-market price. In such a calculation, the negative dumping margins (that is, the excess of export price over the average home-market price) are treated as zero margins; thus they are not balanced against the positive dumping margins. Consequently, dumping is bound to be identified—even if export price is equal to home-market price on average—whenever there exists some variation of export prices across transactions during the investigation period.

A second source of bias is the asymmetric adjustment of sales cost in deriving home-market and export prices on the ex-factory basis. Although all of the sales cost is deducted from the export price, there are restrictions in the deduction of sales cost from the home-market price. The third source of bias is the practice of calculating the average home-market price based only on the remaining above-cost sales, disregarding home-market sales below cost. This practice is based on the view, discussed below, that below-cost sales are not in the ordinary course of trade.

Sales below cost have not been regarded by the four major user jurisdictions of antidumping laws to be part of the normal course of trade since an informal agreement in 1979 during the GATT Tokyo Round. The revision of the Antidumping Code in the Uruguay Round authorizes this view.<sup>4</sup> The standard used to judge whether sales are below cost is the full cost of production and sales, including fixed and variable costs of production as well as selling, general, and administrative costs. When there are extensive below-cost sales, “constructed value” is used as a normal value. Constructed value is the full cost of production and sales plus profit. The frequent use of inflated constructed values has led to the finding of artificial dumping as well as to artificially high dumping margins.

Although dumping is still widely perceived as a form of international price discrimination, in practice below-cost sales have become an increasingly important determinant of dumping. More than 60 percent of all U.S. antidumping cases since 1980 have been based at least in part on allegations of sales below cost (Horlick 1990)—a clear reflection of the increased restrictiveness of cost standards. (See Finger 1992b and Horlick 1990 for historical accounts.) If sales are below cost during the investigation period, the current practice is to use constructed value almost automatically.<sup>5</sup> Moreover, adjustments are rarely made for either business cycles or product cycles, and artificially high profit rates are often used for the calculation of constructed value.

The frequent use of cost standard may also reflect the globalization of competition of capital and R&D-inten-

sive industries. Indeed, industries with higher capital and R&D intensity seem to be involved in more dumping disputes. Table 5.4 shows the industries that most frequently bring charges based on antidumping laws in Canada, the European Union, and the United States. These indus-

TABLE 5.4  
Major user industries of antidumping laws in Canada, the European Union, and the United States

Canada, 1980–91		
Industry	Antidumping cases initiated	
Primary metals	35	(23)
Electrical machinery	18	(12)
Chemical and petroleum	15	(10)
Metal products	12	(8)
Food and beverages	11	(7)
Subtotal	91	(59)
Total	155	(100)
European Union, 1980–89		
Industry	Antidumping cases initiated	
Chemical products	161	(42)
Primary metals	57	(15)
Nonelectric machinery	34	(9)
Electrical machinery	33	(9)
Wood products	19	(5)
Subtotal	304	(79)
Total	385	(100)
United States, 1979–89		
Industry	Antidumping cases initiated	
Primary metals	185	(41)
Chemical products	69	(15)
Metal products	39	(9)
Nonelectric machinery	27	(6)
Electrical machinery	24	(5)
Subtotal	344	(76)
Total	451	(100)

Note: Industrial classifications roughly follow the two-digit Standard Industrial Classification (SIC). Numbers in parentheses are percentages of total cases.

Source: OECD (1993).

tries, which account for 60 to 80 percent of all antidumping investigations, are all relatively capital- or R&D-intensive with the exception of the wood products and food and beverages industries. The primary metals and chemical products industries are jointly responsible for more than 60 percent of the antidumping cases in the United States and the European Union, and together with electrical machinery, they appear among the top five user industries in all three jurisdictions.

#### *Definition of domestic industry and standing*

Determining the scope of the domestic industry competing with imports is necessary to evaluate injury. The Antidumping Code defines "domestic industry" as a group of domestic producers (that is, firms engaged in local production) that produce the whole or a major proportion of like products (that is, similar to those allegedly being dumped).<sup>6</sup> The scope of the domestic industry is in turn determined by the scope of the like product. "Like product" as used in the code implies physical rather than functional likeness.<sup>7</sup> This interpretation, if adopted, would lead to a narrower definition of the market than that adopted in antitrust analysis, which focuses on substitute products based on their price elasticity of demand or consumers' response to a sustained price increase. In practice, however, the scope of like products has often been interpreted broadly.<sup>8</sup> When an affirmative determination of injury from import is relatively easy to obtain, there are strong incentives for domestic producers to argue for a broader definition of the like product. Such an interpretation of like product has occasionally resulted in the imposition of antidumping duties on products that domestic producers could not supply competitively.

An antidumping investigation is initiated when a firm that has standing brings a claim on behalf of the domestic industry. To have "standing," the firm requesting the investigation first must produce like products. This has become an important issue, especially as firms globalize their operations. According to the current GATT rule, the final assembler of product components has no standing to request an antidumping investigation of imported components unless the assembler also produces components domestically.<sup>9</sup> In the United States, however, standing has been assumed to exist for any petition filed unless a majority of the industry expresses opposition (Horlick 1990).<sup>10</sup>

Second, the petitioning firms must secure support from domestic industries. The GATT only recently provided guidelines on the level of domestic industry support necessary for a petitioning firm to obtain standing. The Antidumping Code, as revised in the Uruguay Round,

provides relatively clear albeit arguably weak conditions on standing: Domestic producers supporting the petitioning firm's case must dominate those opposing and must account for at least 25 percent of domestic production.

#### *Determination of material injury*

GATT Article VI states that "dumping is to be condemned if it causes or threatens material injury to an established industry or materially retards the establishment of a domestic industry." As clarified in the Antidumping Code, antidumping duties may be levied only against injurious dumping. But the code provides no clear definition of "material injury."<sup>11</sup> The code specifies two major factors that must be taken into account in the determination of injury: (a) the volume of the dumped imports and their effect on prices in the domestic market for like products, and (b) the consequent impact of the imports on domestic producers. With respect to the volume of the dumped imports, the code stipulates that whether there has been a significant increase in dumped imports, either absolute or relative, must be considered. However, the code does not stipulate that such an increase is a necessary condition for a finding of material injury caused by dumped imports. The Antidumping Code, as well as the national legislation based on it, therefore allows a very broad interpretation of material injury.<sup>12</sup>

The room for interpretation of the meaning of material injury is illustrated by the fact that commissioners of the U.S. International Trade Commission, an official body responsible for injury determination, have diverged widely on their findings on the degree of injury in the same antidumping cases. Among the 14 commissioners studied, three found material injury in less than 30 percent of the cases, whereas four found injury in more than 80 percent (Baldwin and Steagall 1993). Although the commissioners' voting behavior clearly reflects their individual trade policy orientation, it is the vagueness of the definition of material injury that allows for such wide variations.<sup>13</sup>

As argued in the next section, dumping that does not divert business from domestic industry to foreign exporters is unlikely to harm the importing country's welfare even under imperfect competition. However, according to the current GATT antidumping regulation, such dumping can nonetheless be judged as injurious since by reducing the domestic price it results in lower domestic industry profits.

The Antidumping Code explicitly states that injuries caused by other factors must not be attributed to dumped

imports. But since the code does not specify a significant increase in dumped imports as necessary to prove material injury, levels of total import—covering both dumped and undumped imports—as well as the general economic conditions in the importing country can significantly affect the outcome of material injury investigations. In fact, such has been the case in decisions by the U.S. International Trade Commission. Baldwin and Steagall (1993) found that a higher ratio of import penetration increased the probability of an affirmative decision, even controlling for the impact of the rate of increase of the dumped imports. Similarly, their analysis of countervailing duty cases showed that the real GDP growth of the U.S. economy has significantly affected the probability of affirmative decisions on serious injury.

To judge the existence of material injury when imports are dumped by several exporters from a single country or from different countries, the major user countries assess the effect on the domestic industry on a cumulative basis. Even if each individual exporter does not cause material injury, antidumping measures can still be applied.<sup>14</sup>

In the application of competition policy, however, injury must be demonstrated for each defendant unless there is collusion among the defendants. Cumulation, therefore, is clearly not consistent with competition policy. Yet the new Antidumping Code authorizes the practice of cumulation under broad conditions.

#### *Implementation of antidumping measures*

Not all antidumping investigations lead to the imposition of antidumping measures. Some investigations are never concluded. An antidumping investigation may be suspended or terminated if the exporter voluntarily raises its export price or ceases to export. (The Antidumping Code stipulates that any price increase should not be higher than necessary to eliminate dumping margins.) At the request of the exporter or the authorities of the importing country, the injury investigation can be continued. Since the completion of an injury investigation is not mandatory, however, the possibility exists that the exporter will raise its prices when a full investigation would have found that the domestic industry suffered no material injury.

Among the four major users of antidumping laws, only the European Union makes extensive use of price undertakings (that is, commitments by exporters to cease dumped exports). During 1980–89 the number of EU price undertakings was more than 60 percent higher than the number of duty impositions (see Bourgeois and Messerlin in OECD 1993). The United States has used

price undertakings only in rare situations. At the same time, not all affirmative cases have resulted in the imposition of antidumping duties. When such duties could seriously harm the U.S. economy (for example, steel and semiconductor cases), settlements have been arrived at through quotas (such as voluntary export restraints) or special pricing schemes (such as trigger-price mechanisms).

Antidumping investigations may also be terminated by private settlements. The petitioning firm may be willing to withdraw its complaint if the exporter raises prices to the petitioner's satisfaction. During 1979–89 one-quarter of the cases brought by the United States were withdrawn before definitive decisions had been reached (see Shin in OECD 1993). Private settlements, however, infringe on the antitrust law of the importing country when they involve an agreement among domestic and foreign firms for higher export prices.

When an antidumping investigation goes forward and reaches definite conclusions on both the dumping margin and the material injury, the importing country can impose an antidumping duty. U.S. law makes the imposition of a duty mandatory, whereas both Canada and the European Union allow for administrative discretion based on the "public interest." According to the EU law, the most important determining factors include the interests of the domestic industry, users, and consumers.<sup>15</sup>

Yet public-interest considerations have rarely affected the imposition of antidumping duties in either Canada or the European Union. The interests of the European Union, for example, have in practice tended to be equated with those of the industries protected by antidumping measures (see Bellis 1990). Nevertheless, duties were not imposed in several cases because of concern that downstream industry would be harmed.<sup>16</sup> Moreover, in a June 1992 decision in *Extramet Industries v. the Council of the European Communities*, the European Court of Justice ruled that the council had failed to give proper consideration to possible distortion of competition in the European Union and ordered the duty annulled.<sup>17</sup>

There are major international differences in the method of assessing antidumping duties. Whereas duties are prospective in Australia, Canada, and the European Union, they are retrospective in the United States. In the case of prospective duties, importers know the amount of antidumping duty they will be required to pay before they import the goods—a major advantage. Australia and Canada calculate the duty as the difference from the predetermined normal value; the European Union calculates the duty as a fixed percentage of the import price. In a

retrospective system, by contrast, the duty is determined only after goods have been imported and an annual review has been conducted. The uncertainty of this system discourages imports. A major advantage, however, is that retrospective duties can reflect subsequent changes in home-market price and production cost. When normal value declines, the importer is assessed a correspondingly smaller antidumping duty even if the export price remains the same.

In the Australian, Canadian, and U.S. systems, importers can avoid paying antidumping duty if the exporter raises its export price to the level of normal value. Although there is a refund provision in the EU regulation, few refund applications are made because the provision is quite restrictive (Bellis 1990). It requires the exporter to raise its price by the sum of the dumping margin and the antidumping duty when the importing company is related to the exporter.<sup>18</sup> Consequently, antidumping duty is levied in the European Union even if the dumping margin is absent for actual imports.

Sunset clauses in Australia, Canada, and the European Union automatically terminate the antidumping measures within a specified period (five years in Canada and the European Union, three years in Australia). Because a sunset clause does not exist in the United States,<sup>19</sup> U.S. antidumping orders have remained in effect considerably longer than those in Canada and the European Union.<sup>20</sup> The new Antidumping Code that resulted from the Uruguay Round introduced a sunset provision requiring antidumping orders to be terminated within five years unless termination would likely lead to both dumping and injury.

### **Welfare implications of antidumping policy**

Antidumping policy can be evaluated in terms of its effect on both global and national welfare. Global welfare is the sum of the economic welfare of both the importing and the exporting country. National welfare as used here means the economic welfare of the importing country. Global welfare approximates national welfare when the importing and the exporting countries commit to identical antidumping rules and apply the rules similarly.

Because the GATT enables such mutual commitment by national governments, the GATT rule on antidumping policy is best evaluated in terms of global welfare. By contrast, because the GATT does not oblige its signatories to use antidumping measures—and each country can use its own discretion within the boundaries set by the GATT—national welfare must also be considered. The discussion that follows focuses first on the global welfare implica-

tions of international price discrimination and sales below cost and then analyzes the welfare implications of antidumping policy on the importing country.

### *International price discrimination*

When an exporting firm faces more elastic demand in the export market, it sets its export price below its domestic price. This normal, profit-maximizing response does not imply anticompetitive motivation. Demand might be more elastic in the export market if a product mirrored home-market preferences better than it did export-market tastes. Domestic consumers then might be willing to pay a higher price than foreign consumers would pay.

Also, the exporter usually has a smaller market share in the export market than it does in its home market (due to transportation and other export-related costs). In this situation, the exporter may be willing to accept a lower price–cost margin in the export market. And if the importing country is a large economy, enabling many firms to profitably enter the market, the market in the importing country may be more competitive than the exporter's home market. In this situation, too, the exporting firm would have a smaller market share and might accept a lower margin in the export market.

Setting the export price below the domestic price—that is, international price discrimination—is possible only if there are costs or restraints to international arbitrage, such as high transportation costs, trade barriers, or resale restrictions by suppliers. International price discrimination is not consistent with maximum global welfare. When the home-market price (PD) is higher than the export price (PE), the marginal switch of sales from the export market (E) toward the home market (D) increases global welfare directly by the amount (PD – PE), since price in each market signifies the marginal value of consumption and invites expansion of the import-competing industry in the export market, further increasing global welfare.<sup>21</sup>

Yet prohibiting international price discrimination through antidumping regulation does not guarantee an improvement in global welfare. Antidumping regulation can reduce global welfare by reducing global output if, to satisfy the regulatory constraint, the exporting firm increases the export price without lowering the home-market price. The extent to which the exporter raises the export price and lowers the home-market price depends on a number of factors, including the market's size and the price elasticities of demand.<sup>22</sup> The larger volume of home-market sales makes it more attractive for the exporting firm to raise the export price, whereas the more

elastic export demand (a cause of dumping) makes lowering the home-market price more attractive.

One factor favoring an export-price increase is that antidumping action is permitted only if both dumping and injury to the importing country's domestic industry can be proved. Insofar as increasing the export price can relieve both constraints—while reducing the home-market sales price does not—antidumping regulation encourages higher export prices rather than lower home-market prices.

#### *Sales below cost*

A firm may set its export prices below cost without predatory intent in several situations. In all of these, the economic cost perceived by the exporting firm becomes significantly lower than the accounting cost of production, and competition results in below-cost sales by forcing the firm to price close to its economic cost.

- In industries with a high proportion of fixed and sunk costs, market prices may go below the accounting cost, particularly when demand is depressed and excess capacity develops. Such dumping, often called cyclical dumping, is most likely to be observed in industries that are both capital-intensive and cyclical (for example, the investment goods industry) or that have relatively rigid employment levels.
- When there is a learning curve for either production or consumption, the true marginal cost is below the current marginal cost of production, since current production generates information useful in reducing future production costs. Moreover, a firm may need experience merely to know the level of its own productivity and to be able to make correct production decisions in the future (see Clarida 1993).
- A firm's cost burden per unit of production during start-up or expansion (for example, amortization of R&D, capital goods investment, and other fixed costs) is significantly greater than it is over the life of the product. As a result, the economic cost of production may be significantly lower than the accounting cost of production, and goods may be priced below their accounting cost.
- The economic value of a firm's investment can depreciate significantly under unfavorable economic conditions. For example, the emergence of a competing product may make the existing product obsolete, or appreciation of the domestic currency may lower the value of export-oriented investment. In such cases, the exporting firm may be unable to price its product high enough to recover its initial investment.

When export prices fall below cost due to these non-predatory motivations, antidumping measures reduce

global welfare by forcing the exporting firm to increase its export price—and thereby reduce supply. Although import-competing firms might respond by expanding production, this increase would not compensate for the contraction of supply by the exporting firm. Even if antidumping measures are not actually applied, they reduce global welfare because the fear of an antidumping suit can force an exporting firm to restrict its investment.

When import-competing industries have significantly lower marginal costs of production than exporting industries, an antidumping measure may improve global welfare. If, for example, the importing country faces serious unemployment problems, and thus has a very low (social) marginal cost of production (relative to that of the exporting country), antidumping actions may increase global welfare by shifting output to the importing country and reducing unemployment, even if global output declines. Such a possibility does not provide a justification for the antidumping action per se, since neither the dumping nor the antidumping response is intrinsically linked to conditions in the labor market. Measures directly targeting the sources of unemployment are preferable.

Sales below cost may take place *with* predatory intent—that is, an exporting firm may seek to drive competitors out of business by increasing its supply to such an extent that the market price falls below the marginal cost of production.<sup>23</sup> Then, once the exporting firm has monopolized the market, it may raise its price to obtain monopoly profits. But predatory dumping seems only a rare possibility. To be a rational strategy, both concentrated market structure and high entry barriers are needed. Yet, as a study by the Organization for Economic Cooperation and Development (OECD 1993) found, in most U.S. and EU antidumping cases, the relevant domestic market was competitive, the import share was low, and there were several competing foreign enterprises, often from many countries. Nonetheless, if predatory dumping did occur, it would most likely reduce global welfare because overproduction takes place in the predatory stage, and significant underproduction occurs once the firm has gained a monopoly.

It is clear that current antidumping policy overprotects domestic industry against the risk of predation. Antidumping regulations usually do not take into account the extent of actual or potential competition, and they evaluate the pricing behavior of foreign enterprises based on the full cost of production, not the marginal cost. (The concluding section of this chapter suggests reforms that would make antidumping legislation more consistent with competition.)

*Dumping and the importing country*

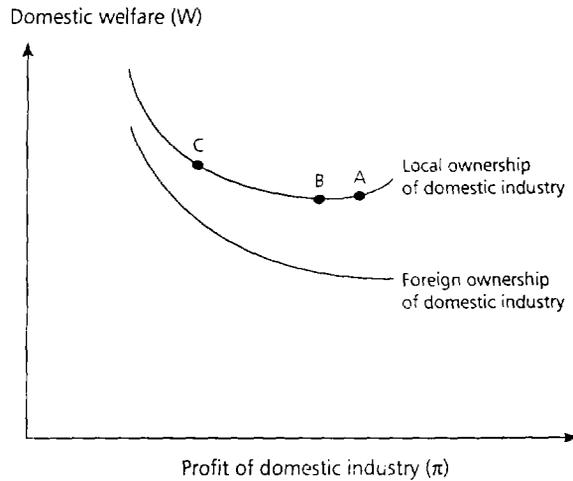
The GATT assigns to the importing country the explicit right to take antidumping measures.<sup>24</sup> Yet antidumping measures reduce the welfare of the importing country even more than they do global welfare. First, unrestricted imports are an important source of competitive discipline, with or without dumping, especially for smaller economies with limited domestic competition. Antidumping measures would enable those domestic enterprises whose competitive positions have fallen relative to that of foreign enterprises to recover lost markets and profitability without making productivity improvements.

Second, low import prices in principle improve the welfare of an importing country. The injury-related welfare cost to the domestic industry is smaller than consumers' and user industries' welfare gain, as long as domestic distortions are small. Injury to the domestic industry results from both the fall in prices (price injury) and the fall in output (output injury). Price injury is always offset by an equivalent consumer gain, and consumers can also enjoy the added benefit of the low import price—terms of trade gain. Output injury is bound to be negligible relative to the terms of trade gain because the output price is close to the marginal cost of production (again, assuming domestic distortions are small).

Several domestic distortions can make output injury non-negligible in static welfare calculation. Injury to the domestic industry may exceed consumers' welfare gain when the following distortions are large and the level of imports low:

- *Noncompetitive product markets.* When the domestic market of a specific industry in the importing country is not competitive even if import is free (for example, in the case of a globally oligopolistic industry), price exceeds the marginal cost of production. As a result, the decline in output in this industry leads to a reduction in rent.
- *Noncompetitive factor markets (particularly labor markets).* When wage is set noncompetitively due to either the monopoly power of unions or efficiency wage considerations in a specific industry, price again exceeds the true marginal cost of production. The decline of employment in the industry leads to lower workers' wage "rent."
- *International differences in production cost structures and labor market incentives.* When the importing country has a synchronous business cycle with the exporting country in certain sectors, free trade can increase unemployment in the importing country through dumping during business downturns in those sectors (see Ethier 1982). This scenario may result if the industries of the exporting country have a high proportion of fixed costs, and the

**Figure 5.1 Effects of dumping on domestic welfare and industry profits under international duopoly**



Note: Cournot-Nash equilibrium is assumed for duopoly competition.

importing country tends to generate large unemployment during business downturns.

When domestic distortions are substantial, dumping may reduce the economic welfare of the importing country because the welfare effect of output injury becomes non-negligible. Such possibility, however, does *not* justify current antidumping policy for several reasons. First, the effect of dumping on the welfare of the importing country can be made positive if domestic distortions can be sufficiently reduced. Although such interventions are not always possible, some distortions, such as entry regulations by the government to protect noncompetitive markets or excessive unemployment compensation, are policy generated and therefore can be reduced by the government.

Second, the effect of dumping on the welfare of the importing country does not necessarily become more negative as either the injury to the domestic industry or the dumping margin increases. When the injury to the domestic industry is large due to low import price, the gain for consumers and for user industries also tends to be large. Moreover, the latter gain becomes increasingly important as import price declines because the level of imports increases.

Figure 5.1 illustrates how the economic welfare of the importing country changes as the export cost of the foreign industry changes, assuming the Cournot-Nash equilibrium of duopoly competition.<sup>25</sup> When the domestic monopoly firm suffers a small injury (A → B), domestic welfare (W) also declines; when it suffers a large injury (A → C), domestic welfare increases because the size of imports becomes large. As this figure suggests, there is no

uniform relationship between the injury to the domestic industry and the economic welfare of the importing country. Globalization of industry ownership tends to further weaken this relationship.

Third, if antidumping measures are taken only when they increase the welfare of the importing country, they often harm the welfare of the exporting country more than they benefit the importing country. This is because antidumping measures restrict global output and thereby tend to reduce global welfare. Since every country both imports and exports goods, all would stand to gain from the restrained use of these measures.

One might question whether government revenues from antidumping duties make the net effect of antidumping measures positive for the importing country. Optimal tariff theory suggests that the welfare of the importing country increases if the country is able to improve its terms of trade by imposing a tariff because exporters may absorb the tariff to maintain their market positions. This conclusion does not apply to antidumping duties, however, since the size of the duty is determined endogenously by the dumping margin. Exporters have no incentive to reduce their export prices after the antidumping duty is imposed because a lower export price is completely offset by a larger duty and has no effect on the duty-inclusive import price, which would equal the normal value.

### **New issues and recent debates**

Recent debates on antidumping policy have focused strongly on its relationship with competition policy. There are several reasons for this focus. First, contemporary experience as well as economic analysis have uncovered that the anticompetitive effects of antidumping law can be much stronger than suggested by conventional analysis. Second, there have been several developments to substitute antidumping law with regional application of competition law. Third, some seek new justification for antidumping law in the global enhancement of competition.

#### *Strategic use of antidumping law as an anticompetitive device*

Antidumping law can be used in an anticompetitive manner: first, as a facilitating device for joint price hikes; and second, as a strategic weapon of a domestic firm to exclude foreign competitors.

*Facilitating device for joint price increases.* Since an antidumping measure, once introduced, forces the exporting firm to raise its export sales price by setting a minimum price, the measure severely limits price compe-

tion in the domestic market of the importing country. When the price of the exporting firm's product increases in a credible way, it is likely that domestic competing firms will also raise their prices. Although such an effect is anticompetitive, it may be inevitable if removing the material injury to the domestic industry is considered necessary.

The anticompetitive effects of antidumping law, however, can be much stronger than those caused by the unilateral price increase of the exporter in response to the imposition of duty. As Prusa (1992) has pointed out, a domestic firm may use antidumping law both as a threat to force an exporting firm to raise its prices and as a cover from domestic antitrust law in order to implement coordinated price increases. Based on a study of U.S. antidumping and countervailing duty cases in 1980–81, Prusa reported that even in cases in which petitions were withdrawn, imports declined as much as they did when duties were actually levied. The exporting firm may choose to increase its price rather than to incur the costs associated with a dumping investigation and the risk of high antidumping duties. Moreover, the threat of antidumping action may serve as an effective deterrent against deviation from a price cartel by domestic and foreign firms (Staiger and Wolak 1994a). Thus, even in cases where material injury is not likely to be established, the antidumping law has the effect of facilitating joint price increases by competing firms in the market of the importing country.

An explicit agreement between import-competing firms and exporting firms whereby the former agreed to withhold or withdraw antidumping petitions in exchange for price increases by the latter would constitute private restraint of trade and therefore violate the antitrust law of the importing country. The guidelines of the U.S. Department of Justice, for example, clearly state that "agreements among competitors that do not comply with the law, or go beyond the measures authorized by the law, do not enjoy antitrust immunity" (U.S. Department of Justice 1995).

What steps could be taken to reduce the risk of such anticompetitive effect of antidumping law? First, the government of the importing country should use tighter criteria in calculating the dumping margin, evaluating material injury, and determining the causality between the two. Tighter criteria would make it more difficult for an antidumping action to be used as a punishment device. Reducing the size of the antidumping duty—through, for example, the more disciplined use of the below-cost sales standard and the use of injury margin—is particularly important.

Second, to discourage sham petitions, petitioners should be required to submit substantial evidence before the government of the importing country initiates an investigation.<sup>26</sup> Third, competition policy should be made available as a deterrent to coordinated price increases. Antidumping petitions should not provide opportunities for domestic firms to exchange information so as to maintain high domestic prices. Nor should domestic and exporting firms be allowed to enter into an agreement for the increase of an export price.

*Predatory weapon.* A domestic firm can use antidumping law as a predatory weapon to shut out exports by foreign firms. By expanding its output, it could cause domestic prices to fall below foreign firms' current cost of production. Insofar as the antidumping law forces foreign firms to price their products above their current cost of production, they would be excluded from the market. Such a predatory strategy is rational if denying market share to foreign firms provides significant competitive advantage to the domestic firm, by enabling it to ascend the learning curve more quickly than foreign firms (see Gruenspecht 1988). Such an advantage may even enable the domestic firm to monopolize the domestic market.

Under normal circumstances, predation is rarely more profitable than accommodation, since predation requires a large expansion of output. Antidumping law can make such a strategy more viable, however, by allowing the domestic firm to exclude the foreign competitor by expanding its output only to the point where price falls below foreign firms' current accounting cost of production.

To reduce the chance that antidumping law will be used for predatory purposes, the below-cost sales standard should be used in a disciplined manner. First, the standard must take into account learning and other dynamic factors that make the accounting cost of production substantially larger than the true economic cost. Second, the competitive consequences of antidumping measures should be carefully evaluated where domestic markets are highly concentrated. Third, the competition policy authority of the importing country should be fully aware of the constraints on foreign competitors imposed by antidumping law, which allows a domestic firm to monopolize the market even if prices are significantly above the marginal cost of production. When antidumping law is binding, the standard presumption that predation does not occur if price exceeds the marginal cost of production does not hold.

### *International predation*

International predatory pricing could be regulated by antitrust law as well as by antidumping legislation. Some scholars believe that antidumping laws should be repealed and antitrust law used instead (see Ordoover, Sykes, and Willig 1983). Such substitution would be welfare enhancing since—due to its focus on injury to the domestic industry rather than injury to competition—antidumping law tends to overprotect domestic firms from the risk of predation.

Many industrialized countries and country groups, including the United States and the European Union, take the stance that domestic antitrust law is applicable to anticompetitive conduct by foreign firms, including predatory pricing, whenever domestic competition is restricted in an important manner.<sup>27</sup> But there are constraints on the international application of antitrust law. Because individual countries' antitrust authorities have no legal mandate in foreign jurisdictions, they are not allowed to conduct the investigations required to prove predatory pricing. Some countries, including Australia and the United Kingdom, have enacted statutes that block extraterritorial applications of competition law and can prevent domestic producers from complying with orders by foreign authorities, including providing information needed to prove anticompetitive conduct.

No such constraint exists with antidumping law, which enables the government of the importing country to collect the data it requires from foreign firms. Because GATT Article VI allows the government of the importing country to impose antidumping duty based on its own judgment, the government can impose duties on an exporting firm unilaterally, based on available information.<sup>28</sup> Under the GATT, the action of the importing country cannot be blocked by the government of the exporting country. As for the remedy, the antidumping law may be as effective as antitrust law. The process leading to the imposition of the duty is relatively swift, and the duty forces the foreign firm to raise prices above full cost through a customs-clearing process. Unlike antitrust law, however, antidumping law does not impose punitive measures such as treble damages or surcharges.

The constraints on the international application of antitrust law could, however, be overcome by international agreements. In fact, some regional arrangements—such as the European Union, the European Economic Area, and the Australia/New Zealand Closer Economic Relations Trade Agreement (ANZCERTA)—have led to the suspension of intraregional applications of antidumping laws, with the understanding that antitrust law can be

effectively applied on a regionwide basis. ANCERTA, which took effect July 1, 1990, seems to have succeeded in removing antidumping procedures under the least common institutional setup. It empowers the competition policy authorities of the two countries to obtain evidence from and issue orders to firms in the other country. But suspension of antidumping law has generally taken place only in the context of fairly deep economic integration, since it requires not only harmonization of competition policy within a region but also mutual recognition of extraterritorial or supranational application of competition law. The U.S.–Canada Free Trade Agreement and the North American Free Trade Agreement have not succeeded in replacing antidumping law by antitrust law.

Even if full substitution of antidumping law by antitrust law is not feasible, reform of antidumping law is still possible if it is agreed that the sole objective should be the prevention of international predation. This is because there are unreasonable discrepancies between antidumping policy and competition policy in terms of their standards for evaluating alleged anticompetitive conduct. Some discrepancies may be justified, as suggested by Ordover and others (1983). First, weaker foreign antitrust laws may permit substantially greater cooperation between firms in their home market, which in turn may make coordination of overseas activities easier. Second, many countries exempt export cartels from the applications of their antitrust laws. If ineffectiveness of antitrust laws results in collusion among foreign firms in the importing-country market that cannot be effectively prevented by the domestic competition law, foreign firms then must be viewed as a single entity rather than competing entities.

But such justifiable differences between antidumping and antitrust laws do not prevent the importing country from taking steps to make antidumping policy more consistent with competition policy, including the following:

- Using market structure standards (including entry barriers) to evaluate the risk of injury to competition, thus avoiding the use of antidumping measures when the structure is competitive or entry barriers are low.
- Cumulating material injury by many exporters only if they are in collusive predation.
- Using significantly tighter criteria on below-cost sales in evaluating predatory intent. For example, to be consistent with antitrust analysis, marginal cost should be used as the standard instead of full cost of production, and appropriate adjustments should be made for learning-curve and promotional motivations.

As another alternative to antidumping law, Deardorff (1990) has suggested that the importing country should tax away the monopoly profits gained by predation to encourage a foreign predator firm to abandon its strategy. The advantage of this approach is that the importing country could fully realize the gain of cheap imports as long as they were not predatory. Ex-post taxation may not be credible, however. First, the foreign firm that has monopolized the market can also threaten the importing country, since by suspending exports it could significantly harm the importing country's interest. Second, when the instrument available is limited to proportional import duty or subsidy, subsidization rather than taxation of imports may be the optimal policy. Finally, once competing firms have left the market, it may become difficult to collect evidence on predation.

#### *Globalization of industry and antidumping policy*

The globalization of industry poses three new issues for antidumping policy. First, it is claimed that an exporting firm can "circumvent" antidumping measures by shifting the location of final assembly or parts and materials processing from its home country to the importing country or to a third country. Both the United States and the European Union introduced "anticircumvention" measures in their national antidumping regulations to allow extension of antidumping duties to parts assembled outside the exporting firm's home country. Such measures have been controversial, however, since their consistency with GATT is questionable.

Second, the globalization of industry has increased the number of markets in which competition takes place. Industries compete in their domestic market in intermediate goods as well as final goods, and they compete in their domestic market as well as in third-country markets. As global sourcing of inputs has become an increasingly important competitive practice in electronics, automobile, and other industries, pressure to expand the scope of antidumping measures for input dumping has also increased. Globalized competition has also made third-country dumping an important issue, and preventing third-country dumping was one of the major points in the U.S.–Japan Semiconductor Agreement.

Third, the globalization of industry has made the identification of domestic industry with national ownership increasingly inadequate. Thus, antidumping measures that protect domestic production do not necessarily protect national enterprises. This last point may have important implications in those countries with industrial and regulatory policies targeting the development of national industry.

*Globalization and "circumvention."* The U.S. Omnibus Trade and Competitiveness Act of 1988 allows antidumping duty to be extended to imported parts and components from which a product similar to one subject to a U.S. antidumping order can be assembled or completed. The U.S. law requires no investigation to prove injurious dumping. It also contains an anticircumvention provision that allows antidumping duty to be extended to goods completed or assembled in third countries and then shipped to the United States, also with no proof of injurious dumping. The EU regulation includes a similar anticircumvention provision.

The consistency of these provisions with GATT is highly questionable. GATT Article VI allows antidumping duty to be imposed on imports only if injurious dumping has been established. In fact, the GATT panel ruled in 1990 that duties the European Economic Community had imposed on Japanese parts for anticircumvention purposes were unjustified and violated GATT Article III on national treatment. The panel also concluded that anticircumvention measures are not covered by GATT Article XX, which allows governments to take measures necessary to secure compliance with national laws or regulations. This is because Article XX does not allow governments to prevent enterprises from taking actions designed to avoid incurring an obligation, for example, by transferring production to the duty-levying country. Shifting the location of production in response to an antidumping duty, the panel held, cannot be viewed as a violation of the GATT.

Circumvention of antidumping law is desirable from a welfare standpoint, except when it involves predation. Because circumvention mitigates the restrictive effects of antidumping measures on competition and output, it generally increases global output and welfare as well as the welfare of the importing country. Circumvention efforts by exporting firms indicate that there is competition in the market.

*Global competition and antidumping policy.* Imposition of antidumping duty on parts and components can significantly affect the competitiveness of downstream industries. It could be argued that when dumping is in the form of pure international price discrimination (that is, a low export price relative to the home-market price), antidumping duties would simply offset the artificial competitive advantage of the importing country's downstream industry. As discussed earlier, however, current antidumping policy is significantly biased toward a finding of larger dumping margins and is dependent on artificially constructed values for determining margins.

Consequently, antidumping measures may well harm the international competitiveness of downstream industry. It is not surprising that U.S. firms such as IBM and Apple have expressed strong concerns about antidumping duties on semiconductors and flat panel displays.

The economic loss of downstream industry from antidumping duties, like consumers' welfare loss, is generally larger than the gain of the domestic parts and components industry. Thus, even from the viewpoint of producers, increasing global interdependency of the industries of different countries may make liberalization of antidumping policy more desirable. However, pressure has also increased to expand the use of antidumping measures against input dumping.<sup>29</sup> Pressures from global competition nonetheless should be used to promote liberalization of antidumping measures rather than their expansion.

Globalization of competition also has made third-country dumping an increasingly important issue. Third-country dumping is best understood using the example of three hypothetical countries—countries A, B, and C. Dumping by country A's industry in country C (third-country market) injures the export interests of country B. The industry of country B then demands removal of the injury from such dumping.<sup>30</sup> Article XIV of the GATT Antidumping Code provides a mechanism for addressing third-country dumping. It allows an importing country with no competing domestic industry (country C) to impose antidumping duty based on the request of another country (country B).

It is clear that third-country dumping both reduces the economic welfare of the competing exporting countries and increases the economic welfare of the importing countries, with positive net welfare in nonpredatory cases. From a global welfare point of view, it is therefore important to be cautious in using the provision of Article XIV. The importing country should respond to the request only when it judges that low import price harms its interest because it endangers competition.

*Diversification of ownership and antidumping policy.* The diversification of ownership due to direct foreign investment, especially in industrializing countries, has important implications for antidumping policy since such policy cannot discriminate on the basis of ownership. Most importantly, injury to domestic industry becomes an irrelevant criterion for evaluating the economic impact of dumping on the importing country (except, again, when dumping is predatory). If owners of capital are the dominant stakeholders of the domestic industry and if they are primarily foreign, the injury to the domestic industry is

excluded from the calculation of the importing country's welfare, unless competition is at stake. In this situation, the importing country can only gain from dumping, even if the market is not fully competitive (see figure 5.1). The increasing diversification of ownership therefore calls for antidumping policy to be focused on competition.

#### *The global competition rationale*

Many support the view that antidumping policy is justified as a corrective response against distortions of global competition by an exporting firm that faces either weak competition or a policy of strategic import protection by the exporting country in its home country. But is dumping a good indicator of these distortions, and can antidumping policy contribute to their removal?

The view that dumping is the product of weak competition in the exporting country was expressed by former chairman of the European Economic Community, Willy de Clercq:

Dumping is made possible only by market isolation in the exporting country, due primarily to such factors as high tariffs or non-tariff barriers, and anti-competitive practices. This prevents the producers in the importing country from competing with the foreign supplier on his own ground, while allowing him to attack their domestic market by sales which are often made at a loss, or are financed from the profits made from the sale of the same or different products in a protected domestic market (Financial Times, November 21, 1988).

It is clear that exporting firms protected by import barriers, such as quantitative restrictions, can engage in international price discrimination and set higher prices in the domestic market than those in the international market. Anticompetitive practices such as cartelization of the domestic market also result in domestic prices that are higher than international prices. Even if there are barriers to competition abroad, however, antidumping policy is generally not a solution. First, the importing country does not generally get hurt from cheap imports, as explained earlier. Second, the main effect of antidumping policy is to reduce competition in the importing country, which is antithetical to global competition policy. This effect has become more important with the increasing use of the below-cost sales standard in determining dumping margins.<sup>31</sup> Furthermore, international price discrimination is not always caused by the absence of competition in the exporting country. Absence of competition in the

importing country can also cause price-discriminating dumping, as Weinstein (1992) demonstrated recently.

Consider the situation in which country A has one firm and country B has two competing firms. Assume that all three firms have an identical unit cost of production ( $C$ ). If there is no international trade, country A has a higher domestic price ( $PA$ ) than country B ( $PB$ ). Also assume that there is a non-negligible transportation cost ( $t$ ). If  $PA > C + t > PB > C$ , the firm in country A finds its export to country B unprofitable, whereas the two firms in country B find their exports to country A profitable. Exports by the two firms in country B then can be deemed dumping, since both firms have a smaller market share in country A than they do in country B, implying a larger profit margin for domestic sales than for export sales. If, however, country A has two domestic firms instead of one, firms in country B will find their exports unprofitable, and no dumping will take place.

In this example, dumping is caused by the absence of competition in the importing country and helps to make the monopolistic market more competitive. As this example clearly shows, dumping is not a good indicator of the degree of competition in the exporting country's market. For antidumping policy to have a meaningful role as a global competition policy, one would need direct measures of the barriers to competition in the exporting country's market, and the antidumping measure would have to be contingent on the presence of those barriers.

Dumping also may be caused by an import restriction policy of the exporting country aimed at strengthening a strategic domestic industry that requires static or dynamic economies of scale. Market reservation provides advantages to the domestic industry in global competition. Willig (in OECD 1993) called dumping caused by such policy—either price discrimination or sales below cost—“strategic dumping.”

Strategic dumping may reduce global efficiency not only by its static effects on global output but also by its dynamic effects on the speed of cost reduction since import protection reduces global output, on which incentives for cost reduction depend. (This will certainly be the case if the industry injured by dumping has room for significant cost reduction through learning-by-doing whereas the protected industry has exhausted such opportunities.) Moreover, such dumping over time may also reduce the degree of global competition by permitting dominance by the protected firms.

However, here again, the essence of the problem is not dumping but the country's desire to protect the strategic industry. Focusing on dumping is counterproductive since

dumping, especially below-cost sales, frequently can occur in such industry even without home-market protection. This is because industries with economies of scale tend to have large fixed and sunk costs and large room for learning. Moreover, strategic dumping does not necessarily reduce either the welfare of the importing country or global efficiency.<sup>32</sup>

Finally, even if there are barriers to competition abroad, antidumping policy further reduces global welfare and the welfare of the importing country when it fails to eliminate or reduce such barriers—a highly likely outcome for importing countries with small domestic markets (a situation that would provide limited incentives for the exporting country to reduce protection). When an antidumping measure has no effect on the level of home-market protection of the exporting country, it ends up simply raising import prices. If international predation is the problem, the injury to the domestic industry will be removed and the risk of monopolization reduced. If competition is not at stake, however, such a price increase further reduces global efficiency since it leads to contraction of global output.

### Conclusions

Based on the preceding analysis, this section offers a set of policy recommendations and points to priority areas for research.

#### *Policy recommendations*

- Developing countries should be very cautious in introducing antidumping regulations. Even if these regulations are most rationally used, they tend to bring about a small benefit to the country that administers them since only in limited circumstances do dumped imports significantly harm the national welfare of the importing country. By contrast, antidumping regulations can cause large damage to the importing country when they are abused for protectionist purposes. The experiences of industrialized countries suggest that such risk is large.
- In introducing antidumping regulations, countries should adhere to the new Antidumping Code agreed to in the Uruguay Round as an element of basic discipline. Accession to the World Trade Organization will automatically oblige member countries to adopt the new code.
- Countries should introduce other elements of discipline as well, in view of the fact that the new code does not put sufficient constraints on antidumping measures. Weak discipline can harm developing countries, in particular, since in these countries importing is a more important source of the supply of goods as well as competition,

and foreign-owned firms often have a larger share of domestic supply. Moreover, they typically maintain higher conventional trade protections. Once they are full members of the World Trade Organization's antidumping committee, developing countries should actively advocate for further reform of the Antidumping Code.

- Developing countries should avoid imposing high antidumping duty since the risk of complete blockage of imports and of domestic shortages is high in economies with small domestic markets. This risk can be reduced by introducing more strict rules for calculating dumping margin. In particular, recourse to the cost standard of dumping should be avoided. Developing countries would be well advised to introduce the lesser-duty rule, as Australia and the European Union have done.

- In evaluating injury to the domestic industry, a significant increase in the volume of dumped imports should be considered as a necessary condition for the affirmative decision. Considering only the level of import as a measure of injury will permit domestic industry to seek redress through antidumping measures even when injury is due to domestic factors. Furthermore, it is generally true that the higher the import level, the more harmful is the antidumping measure to the national welfare.

- An injury investigation should be completed even when price undertakings are accepted. If the investigation does not prove dumping, import price ceilings should be withdrawn.

- Countries should introduce a public interest clause that allows them to forgo imposing antidumping duty when the cost to the national welfare is high. The loss to the downstream industry or to consumers resulting from an antidumping measure can far exceed the gain to the upstream industry, when import supplies a large share of the domestic market and domestic goods are poor substitutes for the imports.

- Countries should make clear that submission of substantial evidence on behalf of the domestic industry is required before the government will initiate an investigation, since the mere fact of the investigation can have an anticompetitive effect on the domestic market. An automatic sunset clause terminating antidumping measures within several years would also be a desirable feature (the new Antidumping Code has a five-year sunset provision).

- Finally, countries should make competition policy available as a deterrent to the abuse of antidumping law. Petitions for antidumping measures should not be used as vehicles for domestic firms to exchange information in order to maintain high domestic prices. Nor should domestic and foreign firms be allowed to enter into an

agreement to increase export prices. The consequences of antidumping measures for competition should be assessed in highly concentrated industries, and this assessment should be used in deciding whether the imposition of antidumping measures is in the public interest.

#### *Directions for future research*

The preceding discussion suggests four priority research tasks. The first is an assessment of developing countries' experiences in applying antidumping measures as importing countries. Since developing countries only recently began using antidumping laws, there is no systematic assessment of their experience, which may vary substantially from that of industrial countries, due to differences in size, market structure, level of industrial development, and share of foreign-owned firms.

The major questions to be addressed include:

- Why have developing countries become so active in using antidumping measures?
- What are the major features of their antidumping regulations? Have developing countries avoided the protectionist biases of some industrial countries' regulations? How do developing countries evaluate material injury to the domestic industry?
- When developing countries have applied antidumping legislation, what has been their experience? Have foreign exporters responded satisfactorily to requests for data? How high have the duties been? What has been the impact of duties on trade, domestic industry, competition, and the economy?

Second, an empirical assessment of the welfare effects of antidumping measures, focusing particularly on industrial countries, should also be conducted. Such an assessment could prompt the reform of antidumping policy, which would benefit both industrial and developing countries. The focus of antidumping law primarily on injury to the domestic industry has been an important cause for its drift toward protectionism. Although the U.S. International Trade Commission is working on such a welfare impact assessment, a parallel effort with a global perspective is strongly suggested.

Although many studies have been conducted on trade restrictions such as voluntary export restraints and multi-fiber agreements, few empirical welfare studies of antidumping measures exist since, until recently, the economic impact of antidumping measures may have been dominated by the other trade restrictions.<sup>33</sup> In addition, antidumping measures may have been viewed as perfectly legitimate responses against distortions in competition.

As pointed out in this chapter, however, neither is the case today.

This assessment could address the following questions:

- What have been the economic effects of antidumping measures on domestic industry and consumers?
- What have been the fiscal implications?
- What net welfare loss have antidumping measures caused?
- How have antidumping measures affected the economies of exporting countries?
- How much would each country gain (as both an importer and an exporter) from reciprocal reform of antidumping measures?

Third, there should be an attempt to clarify the following questions regarding antidumping regulations and their economic effects:

- How many antidumping petitions have been withdrawn? Why were they withdrawn? What has been the economic effect of the withdrawn cases?
- How is "like product" determined in practice, and how does it differ from the market definition of antitrust analysis?
- Why is the public interest clause so rarely effective in influencing antidumping measures?
- Do prospective duty collection systems have a different impact than retrospective systems? What is the economic impact of each type of system?
- What is the economic impact of price undertakings versus duty impositions?

Finally, to promote the reform of antidumping regulations, an empirical assessment of the economic impact of several possible reforms would be useful, including:

- Reducing the biases in dumping margin calculations by averaging and zeroing, asymmetric adjustment of sales cost, and use of constructed value.
- Making the criterion of material injury to domestic industry more consistent with economic welfare by accounting for both price and output injury (business diversion) and effects on competition.
- Substituting antitrust policy for antidumping policy.

#### **Notes**

The author thanks Jim Levinsohn, professor of economics at the University of Michigan, and J. Michael Finger of the World Bank for their valuable comments on this chapter, as well as Claudio R. Frischtak for his helpful suggestions.

1. The number of domestic as well as bilateral disputes concerning the consistency of antidumping measures with national antidumping laws as well as with GATT regulations also has been rising. This

reflects the fact that national antidumping regulations often include provisions allowing for a high degree of administrative discretion, which can be abused for protectionist purposes, as well as provisions that are inconsistent with GATT regulations. The general wording of GATT Article VI and the GATT Antidumping Code also have been a source of international disputes.

2. From June 1991 to June 1992, 202 antidumping investigations were begun by the five countries that were the most frequent users of antidumping measures: 76 cases by Australia, 62 by the United States, 25 by Mexico, 23 by the European Union, and 16 by Canada.

3. In the European Union the average duty was 17.8 percent, compared with an average dumping margin of 28.8 percent during the period 1980–89 (Bourgeois and Messerlin in OECD 1993).

4. GATT Article VI does not explicitly define “ordinary course of trade.” However, an explicit provision in the new Antidumping Code allows importing countries to treat below-cost sales as not being in the ordinary course of trade under certain conditions (see table 5.3).

5. Note that the current administrative standard on below-cost sales is often more restrictive than national regulations. The U.S. Tariff Act (Section 773) of 1930 as amended in 1974, for example, stipulates that sales below cost are considered outside the ordinary course of trade if they are made over an extended period of time (conditions that also were adopted in the new Antidumping Code). In 1987 the U.S. Court of International Trade (CIT) found grounds to criticize the practice of the U.S. Commerce Department, which automatically considered the existence of below-cost sales during the six-month investigation period to imply that cost recovery was not feasible within a reasonable period of time. The CIT also ruled in that same year that the practice of disregarding all below-cost home-market sales in calculating the dumping margin, once such sales reached 10 percent of the total, was not justifiable.

6. It is clear from this definition that ownership does not matter, so that foreign-owned firms should be able to seek redress through antidumping measures just as nationally owned firms do. Note in this regard that the U.S. Court of International Trade ruled in 1992 that the fact that a foreign-owned firm performs design and engineering abroad and imports major parts does not disqualify it as part of the domestic industry. See the discussion in the section on the globalization of industry and antidumping policy.

7. “Like product” is defined in the 1994 Antidumping Code as “a product alike in all respects to the product under consideration.”

8. Messerlin and Noguchi (1991) reported that the antidumping office of the European Union had identified only two markets in photocopier products, whereas the competition office had identified three.

9. In the United States, the 1988 Trade Act gave standing to such domestic assemblers in the context of anticircumvention. See the discussion in the section on the globalization of industry and antidumping policy.

10. However, in a 1990 dispute between Sweden and the United States about U.S. imposition of antidumping duties on Swedish

steel products, the GATT panel ruled that the absence of opposition by other domestic producers was insufficient to conclude that the petition had been made on behalf of the domestic industry.

11. The 1967 GATT Antidumping Code required the dumped imports to be a principal cause of the injury to the domestic industry. For the affirmative determination of material injury, however, this requirement was eliminated in the Tokyo Round.

12. U.S. law, for example, defines material injury simply as “harm which is not inconsequential, immaterial, or unimportant.”

13. The United States seems to have the most sophisticated system of injury investigations. Its International Trade Commission uses an econometric model to estimate the economic impact of dumped imports. However, the result of this analytical work does not seem to significantly influence the judgments of those commissioners who have low subject standards of material injury.

14. According to one view (see Bierwag 1990), it is not clear whether cumulation is fully consistent with the GATT, since the GATT provisions characterize dumping as a business practice of individual firms.

15. See Council Regulation (EC) No. 3283/94, December 22, 1994.

16. These cases involved wrought titanium from Japan (1979), furfural from China (1981), and acrylonitrile from the United States (1981).

17. In this case, the petitioning firm was the sole EU producer.

18. This is due to the EU requirement that all costs and profit incurred by a related importer, including the antidumping duty, be deducted in order to derive the ex-factory export price.

19. To obtain an order of revocation, an exporter must show no sales at less than fair value for two years and demonstrate no likelihood of resumption of dumping (Horlick 1990).

20. The Japanese Industrial Structure Council (1994) reported that 39 percent (22 of 56) of currently effective U.S. antidumping orders against Japanese exports have lasted for 10 years or more. There are no such cases in Canada and only one case in the European Union.

21. Given the profit-maximizing strategy of the exporting firm, the marginal switch of its sales between markets does not affect its profit.

22. The second-order effect on profit of the price deviation from the optimal level is proportional to  $2Q' + (P - C)Q''$  where  $Q'$  and  $Q''$  are the first and second derivatives of the demand curve, respectively. For demand with constant price elasticity, this formula is equal to  $-[(k - 1)^2/k] \times Q/C$  where  $k$  is the elasticity of demand ( $k > 1$ ) and  $C$  is the production cost.

23. If a firm's marginal cost of production is above price, it also is clearly above the marginal revenue of production. The firm could then increase its profit by reducing its supply, unless it expects the gain from predation.

24. The GATT is silent on the exporting country's right to take antidumping action. In a dispute between the European Union on the one hand, and Japan and the United States on the other, about the exporting country's right to take an antidumping measure, the GATT

panel did not take a definitive view. In a case involving the U.S.–Japan Semiconductor Agreement, the panel concluded only that the set of measures taken by the Japanese government to stop third-country dumping was inconsistent with GATT Article XI prohibiting the use of quantitative and other nonprice trade interventions.

25. In a Cournot-Nash equilibrium domestic and foreign firms choose their capacities in the domestic market simultaneously given their respective costs of production. The change of the marginal cost of production of the foreign firm ( $dC^*$ ) causes the supply changes of the domestic and foreign firms in the domestic market [ $dq = a(dC^*)$  with  $a > 0$  for the domestic firm, and  $dq^* = -a^*dC^*$  with  $a^* > a > 0$  for the foreign firm]. Given the price derivative of the demand by  $P'$  ( $= \partial P / \partial Q$  with  $Q = q + q^*$ ), the changes of the domestic consumers' surplus ( $CS$ ) and of the domestic firms' profit ( $\pi$ ) are given by

$$(1) d(CS) = -(q + q^*) P' (dq + dq^*) = (q + q^*) P' (a^* - a) dC^*$$

and

$$(2) d\pi = q P' dq^* = -q P' a^* dC^*.$$

Therefore there is a negative relation between consumers' surplus and the profit of the domestic industry:

$$(3) d(CS) = -(1 + q^*/q) (1 - a/a^*) d\pi.$$

The change in the national welfare ( $W = CS + \pi$ ) is given by

$$(4) dW = d(CS) + d\pi = [1 - (1 + q^*/q) (1 - a/a^*)] d\pi.$$

Equation (4) shows that there is a positive relation between welfare and the profit of domestic industry when  $q^*$  (the import) is small. On the other hand, if  $a^*$  is significantly larger than  $a$ , the relation between welfare and the domestic industry's profit turns negative when  $q^*$  (the import) becomes large relative to  $q$  (domestic production). In the case of linear demand,  $a/a^* = 1/2$ , so that the relation becomes negative when  $q^* \geq q$  (that is, the import supply becomes larger than the domestic supply).

26. An investigation itself has the effect of reducing price competition significantly since the exporting firm does not want to be found to cause injury to the import-competing industry by underselling during the investigation period (Staiger and Wolak 1994b). Nonetheless, an investigation is easily initiated in the United States, given only "notice pleading claims of dumping often with little more than U.S. import statistics and petitioner's own costs" (Horlick 1990, p. 111).

27. In 1986 in a suit brought in the United States, *Zenith et al. v. Matsushita et al.*, it was alleged that a group of foreign firms had engaged in collusive predatory pricing in violation of U.S. antitrust laws.

28. This power to impose duties unilaterally can, of course, be abused for the purpose of protection.

29. There was an attempt to introduce offsetting measures against diversionary input dumping in the 103rd U.S. Congress. If input used in the manufacture of a product had been purchased at a dumped price, the provision provided that the diversionary dumping benefit could be offset by antidumping measures. This proposal was not enacted, and it is highly questionable whether such provision is consistent with GATT Article VI.

30. Third-country dumping became a major issue in negotiations for the U.S.–Japan Semiconductor Agreement. The agreement, which called for a commitment by the government of Japan to stop its companies from dumping into third-country markets, was challenged by the European Economic Council and several countries as being inconsistent with the GATT.

31. The view that below-cost sales are financed by the profit gained by the exporting firm in its noncompetitive home market makes no economic sense. The exporting firm sells below costs since competition forces a low price which, however, exceeds its economic cost. The profit-maximizing export price is generally independent of the size of the profit made from domestic sales.

32. According to Willig (in OECD 1993), whether dumping is truly harmful will depend on (a) the existence of home-market protection; (b) the existence of static or dynamic economies of scale in the product supply; and (c) whether excluding exporters from the home market significantly affects rivalry. Although Willig does not explicitly mention the impact of such dumping on competition as a necessary condition for harmful strategic dumping (because he assumes a symmetric case), it is necessary to investigate as well whether such dumping tends to significantly reduce competition by, for example, further strengthening a dominant position of the exporter.

33. One of the few studies is that by Staiger and Wolak (1994b).

## References

- Baldwin, R. E., and J. W. Steagall. 1993. *An Analysis of Factors Influencing ITC Decisions in Antidumping, Countervailing Duty, and Safeguard Cases*. NBER Working Paper 4282. Cambridge, Mass.: National Bureau of Economic Research.
- Bellis, J. F. 1990. "The EEC Antidumping System." In J. H. Jackson and E. A. Vermulst, eds., *Antidumping Law and Practice*. New York: Harvester-Wheatsheaf.
- Bierwag, R. M. 1990. *GATT Article VI and the Protectionist Bias in Anti-dumping Laws*. Boston: Kluwer and Deventer.
- Boltuck, R., and R. E. Litan. 1991. *Down in the Dumps*. Washington, D.C.: Brookings Institution.
- Clarida, R. H. 1993. "Entry, Dumping, and Shakeout." *American Economic Review* 83: 180–202.
- Deardorff, A. V. 1990. "Economic Perspectives on Antidumping Law." In J. H. Jackson and E. A. Vermulst, eds., *Antidumping Law and Practice: A Comparative Study*. New York: Harvester-Wheatsheaf.
- Ethier, W. J. 1982. "Dumping." *Journal of Political Economy* 90: 487–506.
- Finger, J. M. 1992a. "Dumping and Antidumping: The Rhetoric and the Reality of Protection in Industrialized Countries." *The World Bank Research Observer* 7: 121–43.
- . 1992b. "The Meaning of 'Unfair' in United States Import Policy." *Minnesota Journal of Global Trade*: 1: 33–56.

- Gruenspecht, H. K. 1988. "Dumping and Dynamic Competition." *Journal of International Economics* 25: 225-48.
- Horlick, G. N. 1990. "The United States Antidumping System." In J. H. Jackson and E. A. Vermulst, eds., *Antidumping Law and Practice*.
- Jackson, J. H., and E. A. Vermulst. 1990. *Antidumping Law and Practice: A Comparative Study*. New York: Harvester-Wheat-sheaf.
- Japan Industrial Structure Council. 1994. *Unfair Trade Policies of Major Trading Partners*. Tokyo: Research Institute of International Trade and Industry.
- Messerlin, P. A., and Y. Noguchi. 1991. "The EC Antidumping and Anticircumvention Regulations: A Costly Yet Futile Exercise. The Case of the Photocopies."
- Ordoover, J. A., A. O. Sykes, and R. D. Willig. 1993. "Unfair International Trade Practices." In *Journal of International Law and Economics* 15: 323-37.
- Organization for Economic Cooperation and Development (OECD). 1993. *Antidumping and Competition Policy*. Paris: OECD.
- Prusa, T. J. 1992. "Why are so many antidumping petitions withdrawn?" *Journal of International Economics* 33: 1-20.
- Staiger, R. W., and F. A. Wolak. 1994a. "The Effects of Antidumping Law: Theory and Evidence." In A. Deardorff and R. Stern, eds., *Analytical and Negotiating Issues in the Global Trading System*. Ann Arbor: University of Michigan Press.
- . 1994b. *Measuring Industry-Specific Protection: Antidumping in the United States*. NBER Working Paper 4946. Cambridge, Mass.: National Bureau of Economic Research.
- U.S. Department of Justice. 1995. *Antitrust Enforcement Guidelines for International Operations*. Washington, D.C.
- Weinstein, D. E. 1992. "Competition and Unilateral Dumping." *Journal of International Economics* 32: 379-88.

# The basics of consumer protection: principles and policies

Eduardo Engel

Consumer policies are designed to protect consumers from physical or financial damage that may result from personal or household use of goods and services (Lane 1983). Their aim is to support households in their efforts to utilize their resources in an efficient manner. These policies influence the information available to consumers when they buy a good, the skills they possess to process this information, the likelihood that the product they buy results in physical damage, and the avenues open to obtain redress should they be dissatisfied with the purchase.

Those who stand to gain the most from consumer policies are the most vulnerable groups in society, such as the illiterate and the elderly. Not only do the members of such groups usually have less income to satisfy their material needs, they often lack the skills to determine how to spend their resources effectively.<sup>1</sup> In a country without consumer policies, the poor not only have the problems associated with low incomes, but also obtain less value for the money they spend.<sup>2</sup>

The main problems faced by consumers are excessive price and low quality. Excessive price may be due either to market power (a topic beyond the scope of this chapter) or to deceptive business practices, such as products that do not meet their advertised claims. The quality problem arises when attributes of goods and services turn out to be below the standards (explicitly or implicitly) announced by the seller and expected by the buyer, for example, safety and durability. Thus low-quality goods include a ladder whose faulty design puts the user at risk of physical harm, a toy that breaks when a child uses it as the instructions or common usage suggest, and a contractor that takes much longer than convened. Put differently, most problems faced by consumers fall under the heading “hidden quality.” Because of informational asymmetries, what consumers believe they are buying sometimes differs considerably from what they actually purchase.

Some argue that in addition to protection from hidden quality problems, consumers require protection from their own actions. This argument is offered to justify mandatory seatbelt laws, for example. Both kinds of protection differ at a basic level, since only the latter involves a paternalistic attitude toward consumers.

## Basis for consumer policies

Many view consumer policies as a means to promote consumer rights. The following consumer rights are widely accepted:<sup>3</sup>

- *The right to safety.* The right to be protected against the marketing of goods that are hazardous to health and life.
- *The right to be informed.* The right to be protected against fraudulent, deceitful, or grossly misleading information, advertising, labeling, or other practices, and to be given the facts needed to make an informed choice.
- *The right to choose.* The right to be assured, whenever possible, of access to a variety of products and services at competitive prices; and in those industries in which competition is not workable and government regulation is substituted, an assurance of satisfactory quality and service at a fair price.
- *The right to be heard.* The right to be assured that consumer interests will receive full and sympathetic consideration in the formulation of government policy and fair and expeditious treatment in its administrative courts.
- *The right to recourse and redress.* The right of access to proper redress—through swift, effective, and inexpensive procedures—for injury or damage resulting from the purchase or use of defective goods or unsatisfactory services.
- *The right to consumer education.* The right to gain the knowledge and skills needed in managing consumer resources and in taking actions to influence the factors that affect consumer decisions (Bannister and Monsma 1982).

Although formulating consumer rights is an effective way of focusing public attention on consumer issues, the

implementation of those rights cannot be based only on a statement of principles. Consider, for example, the right to safety.<sup>4</sup> There is no such thing as a totally safe product: Many products can cause physical, economic, or psychological harm. When a product becomes “hazardous to health and life” is difficult to judge. From the point of view of consumer protection, the relevant question is not whether products are safe, but whether market forces result in efficient levels of safety in consumer products.<sup>5</sup> A further complication is that because safer products are usually more expensive, requiring safer products can make it impossible for certain consumers—usually middle- and low-income consumers—to afford the cost of the good.<sup>6</sup>

Car safety regulations are a good example. In industrializing countries, auto bodies are thinner and minimum size requirements less stringent than for the same model in industrial countries. When specifying auto safety requirements, authorities face a trade-off between reducing the number of automobile fatalities and making cars available to a larger fraction of the population. Too-stringent requirements will harm middle- and low-income families that would have been able to afford a car had the safety regulations—and the costs of compliance—been less demanding.

An alternative to basing consumer policies on consumer rights is to adopt guidelines for consumer protection, as the United Nations did in 1985.<sup>7</sup> Even though some of the UN guidelines provide useful orientation, others are vague or even misleading. A case in point is guideline 25, which states that “where a standard lower than the generally accepted international standard is being applied because of local economic conditions, every effort should be made to raise that standard as soon as possible” (United Nations 1986, p. 4). The car safety example presented above calls into question the usefulness of international standards for most purposes. And if international standards do exist, it suggests that if they are based on those prevailing in industrial economies, they may be counterproductive.

On the other hand, simple ideas may benefit consumers considerably. For example, fraudulent weights are a problem faced by consumers in many developing countries:

In September 1977, the Consumers' Association weighed loaves of bread from Penang (Malaysia) bakeries. Each fell short of the government standard. . . . Nine brands of rice sold in “39-pound” bags also were caught short. And of 11 brands of soy

sauce, six bottles held half of what the labels claimed. . . . [The association] tested a dozen brands of ground coffee and found that each had less than the required 50% coffee contents. One had just 4.6% (Newman 1981 quoted in Klitgaard 1991).

When a consumer group in Bombay, India, opened a stand at one of the city's busiest markets, it was supplied not with sophisticated information, but with some pieces of basic equipment, including a correctly calibrated scale to detect fraudulent weights (Mayer 1989).

Another example of how information benefiting consumers can be generated at low cost is price surveys. In 1984, shortly after prices of most consumer goods were decontrolled in Zambia, the Prices and Income Commission began carrying out price surveys in Lusaka. Fifty-eight retail outlets were visited: the prices for goods of similar quality varied considerably. For example, the price per kilogram ranged from 0.65 to 2.00 kwachas (K) for onions, K4.00 to K6.60 for “ordinary mince,” and from K0.50 to K2.00 for tomatoes. Klitgaard (1991) provides the following description:

Commission chairman L. S. Chivuno appeared on television and radio and was interviewed in the press. The commission paid for a full-page advertisement in the Zambia Daily Mail that gave all the statistics and named particular stores at both ends of the price range. Commissioner Chivuno decried the 300 percent price differentials on some items. But his main point was simply that “if this kind of information is carried out at given intervals, the information conveyed will assist consumers in being better informed about where prices seem generally to be more attractive (p. 39).”

One of the benefits of requiring certain information to be made available to potential customers is that it helps consumers allocate their resources more efficiently. Credit purchases in developing countries offer a good illustration. Consumers in developing countries often buy durable goods on credit, and the retailer usually provides both the good and the credit. Some consumers have difficulty calculating the true cost of the credit because, first, various indirect costs are involved, and second, calculating a present value is not an easy task for most people. Regulations requiring that the cost of credit be summarized in one index, such as the “effective interest rate” or the present value, help consumers assess the true cost of

the goods they are considering buying and facilitates comparisons.

An understanding of the economics of information is thereby critical to the formulation of consumer policy. Consumers demand information on the attributes of the goods and services they consider buying. The suppliers of this information may be the sellers of the product, third parties such as product testing organizations, or the consumers themselves, who spend time and money trying out products and sharing this information with other consumers.

Viewing information in this light has important consequences for consumer protection. First, it makes clear the desirability of policies that reduce consumers' costs of obtaining information without increasing producer costs significantly. One such policy requires that sellers label their goods not only with the total price but also with the price by a standard unit ("unit pricing"), reducing the time and effort it takes to compare similar products packaged in different quantities.

Second, it raises the question, under what circumstances do the sellers of goods have incentives to voluntarily provide information to consumers that is both truthful and relevant?<sup>8</sup> Third, viewing information as a good also leads to a qualified appraisal of the "right to be informed." Promoting the right to be given the facts needed to make an informed choice ignores the costs involved in generating such information.

### **Market remedies or regulation?**

As with many other issues in economics, consumer policy divides policymakers into two camps. In one camp are those who advocate government regulation; in the other are those who are skeptical of the effectiveness of regulations and seek to rely on market solutions as much as possible. Both groups even refer to the field by different names: those favoring government regulation speak of "consumer policies" while those skeptical of direct government regulation prefer "consumer protection."<sup>9</sup>

The discussion in this chapter avoids differentiating between consumer policies and consumer protection. The view presented here is that consumer issues call for both policies based on the incentives provided by the market mechanism and, where this mechanism does not work appropriately, government regulation.

To illustrate the tension between the two views, consider the ideal world as seen through the eyes of a consumer activist, compared with that of someone skeptical of government intervention. In the consumer activist's "paradise" (adapted from Mayer 1989, p. 135), informa-

tion to make informed purchasing decisions would be easily available from both government-funded product-testing organizations and consumer advisory boards. Convenient neighborhood centers would provide information about any particular purchase and advice about how to file a complaint.

In this world, consumers continually receive valuable information and education through television. Consumer organizations have free access to prime-time television to discuss consumer issues and inform consumers how to avoid rip-offs. No advertising, including political advertising, is allowed on television. And cigarettes, although not banned, carry warnings that smoking kills. Moreover, the government takes measures to protect consumers not only from unscrupulous sellers, but also from themselves. A combination of public and private funds supports a team of trained safety experts who visit homes on request to search out potential safety hazards. The use of automobile safety belts is mandatory, and the government spends the resources required to enforce the belt law.

What would a consumer "paradise" look like from the point of view of a policymaker skeptical of any government intervention? In this world, sellers have a variety of incentives to provide truthful information to consumers. In the case of repeat purchases, it is to sellers' advantage to invest in reputation, since this investment increases their profits. Private product-testing organizations provide useful information about one-time purchases, for example, in widely read consumer magazines. Sellers are also deterred from deceiving consumers because consumers have access to speedy, inexpensive lawsuits. Lawsuits are brought only rarely. And private providers of safety seals not only guarantee, as far as possible, the safety of goods, but also ensure that any reparation for product-related damage is made quickly and at low cost.

In addition, goods are labeled with all relevant information, presented to facilitate understanding. Consumers read this information and make their purchasing decisions accordingly. The result is better-informed decisions without restricting consumer choice. This is desirable since what is dangerous for one consumer may be safe for another. As for advertising, consumers realize when an ad lacks information content and quickly perceive when they are being misled. Producers consequently have no incentive to use deceptive advertising techniques or unfair contract clauses.

Which version of the consumer "paradise" should developing countries strive for? This chapter shows how elements from both idealized worlds can be combined to protect consumers. The policies considered in this chapter

are aimed at modifying the environment faced by consumers, not the behavior of consumers. Policies designed to change consumer behavior are called “consumer promotion policies” and are considered in chapter 7 of this volume. This companion chapter considers such topics as consumer education, consumer redress, and the role of public consumer organizations. It also covers issues such as the political economy of consumer protection and the special case of consumer protection in economies in transition.

### Basic concepts

This section stresses how actual markets differ from idealized, perfectly competitive markets in ways that are central for consumer protection.

While the perfect competition paradigm is useful for other purposes, it is of limited value as an analytical tool for consumer policy. The perfect competition model assumes both rational consumers and costless transactions. A given good costs the same at different stores, and selling and buying do not consume resources. Under rather general conditions,<sup>10</sup> the equilibrium that results is Pareto-efficient, that is, no individual can be made better off without making someone else worse off. Furthermore, all Pareto-efficient resource allocations can be achieved through a competitive equilibrium.<sup>11</sup> Thus any shortcomings of a perfectly competitive economy necessarily relate to the distribution of income and should be rectified by lump-sum transfers.<sup>12</sup>

In a competitive equilibrium, the price consumers pay for a given good does not vary from store to store, and consumers do not spend time and resources informing themselves about prices and quality. Since consumers are omniscient and their preferences immutable, they do not invest resources in assessing the quality and other characteristics of goods they are planning to buy, and producers have no incentives to spend on advertising.

#### *Relevant sources of market failure*

Information on the price, quality, and other attributes of goods is often not easily available. Acquiring information on goods and services poses several problems for consumers: They must decide how much time and resources to spend in acquiring this information, they must process the information, and, in a world of uncertainty, they must make their purchasing decisions. Producers face similar problems. They would like to have information on consumer characteristics that is not readily available. Producers sometimes spend resources to acquire this information; at other times, they design products and contracts motivated by the informational shortcomings

they face. In both cases, there are inefficiencies that would not arise if information were freely available.

More generally, almost every economic interaction, both within an organization and among organizations, involves costs other than the price paid for the good or service provided. All such costs are referred to as “transactions costs.” They are the costs of running an economic system and underlie the sources of market failure relevant to consumer policy.<sup>13</sup> Transactions costs may arise from the need to determine prices and other details of the transaction so as to bring buyers and sellers together. The fee charged by a broker when an investor buys equity is an example of such a cost. The time consumers spend comparing prices at different stores—so called search cost—is another example.

An overwhelming proportion of transactions costs are due to “informational asymmetries.” Consumers and producers frequently do not have access to the same information. Producers often know more about the quality of the good they sell than do consumers; consumers sometimes have information that sellers would like to have. For example, selling a good under conditions in which payment is not collected at the time of purchase, as with credit sales, poses the problem for the seller of determining whether buyers will honor their payments. Sellers thus face a hidden quality problem—assessing the “quality” of the borrower. With perfect information, sellers could predict the future and charge a poor risk accordingly. Companies in the medical insurance industry face the same problem: Consumers know more about their health status than insurers do.

When there are informational asymmetries, a variety of phenomena may arise that are not captured by the perfect competition model. Prominent among them are moral hazard,<sup>14</sup> adverse selection,<sup>15</sup> and the principal-agent problem.<sup>16</sup> All these have in common that the market for a specific kind of information fails to develop and therefore may be viewed as resulting from information externalities.

“Externalities” occur when a producer or consumer affects a third party in a way not reflected by prices. A positive externality results when the action of one economic agent (consumer or producer) benefits another without being rewarded. Because there is no reward, the economic agent undertakes the action to a lesser degree (or less often) than is socially desirable. The situation is reversed in the case of a negative externality. Externalities may be viewed as a case in which high transactions costs result in the failure of a market (that for the externality) to exist. They become relevant when it is expensive to

exclude nonbuyers from the consumption of a good (or “bad”), either because this exclusion is technically impossible or because it requires considerable resources,<sup>17</sup> as when the market for an externality involves a small number of buyers and sellers.

To illustrate, consider a consumer who, by spending time and resources complaining about the defective design of a good, compels the manufacturer to improve it. The improved design benefits the complaining consumer as well as all future consumers of the good. Because the assertive consumer is not rewarded by those who benefit from his or her complaints, he or she does not internalize the effect his or her behavior has on the well-being of others. As a result, the “production” of consumer complaints leading to better products is underprovided.

Information is an example of a good whose production may involve positive externalities. A consumer acquires information up to the point at which the private cost of acquiring an additional bit of information equals the private cost of producing it. Since often many consumers could benefit from this additional bit of information at no extra cost, its social benefit exceeds its production costs; thus information is underprovided in a market economy.

In the absence of transactions costs and informational asymmetries, well-defined property rights lead to efficient resource allocation even in the presence of externalities (Coase 1960). Externalities pose a public policy problem either when transactions costs are large or when property rights are not well defined. When transactions costs are significant, the main remedy for the underprovision of positive externalities is subsidizing the production of the good.<sup>18</sup> However, providing these subsidies involves operational and informational costs that should be compared to the expected benefit before such subsidies are implemented.

An extreme case of a positive externality is a public good. A good is public if it is nonrival and nonexclusive.<sup>19</sup> A good is nonrival if once it has been produced, it can be provided to additional consumers at no additional cost. For this reason, it is not desirable to ration such a good. A good is nonexclusive if people cannot be excluded from consuming it, that is, people cannot be prevented from enjoying the good without direct payment. Thus a public good is a nonexclusive good that provides a positive externality to a large number of consumers. It is neither feasible nor desirable to ration its use.

Another instance of market failure is the presence of market power. Firms with market power charge a price that is above their marginal costs, usually resulting in abnormal profits. Ideally, firms would like to charge every customer

the highest price he or she is prepared to pay for a good; this price is the customer’s reservation price. When producers charge different prices to different consumers for essentially the same good, they are said to be price-discriminating. That sellers are often prevented by legal or informational constraints from price-discriminating perfectly among consumers implies that most consumers pay less than their reservation price for the goods they buy.<sup>20</sup> The difference between the price a consumer pays and his or her reservation price is called the consumer’s surplus.

When the seller and the potential customer determine a good’s price by haggling, they are effectively bargaining about how they will split the total surplus generated by their transaction, namely, the sum of the consumer’s surplus and the seller’s profit (also called the “producer’s surplus”). The outcome is inefficient from a social point of view when an agreement is not reached even though there is overlap between the prices at which both the consumer and producer would attain a surplus. Although this simple framework for viewing a transaction between a buyer and a seller omits a number of relevant issues (for example, it takes the market structure as given), it will prove useful later in this chapter in the analysis of remedies for consumer protection.

#### *Irrationality and misperceptions*

In a perfectly competitive world, the price consumers pay and the quality of the goods they buy are unrelated to how rational they are, since all consumers pay the same price for a given good, and goods are homogeneous. In reality, however, consumers often pay different prices for identical goods. The amount and quality of information available, and many other factors, make the consumer’s decisionmaking task in a market economy formidable.<sup>21</sup> The assumptions that are made about how consumers make decisions play an important role in the analysis of consumer policy.

Many problems faced by consumers involve either important degrees of uncertainty or the need to assess risk correctly. There is ample evidence that human beings have a hard time evaluating risk. When making decisions under uncertainty, people systematically depart from what common sense would consider rational behavior. Some well-documented examples of consumer behavior provide additional evidence of “irrational” behavior—even though a formal theory incorporating them is lacking. These phenomena have important implications for consumer policy.

*Information processing by consumers.* Consumers use information only when the benefit from doing so exceeds

the costs, including the time to gather and process information.<sup>22</sup> Policies would be judged as desirable from a social point of view if the reduction in cost or increased benefits to which they lead exceed the costs of implementation and enforcement.<sup>23</sup>

Many consumer policies are designed to reduce some of the following “costs,” or barriers that prevent consumers from using information efficiently:

- *Knowledge.* Consumers often do not understand available information about a good or service. For example, laundry detergents often claim to have some sophisticated component (such as biosolves), whose meaning and effect are unknown to most consumers. This problem is sometimes exacerbated by producers who try to differentiate their product by adding attributes that serve no purpose other than product differentiation.
- *Effort.* It may take consumers considerable time to discover where to find information about a particular product. For example, consumers could spend a full day going from one supermarket to the next to compare the cost of the basket of goods that they buy regularly; the time required to do so, however, prevents most consumers from undertaking this task.
- *Environment.* The informational environment faced by consumers is often unfriendly. How information is presented may determine whether consumers use it. For example, unit pricing helps consumers compare prices across products packaged in different sizes or quantities.
- *Irrelevant or biased information.* Available information about a product is often not the information consumers are interested in. Or the information may be presented in a misleading way.

*Assessing risk and making decisions under uncertainty.* Consumers often make decisions in the face of significant uncertainty. The expected utility hypothesis is the central assumption in economics about how individuals make “rational” choices under uncertainty. Consider an individual faced with choosing among alternative actions whose impact on her welfare depends on events unknown to her. According to the expected utility hypothesis, she would consider every possible action she could take and would assign cardinal utility to her welfare for all possible outcomes of the uncertain events.<sup>24</sup> She then would calculate the expected utility of every action by appropriately weighting the utilities she assigned by the corresponding probabilities. Finally, she would choose the action that maximized her expected utility.

The concepts of risk aversion and risk premium fit naturally into the expected utility framework. Most people are prepared to pay money to reduce the level of risk to

which they are exposed. For example, most families prefer a lower but secure income to an uncertain, albeit higher on average, income. Individuals with such preferences are said to be risk-averse, and the income they are prepared to give up (on average) to ensure a steady flow of income is the “risk premium” they pay. People who care only about their average income, and not about how uncertain it is, are said to be risk-neutral. They are not prepared to pay a risk premium to ensure a certain income. A risk-averse person prefers a guaranteed \$100 to equal odds on gaining \$200 or nothing. A risk-neutral person is indifferent between both alternatives, since in each case the average return is \$100. Markets for insurance exist largely because most people are risk-averse.

The expected utility framework raises various issues. First, it is clear that few individuals actually assign cardinal utilities to the possible scenarios and then calculate their expected values. There are two complementary answers to this objection. On the one hand, the expected utility hypothesis can be viewed as a working assumption from which empirically testable implications can be deduced (Laffont 1989). Alternatively, as Savage (1954) showed, if individual choice under uncertainty satisfies certain basic properties (axioms), people act as if they maximized their expected utility.<sup>25</sup>

A second objection to the expected utility hypothesis is that individuals must be able to assess correctly the probabilities of uncertain events to calculate expectations. For policy questions, it is useful to distinguish between cases in which there is a reasonable degree of consensus on these probabilities and those in which lack of information on similar events implies that probability assessments are largely subjective (that is, they may vary substantially from one individual to another).<sup>26</sup> The probability of dying of lung cancer if you smoke a pack of cigarettes a day is rather close to being an objective probability. The probability that the Russian economy will be growing fast by the beginning of the twenty-first century is a subjective probability. When probabilities are objective, individuals may make systematic mistakes in their assessments. For example, most people have a difficult time assessing low probabilities: It is hard to differentiate a risk of 1 in 100,000 from a risk of 1 in 10 million, even though the first is 100 times more likely to occur. If the costs associated with both risks are large, this difficulty may lead to important misallocations of resources in risk reduction.<sup>27</sup>

Even if probabilities are subjective, the mathematical rules for calculating complex events based on probabilities of simpler events impose constraints on how a ratio-

nal individual assigns probabilities. For example, if two events never happen simultaneously, then the probability of either one of the two events taking place must be the sum of their individual probabilities (no matter what probabilities are assigned to the individual events). If consumer behavior indicates that consumers are violating the basic laws of probability in making their decisions, then consumers are acting in an irrational manner.

There is a rich literature showing that people make systematic mistakes when making decisions under uncertainty, that these mistakes are made in simple situations, and that they are made by both laypeople and experts.<sup>28</sup> As described above, these mistakes may arise because (a) people do not act as expected utility maximizers; (b) people assess (objective) probabilities incorrectly; and (c) people make systematic mistakes when applying the laws of probability.

Three of the biases that are documented by this literature and are most relevant for this (and the following) chapter are briefly reviewed here.<sup>29</sup>

- *Prominence or salience.* People may either over- or underestimate the probability of an event occurring depending on the event's characteristics. People generally overestimate the probability of dramatic, dreadful, prominent events (such as airplane crashes) and underestimate the probability of regular, less dramatic events. Breyer (1993, table 4) illustrated this point. He compared how the U.S. public and experts at the Environmental Protection Agency rated the importance of 22 health risks associated with environmental problems. The public's ratings were totally unrelated to the experts' assessments.<sup>30</sup>

"Salience" may lead consumers to weigh available information incorrectly when deciding whether to purchase a good. The following example (Nisbett and Ross 1980 cited in Akerlof 1991) illustrates this point:

Let us suppose that you wish to buy a new car and have decided that on grounds of economy and longevity you want to purchase one of those stalwart, middle-class Swedish cars—either a Volvo or a Saab. As a prudent and sensible buyer, you go to Consumer Reports, which informs you that the consensus of their experts is that the Volvo is mechanically superior, and the consensus of the readership is that the Volvo has the better repair record. Armed with this information, you decide to go and strike a bargain with the Volvo dealer before the week is out. In the interim, however, you go to a cocktail party where you announce your intention to an acquaint-

tance. He reacts with disbelief and alarm; "A Volvo! You've got to be kidding. My brother-in-law had a Volvo. First, the fancy fuel injection computer thing went out. Two-hundred and fifty bucks. Next he started having trouble with the rear end. Had to replace it. Then the transmission and the clutch. Finally sold it in three years for junk (p. 2)."

This anecdote adds only one case experience to those considered by *Consumer Reports*, leaving the mean repair records of the two cars virtually unchanged. Yet most prospective car buyers are likely to give considerably more weight to the case described in the above scenario than is warranted by the information it actually contributes.

- *Rules of thumb.* People often use rules of thumb (heuristics) when making decisions under uncertainty. This approach reduces the time and effort to make a decision and may be justified due to the cost and effort involved in processing information, as long as the biases introduced are small.<sup>31</sup> Yet there is substantial evidence that rules of thumb used in practice are based on principles (such as anchoring, representativeness, and availability of instances) that may lead to large and significant biases.

- *The belief in personal immunity.* There is evidence suggesting that most people view themselves as exposed to less risk than the average person. When it comes to the risk of lung cancer from smoking, for example, many smokers rationalize that "it can't happen to me." Needless to say, this implies that most people systematically underestimate their risk levels. We somehow tend to believe that negative events happen to others, not to us. This misperception helps explain why most people do not use safety belts in the absence of a belt law.<sup>32</sup> It also helps explain why interest rates charged by credit card companies in the United States remained almost unchanged during the second half of the 1980s and the early 1990s, even though interest rates charged by banks decreased dramatically.<sup>33</sup> When choosing a credit card, consumers underestimate the probability that they will have to run high levels of debt on it; by the time debt has accumulated, no other credit card company will lend them money.

*Additional examples of irrational behavior.* This section concludes by considering some additional evidence on consumer "irrationality." These examples are relevant to the discussion of the relative merits of alternative consumer protection policies that follows later in this chapter.

- Pratt, Wise, and Zeckhauser (1979) showed that price dispersion for "almost" identical goods (in the city of Boston) was far larger than could be accounted for by

transaction (search) costs or alternative economic explanations. All products considered were listed in the Yellow Pages, so consumers' lack of access to information was not an issue. A survey conducted by the Chilean Consumer Service (SERNAC) in 1992 obtained similar results. Nearby pharmacies charged prices for identical, relatively expensive drugs that often differed by a factor of three. Although consumers could have "shopped around" for drugs, they apparently did not.

- Day and Brandt (1974) studied the effect of the U.S. Truth-in-Lending Act, which forces retail stores to inform customers of the interest rate implicit in sales that are paid in installments. They concluded that a large fraction of consumers do not change their behavior based on this information and consequently pay a much higher interest rate than necessary.

#### *A classification of consumer goods and services*

The ease with which consumers can assess the attributes of a good provides a useful framework for analysis (see Nelson 1970 and Darby and Karni 1973). Goods can be classified in one of three groups:

- *Search goods.* These are goods whose quality can be ascertained before purchase. Stamps, postcards, and dresses are examples of search goods.
- *Experience goods.* The quality of these goods is learned only after their purchase, through use. A book, canned food, restaurants, and suitcases are examples of experience goods.
- *Credence goods.* Consumers rarely learn the quality of these goods. Fire extinguishers, the resistance of a house to an earthquake, and the timeliness of a doctor's intervention belong to this category.

In the case of search goods, consumers can allocate resources efficiently under reasonable assumptions about their rationality and informational environment; they must merely examine goods carefully before buying them. To what extent does the market mechanism ensure that the quality and variety of search goods are close to what is socially desirable? When firms have market power, there is no such assurance.<sup>34</sup> Both the quality and the variety of goods produced may be above or below their socially desirable levels.<sup>35</sup> Policy redressal would be based on antitrust laws and regulation, topics beyond the scope of this chapter.

In the case of experience goods, sellers often have considerably more information than buyers. The main consumer protection issue is whether adequate information is made available to consumers at a low cost. Deciding whether regulation is necessary requires an

understanding of the incentives that firms and other private agents have to supply information to consumers. This topic plays a central role in the next section on market-based remedies.

Credence goods generally require government intervention. The market test is usually not strong enough to deter producers from opportunistic behavior, since the worst threat a producer faces is often bankruptcy, although the potential damage of such behavior may be considerably larger.<sup>36</sup> Government intervention may take a variety of regulatory remedies.<sup>37</sup>

#### **Market-based remedies**

The main consumer policy issue in the case of experience goods is whether producers have incentives to provide quality products and, if they do, how they convey this fact to consumers. This section considers several market-based remedies that help protect buyers of experience goods.

#### *Guarantees provided by sellers*

Some producers fully compensate consumers if the quality of an experience good differs from that publicized.<sup>38</sup> Producers have incentives to provide full warranties when (a) the quality of the experience good is easy to evaluate, and (b) the good's performance is not affected by the consumer's behavior and can therefore be attributed entirely to the producer. Under these circumstances producers will provide full warranties because consumers will grow suspicious if they do not. Limited warranties are signals of low quality in this case.

The performance of most goods depends on how buyers use them. For this reason, most goods have only a limited warranty or no warranty at all. Providing a full warranty for such goods would lead to a moral hazard problem: Since buyers have no incentives to internalize the cost of using the product carelessly, they will be more careless than they would otherwise be. Furthermore, firms offering full warranties would attract "high-risk" consumers, thus also leading to an adverse selection problem. By requiring consumers to share part of the costs of performance that is below the promised level, producers induce consumers to behave more carefully.<sup>39</sup>

Government regulation that forces producers to provide full (or partial) warranties for experience goods may do more harm than good by leading to moral hazard and adverse selection problems.<sup>40</sup> Yet governments may foster consumers' interests by measures that reduce their information processing costs in evaluating warranties. For example, government regulations could specify a mini-

imum standard that manufacturers must meet to use the term “full warranty”; warranties that fell short of these requirements would have to state that they were limited.<sup>41</sup>

One could argue that such a law might reduce the total provision of warranties by producers, who might not want to signal that they were not prepared to stand fully behind their products. However, in the United States the duration, scope, and remedies in warranties improved after passage of the Magnuson-Moss Warranty Act of 1975.<sup>42</sup> One possible explanation is that manufacturers benefit from standardization, since it reduces the cost of signaling to consumers that the producer believes the good it manufactures is of high quality.

The potential benefit that may accrue to consumers from standardization is illustrated by a proposal that Chilean authorities are considering for car sales. It would require sellers to list in a standard format the duration and kind of warranty that applies to various car parts are under—a simple measure that would make it easier for consumers to compare different cars.

Warranties may also serve as quality guarantees and assurances, rather than insurance policies. The trend in retailing in industrial countries is for retailers to take back products and offer customers a refund, with no questions asked. Some products even have a double-the-money-back guarantee if the buyer is at all dissatisfied. Such practices can be viewed as extensions of the principle, “the customer is always right.”

Two issues related to warranties are particularly relevant in many developing countries. First, warranties are useless in informal markets, where sellers cannot be held accountable for defective goods. This explains why search goods are considerably more likely than experience goods to be sold in informal markets. It also explains why informal sellers find ways of turning experience goods into search goods. For example, potential buyers of watermelon in informal markets in developing countries often are allowed to taste a thin slice.<sup>43</sup> In this way street vendors assure consumers of the quality of their produce and, for all practical purposes, watermelons become search goods.

One could argue that, in the case of experience goods sold in informal markets, manufacturers could provide warranties directly, thereby solving the problem. Yet precisely because of the reputation of brand names, informal retailers often forge brand name labels, which forces manufacturers to guarantee only goods bought at formal outlets.<sup>44</sup>

A second issue, especially in countries where illiteracy rates are high, is that consumers may have difficulty understanding warranties. This provides an additional

argument in favor of a simple standardization, such as that described above, for the term “full warranty.” Both examples point to the importance of consumer education, a topic considered in chapter 7.

#### *Information provided by sellers*

Warranties for experience goods may offer little protection for consumers when enforcement costs are high relative to the value of the good. Such costs may be high if, for example, the legal system is ineffective or if it is difficult to assess whether the good met the promised performance standards. In the second case, adequate compensation may pose a problem. In the absence of warranties, the issue is whether manufacturers have incentives to provide adequate information about product characteristics and prices to consumers.

*Information about quality: signals, bonds, and reputation.* Information about quality is communicated very differently for one-time versus repeat purchases. When the buyer and seller interact only once and the seller cannot be sued for faulty quality (or transactions costs are prohibitively high), market demand is unable to discriminate among different qualities offered. The producer therefore has no reason not to provide the lowest possible quality.

Producers of high-quality goods may invest in reputation to signal consumers that their products are high caliber. The rationale behind such expenditures is that the producer, knowing that the product is of high quality, is prepared to sustain initial losses. Were the product of poor quality, consumers would eventually discover the truth, and the producer would never recover the initial expenditure. The advertising campaign may convey the producer’s desired message even if it gives no real information about the product. The fact that the advertising is launched shows that the product is worth promoting, because advertising is not advantageous for low-quality experience goods.

Posting a bond that guarantees the performance of a good or service is another way a producer can build reputation and signal the quality of its product to consumers. The producer forfeits the bond if the service is defective. For example, contractors often must post a bond stipulating that a project will be completed by an agreed date and in an agreed manner.

Although bonds often give adequate incentives for sellers to provide quality products, they present a number of limitations, some particularly relevant in developing countries. First, sellers may not have the financial resources to post a bond.<sup>45</sup> Second, in the case of experience goods, the price of a bond that provides adequate

incentives for the producer is inversely proportional to the probability that consumers will detect product defects; thus large sums are required when the probability of detecting a defective good is small. Third, the transactions costs involved in determining whether the good was provided as agreed may be high, especially if the legal system is not efficient.

In the absence of warranties and bonds, high-quality goods are likely to be produced in the case of repeat purchases. When purchases are made fairly frequently and customers can quickly learn the quality of the good, manufacturers have incentives to continue producing high-quality products. By so doing they build a positive reputation among consumers and eventually may charge a "quality premium." Fear of lost sales and consumer complaints deters manufacturers from lowering quality.

To illustrate the difference between one-time and repeat purchases, consider the case of life insurance. *The Economist* concluded its gloomy appraisal of the British life insurance industry as follows: "The industry remains riddled with bad practice, and will continue to be as long as most sales take place after brief encounters between ill-informed customers and unfamiliar salesmen chasing hefty commissions."<sup>46</sup> The solution offered by the British magazine was bancassurance. Customers buy this insurance from a bank that has formed an association with an insurance company. The insurance shifts to the insurance company the incentives provided by the long relationship between the customer and the bank, thereby solving the moral hazard problem.

In Chile every worker must save 10 percent of his or her earnings in a savings account that is managed by a private pension fund. On retirement, workers can choose between a sequence of phased withdrawals and an annuity. The resulting annuities market presents problems similar to those of the British life insurance market. In the case of annuities, transactions take place once in a lifetime, commissions are hefty, and advertising expenditures large.<sup>47</sup> Moreover, retirees are charged a much higher price than that for similar assets sold under repeat-purchase conditions (Bitrán 1994). There also is a wide variety of annuities, which consumers have difficulty comprehending.

Bancassurance is not a viable solution in Chile because a large fraction of the population does not have a bank account. The Chilean Congress is considering whether to standardize the annuities market in order to make it easier for customers to make comparisons (Bitrán 1994).<sup>48</sup> The risks of the various annuity providers would be rated, and retirees would choose at least three firms to bid for their annuity. The bidding process would automatically

include all annuity providers whose ratings are equal to or greater than the average rating of those chosen by the customer. Annuity sellers thus would have an incentive to follow a strategy of high capitalization and low prices, without incurring large advertising expenditures. Entry barriers would also be dramatically lowered in this system. Once bids were in, the buyer would be free to choose either the best offer among all bids (within a 1 percent range) or the best offer among those annuity providers selected initially.

Although this discussion has emphasized the distinction between one-time and repeat purchases, it is not strictly necessary for consumers to purchase a product frequently for reputation to take on importance. Consumers often have access to some information before buying a good. They may ask friends and relatives who have bought the good, they may rely on the advice of a vendor whom they trust, they may trust certain brand names, they may read product-testing magazines, or they may perform some simple tests. In any of these cases, the quality of experience goods increases with the number of informed consumers.<sup>49</sup> Furthermore, as long as some customers are informed, high prices may signal high quality to uninformed customers.

*Information about price: sales, bargains, and rip-offs.* There is widespread evidence that consumers pay substantially different prices for similar (even identical) goods and that price dispersion is growing over time. In the United States the dollar value of markdowns on all merchandise sold in department stores as a percentage of dollar sales increased from 8.9 percent in 1975 to 16.1 percent in 1984 (Pashigian 1988).<sup>50</sup>

To determine whether price dispersion (spatial or temporal) justifies a specific consumer policy response, it is necessary to understand why people pay different prices for similar goods. Much price dispersion is probably the result of second-degree price discrimination. Since customers' reservation prices are not known to sellers, and sellers are often unable to charge different prices to different groups of customers (due to informational and legal constraints), sellers may charge a high price initially (so as to capture customers prepared to pay more), then lower the price, and eventually put the good on sale.<sup>51</sup> Since most customers prefer to consume a good sooner than later, some will buy at the higher price. Alternatively, random sales benefit most of those customers who spend time looking for bargains; these customers are usually those with lower reservation prices.

Price dispersion may also be related to positive search costs. Information about prices at all stores selling a par-

ticular good is usually not available. Consumers must spend time (and other resources) to gain this information. A market equilibrium with price dispersion may emerge if consumers' costs of obtaining information differ substantially (Salop and Stiglitz 1977). Although those selling at low prices will have incentives to provide price information to consumers up to the point where the marginal expected benefit from so doing is equal to the corresponding marginal cost, from the point of view of consumer protection, the issue is whether this leads to sufficient price information. In fact, the amount of information is often inadequate. At the same time, there is an externality associated with acquiring information.<sup>52</sup>

The problem often faced by consumers with regard to price dispersion is ensuring that once a cheap outlet has been found, the good has the announced quality. When the price of a good is much lower than a consumer would normally expect to pay, there is a high probability that the product is defective.<sup>53</sup> Even though legislation usually requires that information on substandard articles be provided, it may be difficult to enforce such laws (for example, for goods sold at going-out-of-business sales). Consumers are generally suspicious of sales in which goods are sold at exceedingly low prices (see Leff 1976). For that reason some sellers try to convince buyers that there is a valid reason for the sale—which is why fire sales, going-out-of-business sales, and the like attract more customers than an announcement that a group of substandard products has just arrived.

#### *Information provided by third parties*

Even if sellers and buyers interact repeatedly, buyers often have good reason for not trusting the information sellers provide. When an assessment of the quality of a good involves substantial subjective judgment, it may be to the advantage of both sellers and buyers if third parties provide the information.<sup>54</sup>

Consider, for example, the case of investors wishing to assess the quality of a security. Most investors face a formidable task if forced to assess the probability of default on a bond. They are therefore willing to pay for a reliable assessment of the bond's likely performance. Many investors rely on the ratings of private firms, such as Moody's and Standard and Poor's.<sup>55</sup>

Another example of third-party information is quality certification systems, such as the ISO-9000 series.<sup>56</sup> ISO certificates provide easy checks on the quality of suppliers of intermediate goods, thereby helping manufacturers select suppliers (especially foreign suppliers).<sup>57</sup> In industrializing countries, manufacturers in the export sector

were the first firms to demand ISO certification. A manufacturer that obtains ISO certification often demands that its local suppliers do the same. The number of manufacturers with ISO certification in developing countries is likely to grow dramatically in coming years.

Because many products have a complicated construction, making it difficult for consumers to evaluate their safety, a market has been created for private providers of safety seals, such as the Underwriters Laboratories (UL) seal in the United States and the Safety Goods (SG) seal in Japan.<sup>58</sup> Such institutions sometimes receive government funding; the German Institute for Industrial Standards (DIN), which issues the "DIN-tested" seal, is one example.<sup>59</sup>

A major difference between quality control certification (such as ISO) and safety seals (such as UL, SG, and DIN) is that only in the second case does the certifier bear responsibility for injuries that may occur.<sup>60</sup> Whether such seals will become more important in developing countries is an open question. Early experiences have not been very successful. For example, poorly funded government organizations in India were unable to prevent forgery of both local and foreign seals.<sup>61</sup> Private safety seal providers (both local and foreign) will no doubt emerge in developing countries once their legal systems can handle effectively the liability issues involved.

Whether private institutions have incentives to provide information depends on the extent to which they can derive the full benefits of this information. In the case of safety seals, firms pay for certification directly, thereby avoiding any free-rider problem. In the case of consumer information magazines, however, buyers may benefit from these magazines without paying for them, either by reading them in a library or by borrowing them from a subscriber.<sup>62</sup> Thus consumer magazines will provide less product information than is socially desirable.

Product-testing organizations are an important feature of consumer protection in industrial countries. The magazines these organizations publish provide consumers with useful information on product features and quality. Some of these private organizations are making special efforts to foster consumer protection in developing countries. For example, the International Organization of Consumer Unions (IOCU), which has offices in Chile, London, Malaysia, and Zimbabwe, has promoted the growth of product-testing organizations in developing countries. Clearly, there are important economies of scale in sharing testing methods and results across countries. This raises the issue of how transnational consumer organizations (such as the IOCU), which generate external benefits

beyond the border of any particular country, should be funded. Multilateral organizations may have an important role to play in preventing countries from free-riding.

An important issue related to the private provision of information is the incentives that providers have to be truthful. The concepts developed earlier in this section also apply to private information providers; they sell a good that could be considered an experience good (and often a credence good).<sup>63</sup> Private providers of information must invest in building a reputation, which takes time and resources. This is relevant, for example, for new consumer magazines. A careful choice of private testing institutions and cooperation with institutions that have gained international prestige may help new consumer magazines build a reputation among both consumers and producers.

It is easy to make the case for subsidizing private organizations that provide information to consumers.<sup>64</sup> Argentina's consumer protection legislation passed in October 1993, for example, provides state subsidies for private consumer organizations.<sup>65</sup> Such funding should depend on the interest that consumers demonstrate in the information provided by such an institution. Thus the government funding received by a consumer magazine should increase with its circulation.<sup>66</sup>

Subsidies for organizations that provide information to consumers help the organizations internalize the benefits they provide. By increasing the number of informed consumers, subsidies also benefit uninformed consumers (a positive externality), since informed consumers induce (or allow) firms with market power to produce high-quality goods. Thorelli, Becker, and Engledow (1975) referred to informed consumers as "information seekers." They noted that "they, rather than the average consumer, are keeping producers on their toes. They, more than others, fight the battle for better products, for honesty and decency in business practice, and for more truthful and informative advertising."

### **Regulatory remedies for consumer protection**

Governments are not benevolent social planners striving to maximize society's welfare.<sup>67</sup> Government policies may lead to corruption, rent seeking, and waste. They also may create groups of powerful and privileged bureaucrats. Regulatory remedies for consumer protection must take these limitations into account.

#### *Protection of health and safety*

Some producers sell products that put consumers at unreasonable risk. In some cases, producers fail to take adequate steps to prevent potential hazards. The Chevrolet Corvair

case, which catapulted Ralph Nader to the leadership of the U.S. consumer movement in the mid-1960s, and the thalidomide cases in Europe are well-known examples of manufacturer negligence. Frequent food poisoning episodes in developing countries (and sometimes in industrial countries, such as the Kanemi rice oil and the Morinaga powdered milk cases in Japan) are additional examples.

As noted earlier, there is no such thing as a totally safe product. Most products can cause physical, economic, or psychological harm. From the point of view of consumer protection, the relevant question is not whether products are safe, but whether market forces provide incentives for producers to ensure "efficient" levels of safety in consumer products. Markets do not provide efficient safety levels when there is a serious source of market failure (usually rooted in some significant transaction cost). Examples are hazards that are imposed on others (negative externalities), such as environmental risks, and product risks unknown to the consumers who bear them (due to asymmetric information). Another reason government action may be justified is consumers' misperceptions of low-probability events.<sup>68</sup>

In the absence of voluntary actions by producers, there are two approaches for protecting consumers' health and safety when markets fail to do so efficiently: liability law and regulation. Liability law and government regulation are different both in their design and in the way they promote consumer safety. Liability law compensates those injured, whereas government regulation is best designed to provide incentives for appropriate prevention. Regulatory requirements are superior to liability law when producers are unable to pay fully for the harm they cause (Shavell 1984). This situation is more pervasive in developing countries, where local producers often are not accountable (such as those selling in informal markets) or would go bankrupt if sued successfully when transactions costs are so high that producers are seldom sued, regulation is more effective than liability law (Shavell 1984). Conversely, when the authority's information on risk is poor, liability law can be expected to work better than government regulation.<sup>69</sup>

This section argues that, even though developing countries would benefit from reducing the transactions costs involved in using liability law to promote consumers' interests, they should learn from the mistakes of some industrial economies and avoid excessive reliance on liability law. A well-designed regulation or appropriate incentives for private provision of safety often can be considerably more successful than liability law as well as much less costly and more efficient.<sup>70</sup>

*Liability law.* It is impossible to write contracts that consider all possible contingencies, given the transactions costs that would be incurred. It is economically more efficient to settle the costs associated with risks with a low probability of occurrence only after adverse events actually take place.

In many developing countries, consumers are still the main bearers of risk (*caveat emptor*), yet as consumer protection develops, these risks are borne increasingly by producers (*caveat vendor*). For example, the consumer protection law approved in Brazil in 1990 places the burden of proof on producers (or importers).<sup>71</sup>

Consideration of transactions costs (such as legal fees and liability insurance) for consumers and producers is important in evaluating the efficiency of a consumer protection system based on liability law.<sup>72</sup> Data from countries that rely heavily on liability law to encourage product safety indicate that these costs can be astronomical. For example, liability costs account for 17 percent of the fares paid by passengers on the Philadelphia mass transit system and approximately 20 percent of the retail price of a ladder in the United States (Viscusi 1991a, p. 8). Net compensation to U.S. victims constitutes only about half of the total expenditures for tort litigation.

In most countries, liability is imposed according to two basic standards: strict liability and negligence rule. Under strict liability, producers are responsible for the damage their products cause regardless of how much care they took in product design and manufacture; the new law in Brazil belongs to this class. Such a liability law forces producers (or importers) to internalize all safety costs and may lead to an inefficient outcome when consumer behavior plays an important role in product-related accidents. The moral hazard problem under strict liability is serious and may cause producers to spend too many resources to provide safe products.

Negligence rule imposes some reasonable level of care on producers and requires defining the standards that correspond to "reasonable care." Producers are held to standards regarding product design and manufacture as well as warnings that their products carry.

Breyer (1993) summarized most experts' current thinking on the U. S. liability system:

[The tort system] leaves the determination of "too much risk" in the hands of tens of thousands of different juries who are forced to answer the question not in terms of a statistical life, but in reference to a very real victim, needing compensation in the courtroom before them. The result is a system much

criticized for its random, lottery-like results and its high "transactions costs" (i.e., legal fees) which eat up a large fraction of compensation awards (p. 59).

Viscusi (1991a) illustrated the excesses of the U.S. liability system with some vivid examples. In one case a physician who fell off his horse at a country club and fractured his right arm, sued the club and was awarded \$6.3 million in damages in an out-of-court settlement. In another case a Philadelphia psychic who claimed she had lost her psychic powers after undergoing a scanner exam was awarded \$1 million.

In developing countries whereas consumers' access to redress from private providers of goods and services is often limited and expensive, obtaining redress from public utilities, municipalities, and other government agencies is usually impossible. Describing the situation in India in the early 1980s, Galanter (1985) noted that "disasters large and small in India typically have no legal consequences." Redress from government can be expected to develop in parallel with civil servants' accountability. Major legal (and sometimes even constitutional) reforms are needed in many developing countries. India's 1986 Consumer Protection Act gives consumers access to redress when government goods and services are involved.

Protecting consumers from medical malpractice poses a major challenge for those designing consumer protection policies in developing countries. The asymmetry of information in this case is extreme: Patients have almost no way of assessing to what an extent a negative outcome is due to physician negligence. A difficult agency problem thus exists. Resolving such issues requires assessments provided by other physicians, a task that is often impracticable.

Protecting consumers from negligent physicians can be an expensive task, as the medical malpractice situation in the United States illustrates. When physicians fear the threat of medical malpractice charges, they tend to order more tests than they might otherwise (defensive medicine). Consumers have few incentives to resist such tests since they do not bear medical costs at the margin because medical bills are covered by either government-financed public health services or private insurers. The high cost of health care in the United States can be partly attributed to the medical malpractice liability system.

Successful medical malpractice cases are rare in most developing countries. For example, in Chile it took a case involving a supreme court judge for the legal system to award more than token punitive damages. The judge had

a hip problem, and the physician charged with negligence had performed surgery on the wrong hip. This case also offers an illustration of how the medical profession can close ranks behind a threatened colleague.<sup>73</sup>

Most industrializing countries would benefit from the positive incentive effect on product and service safety that would result from significant expansion of liability. Yet beyond a certain threshold, the liability system stops acting as a deterrent and ends up increasing the prices of goods and services with no positive offsetting effect.<sup>74</sup>

*Regulation and consumer safety.* Protecting consumer safety through the liability system is costly, since it requires use of the normally expensive legal apparatus. Direct regulation may be a more effective, less costly way to deter accidents. The effectiveness of government regulation is limited, however, because practical constraints preclude focusing regulatory attention on more than a small percentage of products. This contrasts with liability law which, at least in principle, is all inclusive.

Government regulation may be quite powerful in some markets. It can take the form of quality control, minimum quality standards, obligatory disclosure, occupational licensing and certification (as for doctors and lawyers), safety regulations, recalls, and bans. Government regulations focus on prevention rather than compensation for victims, as does the liability apparatus. Although regulation does provide incentives for producers (such as through fines), there is no direct link between these incentives and reparations to those injured.

Regulators try to limit or reduce exposure to certain potentially risky substances, products, and even people (for example, unqualified doctors). There are some products whose potential risk is considerably larger than their benefits, justifying a ban on their production.<sup>75</sup> The regulatory system should seek to foster choices that informed consumers would make for themselves in a well-functioning market.<sup>76</sup> The regulatory process should be based on risk-benefit analysis.<sup>77</sup>

Government standards and regulations can be classified as routine and nonroutine. The case of routine product safety regulations is straightforward: manufacturers of products in certain categories must follow standard procedures to obtain government certification before the products are sold to the public. In the nonroutine case, once a new risk makes it to the public policymaking agenda, a risk assessment must be performed to determine the extent of the risk.<sup>78</sup> Risk management then involves answering a number of questions. Should the product be banned, regulated, or modified? Should perceptions and valuations be altered through education and

public relations? How much should be spent? Who should pay?

Risk management can benefit from the expertise of civil servants who specialize in safety regulations and work together with experts from a variety of fields on a regular basis. In the case of developing countries, free-riding on the expertise of industrial economies may lower costs substantially. It is in this spirit that the Russian government recently announced that it would rely on foreign certifications to decide which drugs are sold in Russia.<sup>79</sup>

Despite the fact that many products are subject to extensive regulation prior to marketing, important safety hazards sometimes become apparent only after a product is on the market. In the United States when a product is found to be defective or to pose an unreasonable hazard to health or safety, the manufacturer is required to remove the product from the market. This recall can be initiated by either the manufacturer or a regulatory agency; it can involve the removal of a few bottles of contaminated or mislabeled product or the permanent removal of a product from the marketplace.<sup>80</sup> Recalls can be initiated for products that pose any one of three health hazards.<sup>81</sup> Of the approximately 3,000 citations by the Food and Drug Administration during 1973-78, 2 percent were products with defects that could have seriously adverse health consequences, including death (Jarrell and Peltzman 1984). The other types of health hazards that can prompt recalls are temporary or medically reversible health hazards and hazards unlikely to entail adverse health consequences.

Product safety regulations may sometimes do more harm than good (see Viscusi 1985). Viscusi (1984a, 1984b) considered the case of child-resistant bottle caps for certain drugs. He argued convincingly that when safety designs are too complicated, parents often leave bottles uncapped, thus facilitating children's access to drugs. This fact, combined with the lulling effect that safety aids have on many consumers, explains why the percentage of aspirin poisonings attributed to child-proof caps increased from 40 percent in 1972 to 73 percent in 1978.<sup>82</sup> It should be noted, though, that such concerns are more relevant in industrial economies, where safety regulations are considerably more developed and therefore more likely to address problems whose solutions would have small expected net benefits.

#### *Protection of economic interests*

Many government policies are designed to protect consumers' economic interests in the marketplace by facilitating the process through which consumers acquire and

process information. Such policies include requirements regarding the information that sellers provide to potential buyers through advertising or labeling of products. They also include regulations on contracts between buyers and sellers.

*Advertising and promotional practices.* Economists disagree about the purpose and benefits of advertising.<sup>83</sup> Is advertising designed to systematically fool consumers, calling into question the central tenet of consumer sovereignty? Or, by offering consumers a low-cost way to obtain information, does advertising promote competition and help consumers achieve higher levels of welfare?

The adverse view of advertising is not new (see Kaldor 1950, Nichols 1951, and Galbraith 1958). This view claims that advertising persuades and fools consumers by allowing firms to create artificial product differentiation and increasing barriers to entry (Galbraith 1967 and Solow 1967). When firms compete through advertising rather than prices, advertising is wasteful from a social point of view. The example most often cited by proponents of this view is television advertising, which provides little information beyond the availability of the advertised products. Solow (1977) summarized this position:

Sometimes it comes over me that the TV advertiser does not really care what the ad says. In fact what the commercial actually says is almost always utterly irrelevant or completely inane. It cannot be that the advertiser expects anyone to believe a word of it, that Exxon is in business to make the grass grow, that my Sunoco dealer is all that friendly, that I can actually trust my car to the man who wears a star. It is probably much simpler: when I run out of toothpaste I'm going to buy something. What word will come out of my mouth when I walk up to the counter? God knows: but if I have seen Crest more often than Colgate in prime time this past month, I have a sneaking feeling that the odds are I'm going to buy some Crest (p. 269).

Many European countries have enacted policies consistent with the adverse view of advertising (see Mayer 1989). For example, Denmark has banned television advertising, whereas Norway prohibits the advertising of alcohol and tobacco products as well as advertising that portrays women as sex objects. Almost all European countries restrict television advertising to certain times of the day and require that it be shown in time blocks, thereby making it easier for consumers to avoid commercial messages if they wish. This is in stark contrast to the

practice in Latin America and the United States, where advertising is an integral part of popular TV shows, making it harder for viewers to escape its message.<sup>84</sup>

The negative view of advertising motivates laws against false and deceptive advertising. Many industrializing countries, such as Argentina and Brazil, that have introduced major changes to their consumer protection legislation have included such a law. Yet, to be effective, the transactions costs incurred by consumers who invoke it should be small. The burden of proof, for example, must not be on the consumer's side, as in some developing countries where consumers must show that sellers have deliberately misled them. The regulatory methods used to implement such laws are advertising substantiation rules, mandatory disclosure (such as health warnings on cigarette ads), and provisions for corrective advertising in the case of deceptive ads.

The consumer protection law passed in Mexico in 1975 requires advertisements of sales to indicate how long the special prices will be in effect; otherwise, it is understood that the advertised prices are valid until announced through the same media as originally publicized (see Vargas 1989). Unfortunately, sellers (not only in Mexico but also in industrial countries) have found a way around these laws. Their ads announce that a sale will last "while stock is available." This is a problem for buyers who, after spending time and resources to travel to the store, discover the merchandise is sold out.

Some countries, such as Chile and Germany, rely on private industry-supported advertising councils. These councils often have the right incentives, since false advertising can affect the image of an entire industry. Yet they usually lack both the resources and a mandate to enforce their decisions adequately. For example, when the council concludes that one of its members has run a deceptive ad, the member can choose to terminate its membership in the council, thereby limiting punishment to the loss of reputation that may result from publicity about its actions. If the issue at stake is rather technical, as was the case recently in Chile with a long-distance telephone carrier, the indirect cost paid by the firm running the deceptive ads may be small.

The view that advertising brings useful information to consumers dates to Telser (1964). This view holds that advertising promotes the production of high-quality goods, since it makes it easier for manufacturers of such goods to inform consumers about their products.

Based on the "natural experiment" provided by the 50 states of the United States, proponents of this view have shown that products such as eyeglasses and prescription

drugs are more expensive in states that forbid advertising of these products (see Benham 1972 and Cady 1976). Proponents argue that advertising of search goods fosters competition by reducing the cost of learning about competing products, thereby increasing the elasticity of demand.

Because consumers must try experience goods, they view a good they have tried differently than one they have not tried, even if the two are identical. This provides incentives for firms to advertise new products, in the hope of creating a group of captive buyers. The need to promote pioneering brands and consumer reluctance to switch brands are well-documented reasons that firms advertise experience goods (see Bain 1956 and Schmalensee 1982).

A policy of allowing comparative advertising is consistent with the positive view of advertising. Comparative ads make specific comparisons between the product being advertised and its competitors. Comparative advertising is rare outside the United States; a well-functioning liability apparatus (for example, a false advertising law) is required for comparative advertising to work, since a deterrent against false claims by one manufacturer regarding the product of another is needed.

Allowing advertising by professionals, such as doctors and lawyers, is another policy consistent with a positive view of advertising. Such ads are banned in most countries on the ground that they are “unethical” and degrade the image of the profession involved. Studies for the United States have shown that consumers in states that ban advertising for lawyers’ and physicians’ services pay more on average for these services and have a larger dispersion in the fees they pay.<sup>85</sup>

Consumer protection policies regarding advertising may either limit ads (when the negative view is taken) or facilitate ads (when the positive view is held). The United States is the country that has relied most on measures promoting ads. It is therefore not surprising that U.S. expenditures on advertising—which in 1984 represented between 2 and 3 percent of gross domestic product—are considerably larger on a per capita basis than those of European countries, which often rely on policies limiting advertising. Per capita advertising expenditures in the United States are twice those in Canada, four times those in the United Kingdom, and six times those in France (see *World Advertising Expenditures* 1986).

Both views of advertising undoubtedly have some merit. The relevance of each depends on the product, the nature of the consumer target market (for example, highly educated versus uneducated), and the advertising medium.

Furthermore, a full understanding of advertising requires the inclusion of views from other disciplines, such as social psychology.<sup>86</sup> Thus, for example, one possible answer to Solow’s (1977) question about why TV ads convey so little information has to do with consumers’ attempts to reduce or resolve cognitive dissonance, or psychological inconsistencies.<sup>87</sup> As explained by Akerlof and Dickens (1982):

As the advertising practitioners point out, people do have needs and tastes, and they do buy goods to satisfy them. Some of these needs and tastes are quite obscure or subtle; it may be hard to tell when the needs are being met. In such cases people may want to believe that what they have just bought meets their needs. Advertising gives people some external justification for believing just that. People like to feel that they are attractive, socially adept and intelligent. It makes them feel good to hold such beliefs about themselves. Ads facilitate such beliefs—if the person buys the advertised product (p. 307).

If this view is relevant, then one of the basic tenets of economics, namely, that agents wish to be better informed, is called into question.

What recommendations should be made beyond the advice that countries enact laws against false and deceptive advertising that involve low transactions costs for consumers? Should a developing country foster policies that expand or inhibit advertising? The answer will depend on the information conditions of the market, the degree of product standardization, the advertising medium, and the country’s social and cultural norms.

*Packaging and labeling.* An alternative to direct regulation of products (such as banning dangerous products and imposing design standards) is to regulate the information that sellers must provide through labeling and packaging. Requirements typically address one or more of the following types of information: (1) identification (for example, the country in which the good was manufactured), (2) ingredients (for example, sodium content of food), (3) duration of product effectiveness (for example, expiration date of a drug), (4) comparative performance (for example, energy consumption of a particular refrigerator compared with other brands), (5) information facilitating price comparisons (for example, unit pricing and effective interest rates), (6) conditions under which a good is sold (for example, parts that are under warranty), (7) proper use or care and handling (for example, instructions for washing clothes), and (8) warnings (for example, health risks associated with smoking).

Regulating packaging and labeling information is attractive because, in contrast with direct regulation, consumer choice is enhanced. The idea behind this approach is to help consumers make well-informed decisions. Since what is dangerous for one consumer may be safe for another, this approach allows consumers more choice than does direct regulation.<sup>88</sup> Another advantage of information regulation, compared with direct regulation, is that it is inexpensive (since labels are not costly).<sup>89</sup> A third advantage is that, at least in principle, labeling regulations can be required for all products, in contrast to direct regulation, which is necessarily limited to a small fraction of goods.

Despite these attractive arguments, however, the evidence from industrial countries shows that regulating information provision through labeling is not an effective way of protecting consumers. The problems that industrial countries have encountered suggest that labeling will be even less effective in developing countries. However, this conclusion does not imply that information regulation is useless. Since direct regulation is necessarily limited to a small fraction of hazards, information regulation may be useful for many risks that are not regulated directly. Since information regulation is relatively inexpensive, it may be justified on a cost-benefit basis, even if the expected benefits are small. One example is including a listing of the recommended daily allowances of nutrients on food packages.

To use labeling effectively, consumers must read labels, understand the information, and act on it. There is ample evidence that things go wrong at each of these three stages (see Hadden 1991). First, people often do not read labels, among other reasons because they trust goods that are familiar to them.<sup>90</sup> Periodically changing the information or warning has been found to have some success in cases where consumers stop paying attention to a label because of familiarity. This is why many countries require cigarette manufacturers to alternate among several warning labels.

Second, the information contained in many labels is often quite technical and consequently difficult for most consumers to grasp. This problem is particularly relevant in developing countries, where there are both high functional illiteracy rates and language barriers.<sup>91</sup> Using standardized pictograms to convey information on hazardous products offers a partial solution. For example, Canada adopted a uniform system of pictograms that is taught in school; this system enables almost everyone in that bilingual nation to recognize certain hazards immediately. Label standardization reduces transactions costs for both buyers (information processing) and sellers (deciding what to put on labels). There also is an important positive

externality associated with standardization, which grows with the number of sellers that adopt the standard. Standardization is of limited use, however, since the complexity of many risks makes it impossible to simplify a label without omitting information that some consumers would view as important.

Once consumers have read and understood the information provided on product labels and packaging, they must act on it. Acting rationally in risk situations requires consumers to assess correctly probabilities that are quite small. As noted earlier, biases such as prominence and the belief in personal immunity may prevent individuals from acting on such information. In this case, banning the product or establishing product standards offers more effective protection than does information regulation.

Another limitation of consumer protection through information regulation is that it is highly regressive, an effect that is particularly relevant in most developing countries. Understanding information and then acting on it requires skills that relate to a consumer's level of education, which itself is usually strongly correlated with income. Thus high-income consumers benefit the most from information regulation.<sup>92</sup>

A policy that gives a major role to information regulation to protect consumers in developing countries is not likely to be successful. Yet this is not to say that information regulation should be disregarded. Consider, for example, the significant reduction in information processing time and effort that results from simple requirements such as unit pricing and effective interest rates.<sup>93</sup> The cost of providing this information is so low that, even if only a fraction of consumers benefit from it,<sup>94</sup> the benefits definitely outweigh the costs.

*Unfair contract clauses.* The high transactions costs involved in writing contracts explains why many goods and services are sold with adhesion contracts, which buyers can choose to accept or reject. Adhesion contracts frequently include clauses (sometimes in small print) that are unfair to buyers. If it were costless for consumers to understand the terms of an adhesion contract, regulating these contracts would not be necessary. However, consumers often either do not read or do not understand the terms of a contract. This problem is particularly relevant when sellers can put (psychological) pressure on buyers, as is true for example, for door-to-door sales. One might argue that, in the absence of regulation, consumers would eventually learn from their mistakes. However, this view ignores that this learning process entails high costs for consumers and that sellers may find new and creative clauses that are unfair to consumers.

Regulating some basic aspects of adhesion contracts is justifiable, as is enacting laws that prevent unfair sales practices. The United States was the first country to impose a cooling-off period for door-to-door sales (or, more generally, for sales that do not take place at the seller's usual place of business). Since the U.S. law was enacted in 1972, most European countries and some developing countries (for example, Mexico in 1975 and Brazil in 1990) have followed suit. These laws specify that door-to-door sales must be formalized by a written contract that is binding only a specified number of working days after it is signed or the good is delivered. The cooling-off period is usually five to seven working days, during which time the consumer may rescind the contract with no liability.

A law that provides a cooling-off period for door-to-door sales helps protect consumers from abusive contract clauses at a very low cost. However, for such a law to be effective, door-to-door salespeople must work for an organization that has a formal address, and buyers must remember to ask for the written contract and be able to verify (quickly, by phone) the authenticity of the business address on the contract. Even though desirable, such laws may be expected to be less effective in developing countries with large informal sectors and significant functional illiteracy rates.

Fair credit laws can also protect consumers from abusive practices. These laws aim at making sure that consumers know the true cost of buying on credit, which includes indirect costs such as invoicing and operating charges. The effective interest rate, which summarizes all direct and indirect costs, makes it easier for consumers to compare different credit alternatives and to compare these alternatives with buying with cash.<sup>95</sup>

A fair credit law is desirable in any country. But such a law has limits: It aims mainly at reducing (substantially) consumers' time and effort to compute and compare alternative paying schemes, *not* at imposing ceilings on interest rates or limiting the free market for credit in any other way. An example of a law that went too far is the Mexican Federal Consumer Protection Act of 1975. This law gives the secretariat of commerce the authority to establish maximum rates of interest as well as to ensure that additional charges and interest are not incorporated into the prices of goods and services (see Vargas 1989).<sup>96</sup>

An example of laws that are aimed at unfair sales practices and that have not been very successful are so-called lemon laws (see Nicks 1986). Connecticut enacted the first lemon law in 1980 to help consumers who purchased new cars with serious defects that could not be readily

repaired. By the end of 1986, 40 other states had enacted such a law. Smithson and Thomas (1988) showed that the value consumers give to such laws is relatively small (as low as \$2 for compact cars), since well-established and inexpensive consumer arbitration mechanisms can be just as effective. Because such mechanisms generally do not exist in developing countries, the value of lemon laws would be higher in these countries.

### Major open research topics

Consumer protection benefits from many disciplines in economics and other fields. New developments in micro-economic theory, industrial organization, law and economics, policy analysis, international trade, institutional design, marketing, and psychology often lead to more effective consumer protection policies. This section considers the open research topics in consumer policy.

#### *Expanding the conceptual framework*

What assumptions should be made about consumer behavior when analyzing consumer policy? There is a constant tension throughout this chapter between considering consumers as rational people facing positive transactions costs and assigning an important role to consumers' misperceptions or outright irrational behavior. Where is the balance between the two approaches? Can consumer rationality be quantified? Recent work on the psychology of decisionmaking under uncertainty is promising.<sup>97</sup> However, this school of thought has yet to provide a simple and tractable framework, such as expected utility maximization, that can be used to analyze consumer protection issues. Work by economists looking at concepts in social psychology from an economic point of view has also been helpful.<sup>98</sup> More work of this kind is clearly needed.

All consumers exhibit different degrees of rationality in their buying behavior. Information seekers come closest to economists' ideal of the rational decisionmaker. The relative effectiveness of market remedies as compared with government regulations usually grows with the fraction of consumers who act rationally. Simple models, motivated by empirical data, in which consumers differ in their degree of rationality may be useful for analyzing the effectiveness of consumer protection policies.

#### *Market-based remedies*

Economic theory provides useful insights about when sellers and producers can be expected to provide truthful and relevant information to consumers. More work is needed to verify the empirical relevance of these insights.

Much could also be learned from cross-country studies comparing the extent to which sellers provide information voluntarily.

Casual observation indicates that the coverage of warranties for identical goods varies considerably across developing countries. Documenting this fact, and determining its relation to observable characteristics such as education, the legal redress system, and the number of public and private consumer organizations, among other variables, may be useful in determining appropriate strategies for promoting consumer policy in countries in which consumer protection is embryonic.

#### *Law and economics*

New developments in the field of law and economics can benefit consumer protection considerably. The application of economic concepts to the study of the legal systems in developing countries is a promising research area.

Many developing countries would benefit from a thorough reform of their legal systems. It would not be surprising if reforming the legal system became a high priority once market-oriented reforms are in place. The extent to which radical changes will be possible during such a reform is hard to predict. Nonetheless, studies on the following topics may be useful:

- A study of the comparative effectiveness of jury systems and systems in which judges reach decisions. Such a study should include an analysis of which system produces a more effective outcome for various types of liability cases and of how often each legal system reaches the "correct" decision. Danzon (1991) undertook such an analysis for medical malpractice cases in the United States. Since liability cases are more likely to go to court when the outcome of a trial is uncertain (and both parties differ in their assessment of their chances of winning), a jury system may lead to higher transactions costs than a system in which judges make determinations.
- In some countries, such as the United Kingdom, the losing party in a trial must pay for all litigation costs. Although this practice reduces the number of frivolous cases brought to trial, it also deters litigants who might prevail but who are dissuaded from bringing suit by the uncertainty of the trial's outcome. In most other countries each party pays its own legal costs. It would be useful to study, at an empirical level, the effect of each of these alternatives on the legal system, possibly comparing different countries. The distributive impact of various regimes should be included in such an analysis.

In addition to comparing the percentage of cases that go to trial under each system, the analysis must also assess

the extent to which a fair outcome is achieved and the transactions costs incurred. Any work on this topic must model and estimate the uncertainties involved from the plaintiff's and defendant's point of view, and the actual uncertainty of the outcome, should the case go to trial.

- In some countries the investigations into a case are made by the same judge who later hands down a verdict; in other countries the two roles are carried out by different people. The incentives provided by both approaches differ substantially. Work could be done both to find empirical evidence documenting its relevance.
- Consumer protection seems well suited to common law, in which precedent plays an important role. Remarkably, the Mexican Consumer Protection Act of 1975 relies on common law despite Mexico's strong tradition in civil law. Yet common law has its disadvantages. Specifically, the degree of uncertainty faced by manufacturers is considerably larger than it would be if liability cases proceeded according to well-established codes. Work that pursued these issues in more depth would be useful, particularly if it combined the theoretical and empirical approaches.

#### *Protection of economic interests*

A variety of measures that reduce consumer search costs merit detailed study; for example, the requirement that sellers quote prices by telephone. Such a requirement might substantially reduce consumer search costs, yet it might also facilitate seller collusion in certain markets. Also, enforcement might not be trivial.

More research in the area of advertising could also yield important new insights. For example, studies quantifying the effects of advertising on competition in developing countries might give useful guidelines for regulating (or promoting) advertising so as to protect consumers.

#### **Conclusions**

This chapter has reviewed market-based and regulatory remedies that help protect consumers in a market economy, including legal instruments. These remedies include guarantees offered by sellers and producers' reputation-building efforts, information provision requirements, laws against deceptive business practices, product standards, safety regulations, quality seals, and product-testing magazines.<sup>99</sup> Rather than offer a cookbook with explicit recipes for protecting consumers in developing countries, this chapter has detailed the strengths and limitations of those remedies.

A given level of protection can be achieved at similar costs by very different combinations of policy instru-

ments. Policies on product safety offer an example of why a holistic approach to designing consumer protection policies is important.<sup>100</sup> Product safety can be achieved via product liability law, regulation, actions by producers, or exercise of care by consumers. These approaches differ in their informational requirements, the incentives they create to provide new information about emerging risks, their ability to respond to change, and the costs involved and their distribution. For example, regulation may involve setting standards, requiring testing and disclosure, or banning products completely. Information requirements prevent using regulation to achieve adequate levels of safety for more than a small fraction of goods. Thus, although regulation may be effective in providing adequate safety levels for many products, it must necessarily be combined with other approaches.

#### *Relevant differences in industrializing countries*

Most of the consumer protection literature has been motivated by, and relates to, problems facing consumers in industrial countries. This is so not only because industrial countries spend more on research; consumer protection also is a more important issue in industrial economies (for possible explanations, see chapter 7). This raises the question, what differences between developing and industrial countries are relevant for analyzing consumer policies?

First, developing countries have higher functional illiteracy rates. Consumers find it more difficult to comprehend a label or instruction manual and have a harder time evaluating risks in situations they have not encountered previously. This implies that more resources should be spent in preparing information materials. Using pictograms to communicate risks may be particularly important in developing countries. Higher functional illiteracy rates also imply that product standardization is more desirable in developing economies. For example, regulating use of the term "full warranty" would be particularly important.

Because of their low educational levels, consumers in developing countries are more vulnerable to deceptive business practices, such as false advertising. The welfare of low-income consumers may improve considerably if successful policies to protect them are in place.

Second, institutions that could play an important role in protecting consumers often function poorly in developing countries. For example, it is both difficult and inefficient to achieve adequate access to redress for consumers in a country whose legal system does not function well. Another institutional issue is how well government agen-

cies function.<sup>101</sup> Some countries lack the administrative capability to monitor consumer risks or enforce safety regulations; in other countries the institutional design of government agencies facilitates the co-optation of regulators by those they are supposed to supervise. In such countries promoting consumer protection through market mechanisms may be the most effective approach in the short run.

Closely related is the issue of government accountability. Government agencies in many developing countries bear no responsibility for providing poor or dangerous services. The lack of response of public utilities to consumer concerns is one of the arguments given for their privatization.

Third, many markets in developing economies do not function adequately for both institutional and informational reasons. Consider the car insurance market. When such a market begins to develop, insurance providers have no individual driving records. The problems of asymmetric information and adverse selection are particularly acute at this stage. Informational problems thus limit the size and operation of insurance markets during the early stages of its development.<sup>102</sup>

Thorelli (1982, 1983), who considered developing countries including Thailand, Kenya, and China, concluded that markets in developing countries share three deficiencies: (a) a majority of goods are manufactured locally without adequate quality control, (b) transportation and storage facilities are inadequate for preserving fresh foods, and (c) sellers care little about consumer satisfaction and frequently sell adulterated goods or cheat customers with respect to weights and measures. Since consumers are often poor and uneducated, Thorelli advocates government regulation not only to ensure adequate degrees of safety, but also to provide minimal quality standards.

Thorelli's work in developing countries is less relevant today than it was a decade ago, since globalization has reduced the difference between markets in industrial and developing countries. He equated developing-country consumers with poor consumers buying locally produced goods, ignoring a growing middle class with increasing access to goods from abroad.

#### *Additional policy conclusions*

Some final points are worthy of note here.

*Reputation and quality.* Sellers have incentives to provide truthful information on the quality of their goods when they know that failing to do so will be costly for them. They are more likely to offer truthful information in the case of repeat purchases and for goods whose true

quality the consumer can determine quickly. It follows that policies that foster long-run relations between sellers and customers are advisable.

*Private provision of information by third parties.* Private provision of information by third parties can play an important role in protecting consumers. Product-testing magazines and safety and quality seals provide useful information to consumers about products. When important externalities are involved, as with consumer magazines, government subsidies are advisable.

*Protection from abusive business practices.* Consumers in developing countries would benefit from laws protecting them from abusive business practices. Limitations on the provision of adhesion contracts, laws against deceptive advertising, and minimal requirements for producers' claims that goods are guaranteed, are all measures that can help protect consumers from abusive practices.

*Information provision requirements.* Information provision would seem an attractive way of protecting consumers, since it leads to better decisions without limiting consumer choice. Yet for this approach to be effective, consumers must read the information, understand it, and act on it. Ample evidence shows that things often go wrong at each of these steps. High illiteracy levels in developing countries further limit the value of consumer information. This does not mean that information provision has no role to play in consumer protection. For example, unit pricing significantly reduces the time and effort needed to compare the prices of various brands. The number of consumers who compare prices will grow significantly with the introduction of unit pricing, yet not as much as it would if all consumers were rational.

## Notes

The author thanks Eduardo Bitrán, Peter Diamond, Mark Dutz, Tim Ennis, Juan Escudero, Ronald Fischer, Claudio Frischtak, Daniel Kaufmann, Raúl Leal, Patricio Meller, Pablo Serra, Ennio Stacchetti, and in particular Richard Zeckhauser, for helpful comments and suggestions. The outstanding research assistance of Alexis Camhi and Alejandro Micco is also deserving of mention.

1. Consumer policies in New Zealand and Australia are explicitly aimed mainly at the disadvantaged. See McGregor 1991.
2. The relevance of this argument grows with the degree of correlation between income and people's abilities. Note, however, that some consumer protection policies may benefit high-income consumers at the expense of poor consumers. An example is offered below.
3. See Maynes 1988. The first four rights were introduced by President John Kennedy's influential 1962 address on consumer issues. See Lampman 1988 for details and Nadel 1971 for another

view on the importance of this address. The right to consumer education was formulated by President Gerald Ford in 1975; see Mohr 1988 for details.

4. See Oi 1972, 1977 for seminal works on the economics of safety.
5. The safety level provided is efficient if the cost to society of increasing safety provision slightly is equal to the social benefits this increase generates. Of course, this ignores distribution issues.
6. In this case consumer policies may harm the poor. This may happen more generally with consumer policies aimed at raising the quality of goods. To the extent that quality controls raise both quality and price, poor people may be hurt relative to those who are better off. The poor are unable to afford the higher-quality good and therefore lose when the low-quality good is no longer produced. The better-off benefit from the withdrawal of low-quality goods from the market both because of economies of scale and by avoiding mistaken purchases.
7. See United Nations 1986. Also see Merciai 1986, Harland 1987, and, for the case against the guidelines, Weidenbaum 1987.
8. This question is important because a law requiring sellers to provide all information that a consumer might find relevant is impossible to enforce. After all, information is an unusual good.
9. The word "consumer protection" did not appear in the *New York Times Index* until 1969.
10. Also presumed are no large indivisibilities relative to the economy and universality of markets.
11. By a suitable reallocation of initial resources. This result extends to the case with uncertainty, as long as universality of markets is understood to include markets for contingent claims.
12. So as to avoid interfering with the efficiency properties of the price mechanism.
13. For a more detailed review of many of the concepts covered in this section, see Arrow 1970, Pindyck and Rubinfeld 1992, Milgrom and Roberts 1992, Nicholson 1992, Varian 1992, Kreps 1990, and Tirole 1988.
14. The concept of moral hazard originated in the insurance industry. If a consumer insures his car, the fact of having insurance may lead to careless driving and make accidents more likely. Thus buying the good (insurance) changes the consumer's behavior in a way that makes the "production" of the good more expensive.
15. Consider a group of people with similar driving records. Since these records capture a combination of chance events and driving ability, some people will be better drivers than others and, other things being equal (such as the degree of risk aversion), will be prepared to pay less for insurance than drivers who know that they themselves, not chance events, caused their accidents. Some of the better drivers will decide not to buy insurance, thereby raising costs for the remaining drivers, forcing additional (better) drivers out of the insurance market, and so on. Taken to the extreme, this process of adverse selection may lead to the disappearance of an entire market; more generally, it results in markets that are considerably

smaller than they would be in a world with full information. See the pioneering work by Akerlof (1970).

16. The principal-agent problem is how to find ways to ensure that one individual, the agent, acts effectively on behalf of another, the principal. The problem is relevant only when there is uncertainty and the information available to the two participants is asymmetric. Under such circumstances, the principal cannot infer from observable evidence how effectively the agent has acted on his or her behalf and thus cannot judge the extent to which an observable outcome was determined by the agent's behavior and the extent to which it was caused by events beyond the agent's control.

17. This is the exclusion principle of Musgrave (1959). What is technically impossible at one moment in time may become technically feasible later; consider, for example, peak and off-peak usage of electricity, roads, and telephones.

18. Similarly, negative externalities are addressed through taxation.

19. That a good is public does not mean that it is provided by the government.

20. Perfect price discrimination consists of charging every consumer his or her reservation price. This practice is also referred to as first-degree price discrimination.

21. For example, American consumers can choose from more than 25,000 products at the supermarket; they can read any of 11,000 magazines or periodicals; and they can view more than 50 television stations. See Williams 1990.

22. The economics of information, and the importance of price search, originate with Stigler (1961). See Russo 1988 for a more detailed analysis along lines similar to what follows.

23. Such an analysis should consider how the policies being considered affect the variety of informal channels through which consumers acquire information.

24. Cardinal utility refers to a quantitative measure of the individual's welfare in a particular scenario. It should be contrasted with ordinal utilities, where such assignments are meaningless.

25. One of Savage's axioms is less appealing than it initially appeared to be—an issue I do not explore in this chapter. See Machina 1982.

26. I intentionally avoid the discussion of the objective and subjective interpretations of probabilities. The eclectic approach adopted here seems a reasonable compromise for the policy issues considered in this chapter.

27. This and other issues considered in this section are illustrated with vivid examples in Zeckhauser and Viscusi 1990.

28. For a representative collection of papers from this literature, see Kahneman, Slovic, and Tversky 1982.

29. See Plous 1993 for an up-to-date text on the psychology of judgment and uncertainty, including a detailed exposition of the concepts mentioned below.

30. The public's ratings are summarized by ranking the health risks from most to less dangerous; experts classify the health risks as high, medium, or low. The experts rate nine of the health risks as

high. The sum of the ratings by the public of these nine risks is 105; if the public's assessment were independent of the experts', this sum would (on average) be equal to 103.5. Thus a pure significance test (see Cox and Hinkley 1974) would give a p-value close to 0.50.

31. A closely related concept is that of bounded rationality; see Simon 1955, 1956.

32. Other reasons are the "cost" of creating the habit of using a safety belt and social norms. For more details, see chapter 7 of this volume.

33. See Ausubel 1991. For a dissenting view, see Brito and Hartley 1995.

34. Another important issue is how consumers gain access to information about prices and where goods are sold. This matters for goods in the three categories defined above and is considered in the section on market-based remedies.

35. It would seem that too much variety cannot be bad, yet this ignores the price that consumers pay for goods. When firms have market power, more variety may come at the expense of higher prices for all goods, which may make consumers worse off.

36. Risk-neutral producers may be induced to behave in a risk-loving manner by the asymmetry described above. The situation is analogous to the one considered in Stiglitz and Weiss 1981.

37. Regulations designed to change consumer behavior are considered in chapter 7 of this volume.

38. Even a full warranty system, however, may fall short of protecting consumers. These limitations relate to transactions and enforcement costs, on which the discussion focuses shortly.

39. An example is provided by automobiles. The duration of warranties for different auto parts relates to the extent that deficiencies depend on customer behavior. Also, to reduce the moral hazard problem, car manufacturers issue warranties requiring that cars be serviced by recognized dealers and only with new parts.

40. Strictly speaking, government regulation is not a pure market remedy. Still, it is more natural to discuss this topic here than in the next section.

41. In the United States the warranty disclosure provisions of the Magnuson-Moss Warranty Act of 1975 include the requirement that goods that cannot be repaired within a reasonable time must be replaced or the purchaser given a full refund.

42. Arthur Young & Co. 1979 and Schmitt, Kanter, and Miller 1979, quoted in Mayer 1989.

43. Watermelons are much less expensive in developing countries than in most industrial countries.

44. Manufacturers may have other reasons for not providing guarantees for products sold by informal retailers. First, formal retailers may threaten not to sell their goods. Second, if manufacturers can exert market power, it may be to their advantage not to sell their product at a low price.

45. In such a situation sellers are usually credit-constrained.

46. *The Economist*, December 19, 1992.

47. The high commissions and advertising expenses can be inferred from the fact that workers approaching retirement age receive

phone calls from annuity sellers day and night to offer them trips to Miami and similar incentives for buying a particular annuity. An “informal” market providing the (confidential) home phone numbers of workers nearing retirement has developed.

48. The trade-off between simplifying customers’ information processing and limiting their freedom of choice appears repeatedly when analyzing consumer policy.

49. See Tirole 1988, p. 107 for a simple model making this point. This model borrows elements from Salop 1977, Salop and Stiglitz 1977, and Wolinsky 1983.

50. There is evidence that the U.S. consumer price index may be overestimated as a result.

51. The appropriateness of this and more sophisticated dynamic pricing strategies depends on how consumers gather information and to what extent they act strategically; see Lazear 1986, Pashigian 1988, Pashigian and Bowen 1991, and Bitran and Mondschein 1993, among others.

52. There is convincing evidence that price dispersion in some markets is considerably larger than can be accounted for by any of the explanations discussed here (Pratt, Wise, and Zeckhauser 1979).

53. This is, once again, the hidden quality problem mentioned in the introduction.

54. The case of private parties is considered in this section. Public organizations are considered in chapter 7.

55. Since the firms issuing the bonds pay for the ratings, consumers pay for the ratings only indirectly.

56. These were defined in 1987 by the International Organization of Standardization.

57. ISO certification can be provided by both private and public organizations. Differences among the qualities of certifiers may emerge, as in Brazil, where ISO certification by a foreign certifier is more highly valued than certification by a local group.

58. The latter is provided by the Consumer Product Safety Association, a private Japanese organization.

59. DIN receives approximately 15 percent of its budget from the government, in exchange for which it must give priority to state requests to establish norms in particular fields.

60. Certifiers are usually liable only for accidents that occur within the country of issue.

61. See *ISO Bulletin*, July 1993. Similar problems occurred in Nigeria; see Agege 1987.

62. This problem is more relevant for consumer magazines than for news magazines, since the information provided by consumer magazines becomes obsolete at a slower rate.

63. Government regulation may be called for in the latter case.

64. An alternative is that government agencies provide this information directly; see chapter 7. With the exception of the United Kingdom, governments in Western Europe provide financial support for product-testing organizations.

65. The two major consumer organizations have already benefited

from this provision; see Salgado and others 1994.

66. Complementary measures are needed to prevent such a subsidy from acting as a barrier to entry.

67. See Inman 1987 and Wolf 1988 for reviews of the failures of both governments and markets.

68. Spence 1977 considers a signaling model in which consumers underestimate failure probabilities and shows that safety is underprovided in the free-market equilibrium.

69. Since developing countries can free-ride on the knowledge on risk available in industrial economies, this is not a strong argument in favor of a liability law in developing countries.

70. Viscusi 1991a and the articles in the summer 1991 issue of the *Journal of Economic Perspectives*—especially Cooter 1991, Danzon 1991, Priest 1991, Shapiro 1991, and Viscusi 1991b—include rich discussions of theoretical issues pertaining to liability law and authoritative accounts of the U.S. liability crisis. A compelling argument in favor of regulation when large risks are involved can be found in Breyer 1993. Gerner 1988 and Crandall 1988 consider both government regulation and liability law when discussing consumer safety issues. This section draws from these sources.

71. As mentioned earlier, this poses a moral hazard problem on the consumer’s side.

72. Since liability cases are more likely to go to court when the outcome of a trial is uncertain (and both parties differ in their assessment of their chances of winning), a jury system may lead to higher transactions costs than a system in which judges decide.

73. See *El Mercurio*, November 3 (p. C13), 4 (p. C11), and 5 (p. C11), 1993.

74. Alternative redress mechanisms for consumer grievances that circumvent the costly legal system are discussed in chapter 7.

75. This assumes that some consumers are unaware of the risks, do not read labels, and so on.

76. Zeckhauser 1979, 1985 analyzes food safety regulation in the United States based on this principle.

77. Risk-benefit analysis is not discussed in more detail here. See Crandall 1988 for a survey and Zeckhauser 1979, 1985 for an application to food safety.

78. See chapter 7 for an analysis of what determines whether a risk becomes an issue.

79. See Morgan 1993 for a more detailed discussion of risk analysis and risk management.

80. See Jarrell and Peltzman 1984 for a study of the direct and indirect costs faced by firms subject to a product recall.

81. This is based on the classification of the Food and Drug Administration.

82. This goes beyond the usual moral hazard problem on the consumer’s side, since consumers confuse a reduction in risk with a total elimination of risk, thereby leading to an increase in accidents. This phenomenon is related to misassessments of low-probability events; see the first section of this chapter.

83. See chapter 7 in Tirole 1988 for an insightful discussion. This section draws from this source.
84. The most popular television show in Latin America, "Sábado Gigante," is an extreme example.
85. Note, though, that Rizzo and Zeckhauser 1990 shows that advertising inhibits entry of new physicians.
86. See Becker and Murphy 1993 for a dissenting view.
87. The theory of cognitive dissonance is due to Festinger 1957. According to this theory, people are motivated to reduce psychological inconsistencies. See Plous 1993 for details.
88. The relation between labeling and consumer protection is explored in Hadden 1986, 1991; this section draws from these sources.
89. Deciding what the labels should say and enforcing the information requirements may be more expensive.
90. Most people don't listen to the safety instructions announced on airplanes before takeoff, for example.
91. Another difficulty that consumers in developing countries face is that the labels of imported goods are written either in a foreign language or in unintelligible translations. This point is considered further in chapter 7.
92. This argument ignores the fact that the opportunity cost of the time needed to process the information is lower for those with less education; this effect should be small. Alternative policies for reducing risk usually have the effect of increasing prices. Which approach is more regressive must be determined on a case-by-case basis.
93. Unit pricing is labeling a good not only with its price but also with its price per standard unit (for example, pound or kilogram). See Russo and Leclerc 1991 for estimates of the time savings involved.
94. Because many consumers do not act rationally. See the first section of this chapter.
95. The present value of payments, including direct and indirect costs, can be used alternatively.
96. The International Organization of Consumer Unions also advocates ceilings on interest rates.
97. See Plous 1993 for a comprehensive text on the psychology of decisionmaking, in particular, on prospect analysis, an approach due to Kahneman and Tversky 1979.
98. Examples are Akerlof and Dickens 1982 and Akerlof 1991.
99. Education campaigns were also considered briefly. This topic is treated in more depth in chapter 7, which also considers private and public consumer organizations and special court proceedings.
100. The Japanese consumer policy framework is famous for following such an approach; see McGregor 1991.
101. Needless to say, this is often a problem in industrial countries as well.
102. This is so for two reasons. First, it reduces insurance prices, since fixed costs are shared by a larger pool of customers. Second, the cost of collecting damages is lower when both parties involved are insured.

## References

- Agee, C. O. 1987. "Quality Standards for Products: An Analysis of Standards Legislation in Nigeria." *Journal of Products Liability* 10: 277-300.
- Akerlof, G. 1970. "The Market for Lemons: Qualitative Uncertainty and the Market Mechanism." *Quarterly Journal of Economics* 84: 488-500.
- . 1991. "Procrastination and Obedience." *American Economic Review* 81: 1-19.
- Akerlof, G. A., and W. T. Dickens. 1982. "The Economic Consequences of Cognitive Dissonance." *American Economic Review* 72: 307-19.
- Arrow, K. J. 1970. "Political and Economic Evaluation of Social Effects and Externalities." In J. Margolis, ed., *The Analysis of Public Output*. New York: National Bureau of Economic Research.
- Arthur Young & Co. 1979. *Warranties Rules and Warranty Content Analysis*. Washington, D.C.: Federal Trade Commission.
- Ausubel, L. M. 1991. "The Failure of Competition in the Credit Card Market." *American Economic Review* 81:50-81.
- Bain, J. 1956. *Barriers to New Competition*. Cambridge, Mass.: Harvard University Press.
- Bannister, R., and C. Monsma. 1982. *Classification of Concepts in Consumer Education*. Monograph 137. Cincinnati, Ohio.: South Western Publishing Company.
- Becker, G. S., and K. M. Murphy. 1993. "A Simple Theory of Advertising as a Good or Bad." *Quarterly Journal of Economics* 108: 941-64.
- Benham, L. 1972. "The Effects of Advertising on the Price of Eyeglasses." *Journal of Law and Economics* 15: 337-52.
- Bitrán, E. 1994. Personal communication. April.
- Bitrán, G., and S. Mondschein. 1993. "Pricing Perishable Products: An Application to the Retail Industry." forthcoming in *Management Science*.
- Breyer, S. 1993. *Breaking the Vicious Circle: Toward Effective Risk Reduction*. Cambridge, Mass.: Harvard University Press.
- Brito, D. L., and P. R. Hartley. 1995. "Consumer Rationality and Credit Cards." *Journal of Political Economy* 103:400-33.
- Cady, J. 1976. "An Estimate of the Price Effects of Restrictions on Drug Price Advertising." *Economic Inquiry* 14: 493-510.
- Coase, R. 1960. "The Problem of Social Cost." *Journal of Law and Economics*.
- . 1974. "The Lighthouse in Economics." *Journal of Law and Economics* 17:357-76.
- Cooter, R. D. 1991. "Economic Theories of Legal Liability." *Journal of Economic Perspectives* 5: 11-30.
- Cox, D. R., and D. V. Hinkley. 1974. *Theoretical Statistics*. London: Chapman and Hall.
- Crandall, R. W. 1988. "The Use of Cost-Benefit Analysis in Product Safety Regulation." In E. Scott Maynes, ed., *The Frontier of*

## REGULATORY POLICIES AND REFORM: A COMPARATIVE PERSPECTIVE

- Research in the Consumer Interest*. Columbia, Mo.: American Council of Consumer Interests.
- Danzon, P. M. 1991. "Liability of Medical Malpractice." *Journal of Economic Perspectives* 5: 51-70.
- Darby, M., and E. Karni. 1973. "Free Competition and the Optimal Amount of Fraud." *Journal of Law and Economics* 16:67-88.
- Day, G., and W. Brandt. 1974. "Consumer Research and the Evaluation of Information Disclosure Requirements: The Case of Truth in Lending." *Journal of Consumer Research* (June): 21-32.
- Festinger, L. 1957. *A Theory of Cognitive Dissonance*. Evanston, Ill.: Row, Peterson.
- Galanter, M. 1985. "Legal Torpor: Why So Little Has Happened in India after the Bhopal Tragedy." *Texas International Law Journal* 20: 273-94.
- Galbraith, K. 1958. *The Affluent Society*. Boston: Houghton-Mifflin.
- . 1967. *The New Industrial State*. Boston: Houghton-Mifflin.
- Gerner, J. L. 1988. "Product Safety: A Review." In E. Scott Maynes, ed., *The Frontier of Research in the Consumer Interest*. Columbia, Mo.: American Council of Consumer Interests.
- Hadden, S. G. 1986. *Read the Label: Reducing Risk by Providing Information*. Boulder: Westview Press.
- . 1991. "Regulating Product Risks Through Consumer Information." *Journal of Social Issues* 47: 93-105.
- Harland, D. 1987. "The United Nations Guideline for Consumer Protection." *Journal of Consumer Policy* 10: 245-66.
- Inman, R. P. 1987. "Markets, Governments, and the 'New' Political Economy." In A. Auerbach and M. Feldstein, eds., *Handbook of Public Economics*. Amsterdam: North-Holland.
- Jarrell, G., and S. Peltzman. 1984. "The Impact of Product Recalls on the Wealth of Sellers." In P. M. Ippolito and D. T. Scheffman, eds., *Empirical Approaches to Consumer Protection Economics*. Washington, D.C.: Federal Trade Commission.
- Kahneman, D., and A. Tversky. 1979. "A Prospect Theory: An Analysis of Decisions under Risk." *Econometrica* 47: 263-91.
- Kahneman, D., P. Slovic, and A. Tversky. 1982. *Judgment under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.
- Kaldor, N. 1950. "The Economic Aspects of Advertising." *Review of Economic Studies* 18: 1-27.
- Klitgaard, R. 1991. *Adjusting to Reality*. San Francisco: International Center for Economic Growth.
- Kreps, D. 1990. *A Course in Microeconomic Theory*. Princeton: Princeton University Press.
- Laffont, J. J. 1989. *The Economics of Uncertainty and Information*. Cambridge, Mass.: MIT Press.
- Lampman, R. J. 1988. "JFK's Four Consumer Rights: A Retrospective View." In E. Scott Maynes, ed., *The Frontier of Research in the Consumer Interest*. Columbia, Mo.: American Council of Consumer Interests.
- Lane, S. 1983. "The Rationale for Government Intervention in Seller-Consumer Relationships." *Policy Studies Review* 2: 419-28.
- Lazear, E. P. 1986. "Retail Pricing and Clearance Sales." *American Economic Review* 76:14-32.
- Leff, A. 1976. *Swindling and Selling: The Spanish Prisoner and Other Bargains*. New York: Free Press.
- Machina, M. 1982. "Expected Utility Analysis without the Independence Axiom." *Econometrica* 50: 277-323.
- Mayer, R. N. 1989. *The Consumer Movement: Guardians of the Marketplace*. Boston: Twayne Publishers.
- Maynes, E. S. 1988. "The First Word, The Last Word." In E. Scott Maynes, ed., *Proceedings of the International Conference in the Consumer Interest*. Columbia, Mo.: American Council of Consumer Interests.
- McGregor, S. L. T. 1991. "International Consumer Policy." Mount Allison University, Canada.
- Merciai, P. 1986. "Consumer Protection and the United Nations." *Journal of World Trade Law* 20: 206-31.
- Milgrom, P., and J. Roberts. 1992. *Economics, Organization and Management*. Englewood Cliffs, N.J.: Prentice-Hall.
- Mohr, L. A. 1988. "The Role of Federal Government." In E. Scott Maynes, ed., *Proceedings of the International Conference in the Consumer Interest*. Columbia, Mo.: American Council of Consumer Interests.
- Morgan, M. G. 1993. "Risk Analysis and Management." *Scientific American* (July).
- Musgrave, R. A. 1959. *The Theory of Public Finance: A Study in Public Economy*. New York: McGraw-Hill.
- Nadel, M. 1971. *The Politics of Consumer Protection*. Indianapolis: Bobbs-Merrill.
- Nelson, P. 1970. "Information and Consumer Behavior." *Journal of Political Economy* 78: 311-29.
- Nichols, W. 1951. *Price Policies in the Cigarette Industry*. Nashville: Vanderbilt University Press.
- Nicholson, W. 1992. *Microeconomic Theory*. 5th ed. Orlando: Dryden Press.
- Nicks, S. J. 1986. "Lemon Laws in the United States: More Hype than Help." *Journal of Consumer Policy* 6: 79-90.
- Nisbett, R. E., and L. Ross. 1980. *Human Inference: Strategies and Shortcomings of Social Judgment*. Englewood Cliffs, N.J.: Prentice Hall.
- Oi, W. Y. 1972. "The Economics of Product Safety." *Bell Journal of Economics and Management Sciences* 4: 3-28.
- . 1977. "Safety at Any Price?" *Regulation* (November/December).
- Pashigian, B. P. 1988. "Demand Uncertainty and Sales: A Study of Fashion and Markdown Pricing." *American Economic Review* 78: 936-53.

- Pashigian, B. P., and B. Bowen. 1991. "Why Are Products Sold on Sale? Explanations of Pricing Regularities." *Quarterly Journal of Economics* 106: 1015–38.
- Pindyck, R., and D. Rubinfeld. 1992. *Microeconomics*. 2nd ed. New York: Macmillan.
- Plous, S. 1993. *The Psychology of Judgment and Decision Making*. New York: McGraw-Hill.
- Pratt, J., D. Wise, and R. Zeckhauser. 1979. "Price Differences in Almost Competitive Markets." *Quarterly Journal of Economics* 93:189–211.
- Priest, G. L. 1991. "The Modern Expansion of Tort Liability: Its Sources, Its Effects and Its Reform." *Journal of Economic Perspectives* 5: 31–50.
- The RAND Corporation. 1987. *Trends in Tort Litigation Special Report*. Santa Monica, Calif.
- Rizzo, J. A., and R. J. Zeckhauser. 1990. "Advertising and Entry: The Case of Physician Services." *Journal of Political Economy* 98: 476–500.
- Russo, J. E. 1988. "Information Processing from the Consumer's Perspective." In E. Scott Maynes, ed., *Proceedings of the International Conference in the Consumer Interest*. Columbia, Mo.: American Council of Consumer Interests.
- Russo, J. E., and F. Leclerc. 1991. "Characteristics of Successful Product Information Programs." *Journal of Social Issues* 47: 73–92.
- Salgado, L. H., J. A. Kelly, O. A. Martinez, and G. Pais. 1994. "Competition and Consumer Protection Policy in Mercosur" (in Portuguese). Inter-American Development Bank, Washington, D.C.
- Salop, S. 1977. "The Noisy Monopolist." *Review of Economic Studies* 44: 393–406.
- Salop, S., and J. Stiglitz. 1977. "Bargains and Rip-offs: A Model of Monopolistically Competitive Price Dispersion." *Review of Economic Studies* 44: 493–510.
- Savage, L. 1954. *The Foundations of Statistics*. New York: John Wiley.
- Schelling, T. 1981. "Economic Reasoning and the Ethics of Policy." *The Public Interest* 63: 37–61.
- Schmalensee, R. 1982. "Product Differentiation Advantages of Pioneering Brands." *American Economic Review* 72: 349–65.
- Schmitt, J., L. Kanter, and R. Miller. 1979. "Impact of the Magnuson-Moss Warranty Act." Staff Report. Federal Trade Commission, Washington, D.C.
- Shapiro, C. 1991. "Symposium on the Economics of Liability." *Journal of Economic Perspectives* 5: 3–10.
- Shavell, S. 1984. "A Model of the Optimal Use of Liability and Safety Regulation." *Rand Journal of Economics* 15.
- Simon, H. A. 1955. "A Behavioral Model of Rational Choice." *Quarterly Journal of Economics* 69: 99–118.
- . 1956. "Rational Choice and the Structure of the Environment." *Psychological Review* 63: 129–38.
- Slovic, P., Fischhoff, B., and S. Lichtenstein. 1977. "Cognitive Processes and Societal Risk Taking." In H. Jungermann and de Zeeum, eds., *Decision Making and Change in Human Affairs: Proceedings on the Fifth Research Conference on Subjective Probability, Utility, and Decision Making*. Dordrecht, Holland: D. Reidel.
- Smithson, C. W., and C. R. Thomas. 1988. "Measuring the Cost to Consumers of Product Defects: The Value of 'Lemon Insurance'." *Journal of Law and Economics* 31: 485–502.
- Solow, R. M. 1967. "The New Industrial State or Son of Affluence." *Public Interest* 9: 100–108.
- . 1977. Review of *Swindling and Selling: The Spanish Prisoner and Other Bargains*, by A. A. Leff. *Bell Journal of Economics* 8: 627–29.
- Spence, M. 1977. "Consumer Misperceptions, Product Failure and Producer Liability." *Review of Economic Studies* 44: 561–72.
- Stigler, G. J. 1961. "The Economics of Information." *Journal of Political Economy* 69: 213–85.
- Stiglitz, J. E., and A. Weiss. 1981. "Credit Rationing in Markets with Imperfect Information." *American Economic Review* 71: 393–410.
- Telser, L. G. 1964. "Advertising and Competition." *Journal of Political Economy* 72: 537–63.
- Thorelli, H. 1982. "Consumer Policy in the Third World." *Journal of Consumer Policy* 5: 197–211.
- . 1983. "Consumer Policy in Developing Countries." In K. P. Goebel, ed., *Proceedings of the 29th Annual Meeting*. Columbia, Mo.: American Council of Consumer Interests.
- Thorelli, H., H. Becker, and J. Engledow. 1975. *The Information Seekers*. Cambridge, Mass.: Ballinger.
- Tirole, J. 1988. *The Theory of Industrial Organization*. Cambridge, Mass.: MIT Press.
- United Nations. 1986. *Guidelines for Consumer Protection*. New York: United Nations, Department of International Economic and Social Affairs.
- Vargas, J. A. 1989. "An Overview of Consumer Transactions Law in Mexico: Substantive and Procedural Aspects." *New York Law School Journal of International and Comparative Law* 10: 345–82.
- Varian, H. 1992. *Microeconomic Analysis*. 3rd ed. New York: W. W. Norton.
- Viscusi, W. K. 1984a. "The Lulling Effect: The Impact of Child-Resistant Packaging on Aspirin and Analgesic Ingestions." *American Economic Review* 74: 324–27.
- . 1984b. "Regulating Consumer Product Safety." Washington D.C.: American Enterprise Institute for Public Policy Research.
- . 1985. "Consumer Behavior and the Safety Effects of Product Safety Regulation." *Journal of Law and Economics* 23: 527–54.
- . 1991a. *Reforming Products Liability*. Cambridge, Mass.: Harvard University Press.

REGULATORY POLICIES AND REFORM: A COMPARATIVE PERSPECTIVE

- . 1991b. "Product and Occupational Liability." *Journal of Economic Perspectives* 5: 71–92.
- Weidenbaum, M. 1987. "The Case Against the UN Guidelines for Consumer Protection." *Journal of Consumer Policy* 10: 425–32.
- Williams, L. 1990. "Decisions, decisions, decisions: Enough!" *New York Times*, February 14, pp. B1, B5.
- Wolf, C. 1988. *Markets or Governments: Choosing between Imperfect Alternatives*. Cambridge, Mass.: MIT Press.
- Wolinsky, A. 1983. "Prices as Signals of Product Quality." *Review of Economic Studies* 50: 647–58.
- Zeckhauser, R. J. 1979. "Social and Economic Factors in Food Safety Decision-Making." *Food Technology* (November).
- . 1985. "Measuring Risks and Benefits of Food Safety Decisions." *Vanderbilt Law Review* 38: 539–82.
- Zeckhauser, R. J., and W. K. Viscusi. 1990. "Risk within Reason." *Science* 248: 559–64.

# Beyond the basics of consumer protection

**Eduardo Engel**

The importance of consumer issues grows with mass consumption and mass production. In rural societies people produce most of the goods they consume. In such societies consumer policy plays a limited role.<sup>1</sup> As societies industrialize, people abandon the countryside or village and move to the city, where they buy goods and services from anonymous sellers and must choose among different brands of varying quality and price. As incomes increase, goods become more sophisticated and therefore more difficult to judge. Buying a durable good, such as a refrigerator, requires more skill and judgment than buying a shirt, for example, for two reasons. First, it is harder to assess the quality of a refrigerator. Second, because consumers purchase refrigerators more infrequently than they do shirts, they are less likely to know the refrigerator seller and less likely to trust the information provided. When consumers are not familiar with sellers, they also become more vulnerable to deceptive business practices.

As economies develop and the need for both consumer education and consumer protection increases, consumer policies should become a central concern of policymakers.<sup>2</sup> Educated consumers help a market economy function better. By scrutinizing the products they buy, educated consumers encourage manufacturers to produce high-quality goods. Despite the desirability of consumer policies in industrializing countries, in most cases they play only a secondary role. Understanding why this is so requires considering the political economy of consumer protection policy.

Consumer education also has an important role to play in protecting consumers' economic interests during transition to a market-oriented economy. Once price controls are lifted, the vulnerability to which buyers are exposed increases dramatically. As illustrated by the MMM pyramid scheme in Russia, uneducated consumers easily fall victim to large-scale scams. The chances of success of market reforms is enhanced when citizens perceive the

new economic system as being fair. Consumer education and consumer policies increase dramatically the likelihood that a reform will be successful and the chance that citizens will support it with their votes.

Consumer policies can be divided into two groups, according to whether they attempt to modify the informational environment faced by consumers or to change consumer behavior itself. The former are referred to as consumer protection policies and are covered in chapter 6. This chapter considers policies aimed at changing how consumers behave, which can be called "consumer promotion." Consumer promotion policies are discussed in the first section. The relation between consumer policy and trade policy, and their link with competitiveness and productivity—a topic of growing importance in both developing and industrial economies—is analyzed in the second section. The third section considers the political economy of consumer protection. The fourth gives a detailed account of the MMM pyramid scheme and draws policy conclusions on the role of consumer protection in the former Soviet Union. Topics for future research are suggested in the fifth section, which is followed by a brief conclusion.

## **Consumer promotion**

Regulations protecting consumers work by changing the environment that consumers face. Such regulations have short-run, often immediate, effects. Consumer promotion policies, on the other hand, do not try to change the consumer's environment, but aim to modify the consumer's behavior. They include consumer education and information programs and mechanisms aimed at helping consumers achieve representation and redress, such as consumer organizations.

### *Consumer education and information*

If consumers were better informed and more educated, they would be able to solve many of the problems they

encounter. This is especially true in developing countries, where consumers often lack the abilities needed to uphold their rights. Less regulation is needed to protect more educated consumers. For this reason, consumer education is an important aspect of consumer policy. Another reason concerns consumer safety: Many injuries cannot be prevented through direct regulation. Miller and Parasuraman (1974) concluded that “the fact that at least 80 percent of the consumer product-related injuries may not be caused by defective or unsafe products suggests that consumer education has a very large untapped potential for reducing such injuries.”

Consumer education and information programs can reach consumers through a variety of media. In many countries television programs are devoted to discussion of consumer issues; in some, government consumer organizations are allocated a specific number of prime-time minutes per week.<sup>3</sup> Some private media (newspapers, radio, and television) also devote time or space to consumer issues. When the media are privately owned, however, the possibility of retaliation by companies portrayed in a negative light through the withdrawal of advertising can limit the educational value of such programs.

Product-testing magazines are one of the most important vehicles for educating consumers. Chapter 6 argues that governments in developing countries should subsidize private product-testing magazines. Examples of product-testing magazines in industrial countries include *50 millions de consommateurs* in France, *Rad and Rön* in Sweden, *Which?* in the United Kingdom, and *Consumer Reports* in the United States, published by Consumers Union. The coverage of these magazines varies considerably, from 20 percent of all households (Norway) to 3 to 4 percent (United States). Some are aimed specifically at children, for example, Consumers Union’s *Penny Power*. The potential impact of magazines targeting children is large, since consumer attitudes, habits, and skills, including the ability to evaluate advertising critically, are developed at an early age.<sup>4</sup>

While information regulation is often insufficient to protect consumers—for such an approach to work, consumers must read, understand, and act on the information provided—education campaigns are even more problematic. The standard distinction between information and education programs is that the first are considered *notification* schemes that provide factual information, and the second *persuasion* schemes that convey messages, which may or may not contain factual information and often seek to motivate the public to modify their behavior.<sup>5</sup>

Adler and Pittle (1984) concluded that “the popularity of persuasion campaigns . . . says little about their effectiveness. While we do not challenge the value of all information and education programs, we suggest their popularity rests more on philosophical and ideological grounds than on solid empirical evidence supporting their ability to alter consumer behavior.” The fact that education programs often attempt to break deeply ingrained consumer habits compounds the hurdle that such programs must overcome. When a message conflicts with a consumer’s prevailing belief, the consumer will reject or distort the message to make it palatable.

An example of the limitations of education campaigns is provided by programs in the early 1980s aimed at promoting the use of safety belts in industrial countries. According to the National Highway Traffic Safety Administration (NHTSA), in the United States in the early eighties, 34,000 people were killed each year—and more than half a million suffered moderate to severe injuries—as a result of highway accidents. If all occupants wore seat belts, motor vehicle fatalities would be cut in half and injuries reduced by 65 percent. Because automobile insurance is more expensive for all drivers because some do not use safety belts, government intervention might be considered appropriate. In addition, because most drivers believe they are better-than-average drivers, they may underestimate the probability of having a car accident and consequently view safety belts as unnecessary. The fraction of U.S. drivers using safety belts increased only marginally (from 11.3 percent to 13.9 percent) after a three-year education campaign by the NHTSA. Even in Sweden and the United Kingdom, where such campaigns were most successful, the percentage of drivers using safety belts never exceeded 35 percent. Although such campaigns are usually cost-effective, alternative mechanisms, such as complete passive protection and mandatory seatbelt laws, prove more effective.<sup>6</sup>

#### *Consumer redress*

Consumers should have access to proper redress, through swift, effective, and inexpensive procedures, for injury or damage resulting from the purchase or use of defective goods or unsatisfactory services. Yet consumers with grievances often do not gain redress, in part because the opportunity costs involved in resolving a complaint can be prohibitive.<sup>7</sup>

By giving consumers a voice, a well-functioning redress system can mobilize consumers to play an active role in the marketplace. By contrast, a poorly functioning redress system can lead consumers to exit the market-

place (see Hirschman 1970). A study of consumer complaints in a representative sample of industrial economies found that consumers were dissatisfied with one of every six products they had purchased.<sup>8</sup> About half these grievances (complaints) were voiced to the sellers, and a satisfactory result was reached in 60 percent of the cases. Of those complaints that were not solved satisfactorily, only 1 to 3 percent were brought to a third party such as a consumer association or an attorney. Only 3 to 5 percent of the complaints brought to a third party were pursued further.<sup>9</sup> Finally, only 5 to 10 percent of the legal actions that are brought are settled in court; the rest are resolved through out-of-court settlements.

Schemes designed to enhance consumer redress vary across countries. Some redress schemes depend on government initiatives. Others are sponsored by business.<sup>10</sup> Among the former two broad groups can be distinguished. A first group consists of mechanisms aimed at reducing the transactions costs of using the legal system. A second group consists of a variety of mechanisms designed to circumvent the legal system altogether—sometimes called “alternative dispute resolution mechanisms.”

*Government-sponsored consumer organizations.* The role of public consumer organizations as a consumer coordination mechanism is important in most consumer promotion policies. As Arrow (1970) pointed out, producers collude more often than consumers because they are fewer in number and stand to gain more from coordinating their actions. Lower transactions costs (in this case, bargaining costs) explain why without government action there are more producer associations than consumer associations.

Public consumer organizations can play an important role in both industrial and developing countries by reducing consumers' transactions costs in obtaining redress, gathering and disseminating product information, and representing consumers before government agencies and legislative bodies. In some countries government-sponsored consumer organizations also support networks of citizen advice groups.

Public consumer organizations may also play an important role in enforcing health and safety regulations. In practice, health and product safety regulations are often poorly designed and enforced. Both private and public consumer organizations can play an important role in monitoring regulatory agencies.<sup>11</sup> Where search and experience goods are concerned (see chapter 6), it is enough that these agencies receive and channel consumer complaints on lax enforcement.<sup>12</sup> In the case of credence goods, they must act on their own initiative. In the latter case, where the role of “whistle blowing” is most impor-

tant, the cost of monitoring regulators is also highest. An alternative approach for monitoring regulators is to create an independent “regulatory board,” with an institutional design to insulate it from political pressure.

The effectiveness of a consumer organization in promoting consumers' economic interests varies, and it often depends less on financial resources than on commitment and ingenuity. To facilitate consumers' redress, the Canadian Bureau of Consumer Affairs established a special post office box in 1968, and the French government took the same step in 1978. Consumers' letters are received at local branches of the consumer organization, where staff give information and advice and try to reconcile parties. If a problem cannot be resolved, the letter is forwarded to the appropriate government agency or private consumer organization. For such a system to work, the government organization must be efficiently designed and managed. At its peak, Canada's consumer postal box had 60 staff members who handled 56,000 complaints each year.<sup>13</sup>

*Business-sponsored redress mechanisms.* A variety of redress mechanisms are sponsored by the private sector. One means of redress available to consumers who purchase some products is a toll-free telephone number. More than 200,000 such lines currently operate in the United States, and the concept is making inroads in developing countries as well (see Agins 1990). A related mechanism is repair, replacement, and return policies, as well as warranties offered by sellers. And U.S. companies also frequently establish consumer affairs and consumer service departments to handle customer complaints.

Private arbitration schemes are also used in some countries to resolve consumer complaints. In such schemes, the disputing parties jointly select an arbitrator and agree to abide by the arbitrator's decision. The informality and speed of the procedure (which bypasses the court system) and the finality of the decision often make private arbitration an attractive, less expensive alternative to a court hearing.

Media action lines are yet another redress mechanism. Some newspapers have a consumer action columnist who receives complaints from consumers, contacts the offending business, attempts to mediate a settlement, and publishes the result in the newspaper. Consumer columns can have a powerful deterrent effect. A customer's threat of contacting the columnist may encourage a seller to settle a complaint to the consumer's satisfaction. A limitation of such columns, however, is that newspapers face a conflict of interest when complaints involve their major advertisers.<sup>14</sup> It is not surprising that the complaints pub-

lished in such columns mainly (if not exclusively) involve small businesses or government agencies.

Government agencies can generate media publicity by releasing the results of comparative testing experiments they have solicited or by publicizing the names of companies that have not met government standards. This can be an effective way of promoting consumer protection in developing countries. Simply showing a wide price dispersion for homogeneous goods may alert consumers to the value of spending more resources to search for lower-priced goods. Similarly, exposing the contents of certain food products may prompt consumers to choose more carefully among competing brands.

*Redress and the legal system.* Reducing the transactions costs involved in bringing a complaint is one way of improving consumers' access to redress, particularly in developing countries, where the legal system can be slow and cumbersome. Small claims courts, such as those in the United Kingdom, make the legal system more accessible to consumers by reducing litigation costs, expediting the procedure, and discouraging (sometimes even prohibiting) parties from retaining legal representation. Allowing the consumer choice of jurisdiction is one way to help keep transactions costs low. India's Consumer Protection Act created a system of courts to resolve consumer disputes at the district, state, and national levels (see Kayak 1987). Although these courts do not have all the features of small claims courts and the procedure consumers must follow is cumbersome, they are a step in the direction of ensuring legal redress for India's consumers.

Class action suits in the United States reduce transactions costs by providing a mechanism for consumers who have a similar complaint against a seller or manufacturer to join together in bringing a lawsuit. Class action suits were instituted to provide incentives to lawyers to take cases affecting a large, diffuse group of consumers. An alternate approach is to give consumer organizations a central role when collective action is called for, as Brazil's 1990 consumer protection law did (see Salgado and others 1994). The Brazilian law gives incentives to consumer organizations to seek redress for consumers by exempting them from judicial costs and fees.<sup>15</sup>

The high cost of legal recourse in most countries and the lengthy delay in bringing proceedings to a conclusion have given rise to myriad alternative dispute resolution mechanisms. In many countries government institutions can act as arbitrators in consumer disputes. The recommendations of such arbitrations are usually nonbinding on the parties, although the system in Norway is one exception.

By centralizing information on consumer grievances, government-sponsored consumer organizations can detect when a pattern of deception by a particular seller emerges and take action accordingly. Public complaint boards in Denmark, Finland, Norway, and Sweden provide buyers and sellers a simple, rapid, low-cost alternative to the courts to settle disputes. Although the decisions of these boards are not legally binding, compliance is achieved in 75 to 85 percent of the cases. If the losing party does not comply with the board's decision, the winning party is free to pursue its case in the court system.

Mexico's Federal Consumer Protection Act of 1975 enables aggrieved consumers to follow either of two routes (see Vargas 1989). The situation in Mexico prior to 1975 resembled that of many developing countries today. Warranties and operating instructions of most appliances were written in a foreign language (English), consumers did not know where to go or what to do if a good turned out to be defective, and the courts were unprepared to deal with even the simplest legal complaint related to consumer transactions. Today, the government's Consumer Affairs Office plays a central role in an adjudicatory system to solve consumer disputes. A relatively simple and expeditious binding arbitration procedure is also available. The first stage of the arbitration procedure is carried out in writing, which substantially reduces transactions costs.<sup>16</sup>

### **The relation between consumer protection policy and trade policy**

Trade liberalization in many developing countries has forced exporters and government officials in these countries to adhere to more stringent consumer protection policies. Quality requirements for consumer products are usually greater in industrial countries than in developing ones. Moreover, increased access to high-quality goods makes local consumers more demanding. In meeting these demands, local manufacturers enhance their competitiveness. Consumer protection policies in the exporting country are also fostered. A challenge for developing countries is to reduce the risk that local consumers will be harmed by imported products.

#### *Consumer protection as a nontariff barrier*

Industrial countries often set higher product standards than developing countries would meet in the absence of regulation. The additional costs that developing-country exporters must incur to bring their exports up to these standards make their products less competitive in global markets. For example, since May 1994 the U.S. Food and Drug Administration has required that food imports into the

United States indicate their nutritional contents as well as the processor, packer, and authorized wholesaler. Environmentally friendly ("green") products that consumers in some industrial economies are demanding are another example. The requirements for using a "green" label may increase costs dramatically for a developing-country exporter, thereby eroding its comparative advantage.

Such standards are usually legitimate. But quality standards are sometimes used as a nontariff barrier to block foreign products from industrial markets. The best-known example is the *Cassis du Dijon* case, in which the European Court of Justice ruled that a German regulation preventing the sale of French liqueur in Germany could not be justified on consumer protection grounds (see Dardis 1988). In recent years developing countries' complaints that industrial countries use consumer protection and safety issues as a nontariff barrier have increased. For example, Japan bans fruit imports from countries in which the fruit fly has been spotted, regardless of whether the location where the fruit is grown is free of the insect. Thus Chile is barred from exporting fruit to Japan because a fruit fly was found in Arica, located some 2,000 kilometers from the Chilean central valley where fruit is grown.

The practice of using consumer issues as a nontariff barrier is not limited to industrial countries. Many African countries ban imports of used clothes, giving as the reason that such garments spread acquired immune deficiency syndrome (AIDS). In Tanzania this accusation apparently was initiated by domestic textile producers (personal communication, D. Kaufmann, August 1994).

Given the higher standards prevailing in industrial economies and the fact that these standards are sometimes used as nontariff barriers, developing countries should promote standardization at both the national and international levels.<sup>17</sup> Although most developing countries have instituted standards to ensure the accuracy of weights and measures used in commerce and to protect consumers from dangerous products and substances, they only recently have become aware that standardization can facilitate exports to industrial economies. John Hinds, president of the International Standards Organization, described the situation at a World Bank seminar in November 1992:

The national standards in industrialized economies can be barriers to exporters from developing economies when they do not know, cannot attain, or cannot certify that they have attained the standards of performance or safety required in the

importing market. To compete effectively in the world market, developing countries therefore need to know the standards of the markets they will be exporting to, have internationally recognized testing and certification organizations, and a national quality assessment system conforming to international criteria.<sup>18</sup>

Standardization should go hand in hand with the promotion of the minimal harmonization principle, which the European Union announced after the *Dijon* case. This principle admonishes countries against using standards to bar goods from other countries unless consumer health or safety is at risk.

The coordination of consumer policies across countries could provide a beneficial step. Trading partners aiming at similar levels of quality and safety would benefit from coordinating their consumer policies, since doing so reduces production costs without significantly changing the level of protection provided.<sup>19</sup>

In negotiating free trade agreements and customs unions among themselves, countries would benefit from incorporating consumer policy issues into their negotiations. The European Union has acknowledged the potential advantages of doing so by creating the European Bureau of Consumer Unions to coordinate the consumer policies of the 12 member states. This organization will be challenged to propose standards that are compatible across both highly industrial economies such as Germany and France and less industrial economies such as Greece and Portugal. Countries in the first group may complain that the standards are too lax, and those in the second group may perceive even a relatively weak standard as damaging their export industry.

#### *The competitiveness link*

A popular argument to allay the private sector's concerns about the costs of promoting consumer protection is that these policies benefit producers by raising their competitiveness. Some observers have concluded that without efficient markets, rivalry, and demanding consumers at home, local industry cannot develop the products, skills, and expertise necessary to compete successfully abroad. More demanding consumers help local firms to test new products and spot opportunities for new product development.

More sophisticated consumers at home may also help potential exporters reduce the risks associated with new investments. Consider, for example, a producer in a developing country considering the export of meat. The absence of a slaughterhouse that satisfies the standards in

industrial economies is an important barrier. If local consumers were more demanding, the start-up costs to begin to export would be lower. In general, more demanding local consumers imply that there will be local demand for better-quality products, ensuring exporters a minimum return on a risky export project.

Finally, it would be beneficial if consumer organizations in developing countries stimulated consumer interest in trade matters, for example, by providing concrete examples of unexplained price and quality differences between countries for certain products. The Consumers Association in the United Kingdom goes a step further: It provides advice to consumers on how to obtain the cheaper product, for example, how to import a cheaper car than one that can be purchased in the United Kingdom (see Linke 1988).

#### *Protection of health and safety*

Protecting consumers may become more difficult in developing countries as trade increases because the flow of product information for imported goods may be obstructed.<sup>20</sup> Information may be inadequate for three reasons. First, individual consumers and even government agencies may not know how to contact the home office of the seller or manufacturer for additional information. Second, language barriers reduce the value of the information provided, especially if information is in the language of the exporter. Even if information has been translated, translations are often of poor quality. In addition, many manufacturers do not consider the fact that more than one language is spoken in some developing countries. For example, warning labels on imported pesticides are printed in only two of the 15 official languages in Kenya (see World Bank 1984). Finally, in many developing countries high levels of illiteracy create obstacles to the transfer of information, exposing the population to greater health and safety risk.

It is often more difficult for consumers to obtain redress in the case of imported goods, especially when redress must be obtained directly from the manufacturer. International redress may involve two legal jurisdictions, and international law may apply. In such cases, transactions costs become prohibitively high.

Obtaining redress may also be more difficult because certain safeguards aimed at facilitating redress apply only within the producer's country. For example, safety seal providers (such as UL in the United States, SG in Japan, and DIN in Germany) are only liable for accidents that occur within the country of issue. Developing-country consumers would benefit if safety seals with worldwide lia-

bility were developed. At the same time, however, this step would harm developing-country exporters of similar products, who would face higher liability premiums as a result.

Developing countries can adopt policies to reduce the risks associated with imported goods. First, countries can require that imported goods carry labels and other relevant product information in a form that is understandable to domestic consumers. Resources must be allocated to enforce such a policy and to check the quality of translations. Second, countries can encourage the adoption of an international code of hazards. To the extent possible, this code should be based on pictograms. The desirability of such a code is illustrated by a tragedy that occurred in Iraq in the early 1970s. Iraqi farmers used mercury-coated grain intended for seed to make bread, resulting in 6,000 reports of poisoning. Although the manufacturer had colored the grain with brownish-red dye as a warning, the dye was not permanent. Had warning labels been translated or pictograms used, farmers would have understood the dangers the seeds posed.

Third, to provide incentives for appropriate information provision, the importer should be held liable for product-related accidents. Brazil's consumer protection law includes this provision. Such a policy faces two important limitations, however. First, importers may lack the resources to pay damages. Second, in many developing countries consumer access to legal redress is almost nonexistent. The role of government agencies in providing redress is important in this case.<sup>21</sup>

Clearly, there are trade-offs to consider in deciding whether to allow imports of dangerous substances. For example, Thailand's imports of the pesticide DDT dramatically decreased the incidence of malaria and opened up vast new areas for rice and corn production, thereby increasing local food consumption, improving nutrition, and promoting agricultural exports (see Kinsey 1988). But as a World Bank (1988) report noted regarding warnings on pesticides sold in Kenya, "Even if the consumer can read and understand the warnings, it is not easy to 'avoid contaminating rivers,' nor to 'wash with soap and water after use.' Most users have never seen a physician, and certainly are not able to consult one 'immediately' as advised by the label."

Considerable attention has been paid to the export of unproven or dangerous pharmaceuticals. A double standard in some countries allows their pharmaceutical industries to export drugs whose use is banned domestically (Micklitz 1988). One solution is to require manufacturers to inform responsible authorities in importing countries of the potential danger—the "informed consent"

approach. As long as health authorities are accountable for their decisions to allow the import of pharmaceuticals that have been banned in the country of manufacture, the informed consent approach may prove more successful than other information provision regulations. The main reason why export notification has not been discussed internationally is that the World Health Organization (WHO) has advocated tougher measures to deal with the double standard problem. For this purpose, WHO has designed a certification scheme that gives quality guarantees for exported products. In February 1988 the European Community published a draft resolution on the export of pharmaceuticals. It calls for member states to directly notify WHO of all regulatory actions, voluntary suspensions, and recalls.

### **The politics of consumer protection**

Although the number of people who stand to gain from consumer protection is often significant, consumer protection has a hard time finding political support because its benefits are so diffuse. While the aggregate gains can be large, often no individual has an incentive to become a consumer advocate. Yet, producers can easily mount a force to combat consumer protection legislation.

As the history of consumerism in the United States shows, high-profile, credible, and charismatic consumer advocates can play an important role in promoting legislation favoring consumers (Creighton 1976; Mayer 1989). Ralph Nader helped promote consumer issues in the United States during the 1960s and 1970s, when social activism was high in both developing and industrial countries. Nader believed that U.S. car manufacturers could build safer cars and that the market system would function more effectively if consumer organizations provided the missing checks and balances.<sup>22</sup> Cars manufactured by U.S. auto makers are considerably safer today, thanks to safety regulations prompted by consumer activism. Nader and his Nader's Raiders, as his followers were called, also worked to ensure that the government agencies mandated to protect consumers did not become co-opted by the interests they were supposed to regulate.

During the 1960s and 1970s incentives for the emergence of consumer activists were lacking in many developing countries. Political parties favoring free markets had close ties to business, and thus had little incentive to promote consumer issues. Groups critical of the free market approach were often less interested in improving the market system than in replacing it altogether.

The experience of Chile, which led its neighbors in instituting market reforms, may be repeated in many

other Latin American countries in the near future. Chile returned to democratic rule in 1990, when a coalition of center and left parties came to office. The coalition was divided on the importance of the market mechanism in the economy—less along ideological lines than along generational lines. Younger politicians favored the free market approach, whereas older politicians preferred more government intervention. Although the market-oriented politicians prevailed, a significant proportion of the rank and file supporting the government yearned for more government intervention in the economy.

An anecdote told by the Bolivian undersecretary of commerce shows that this phenomenon is widespread. In 1993, when the governing party of Bolivia announced the establishment of consumer advice and redress offices throughout the country, the government received letters from elderly party members in the provinces, who offered their services to enforce price controls. Enforcing price controls was common work in Latin America in the 1960s, but it contradicted the essence of market reforms in Bolivia in the late 1980s.

Another illustration that governments are aware that the rank and file is still catching up on the spirit of recent reforms is implicit in a four-page pamphlet published by Chile's National Consumer Service (SERNAC). The pamphlet was designed to teach low-income housewives to compare prices, demand quality, check weights, ask for warranties, and so on.<sup>23</sup> It presents a dialogue between two housewives, one of whom argues that, since poor people consume little, they cannot benefit from consumer promotion policies.

In Chile, as in many other developing countries, no political group wants to push consumer issues as a priority. On the right of the political spectrum, close ties to business and skepticism of government regulations prevent the emergence of politicians who will fight for consumer protection. And those at the political center and left are fearful of fully embracing the market system. The first Latin American country to pass significant legislation favoring consumers was Mexico, a country in which the power of groups supporting radical alternatives to the system was never strong.

Once a consumer issue gains a place on the public agenda, however, few political groups are willing to invest political capital in opposing it. For this reason, many Latin American countries have recently enacted consumer protection laws and others are discussing their pros and cons. As market-oriented reforms begin improving consumers' well-being, politicians can be expected to embrace consumer issues.

How do consumer issues gain the attention of public policymakers? Consumer problems become consumer issues as a result of the interplay of the media, the public, and policymakers.<sup>24</sup> Policymakers sometimes pursue a consumer issue because they believe it is important, even if most consumers and the media are unaware of it. Other times policymakers react to public outcries after a disaster, as described in a film produced in 1984 by Consumers Union, *America at Risk*:

[There is] a pattern that seems to recur throughout the history of the consumer movement. First, there is a disaster, as in the case of the diseased meat and patent medicine scandals. Then there is research, investigations undertaken by scientists or by groups like the National Consumers League or by journalists to help expose the problem. Next there is a ground swell of protest by an outraged public, leading to a demand for legislation. Then a regulatory bill emerges, often so flawed by compromise that it takes another disaster and another struggle to get the bill amended.

Consumer issues emerge in a less unidirectional fashion than depicted in the Consumers Union film. In particular, there are clear limits to the ability of the press to lead public opinion on consumer issues. Their dependence on private advertising possibly explains why certain consumer issues are not covered by the media.<sup>25</sup> Much remains to be understood about how consumer issues reach, or fail to reach, policymakers' agendas.

Another question of interest is whether the problems that make it to the policy agenda are those that "should" make it. The approval process for new drugs illustrates an unfortunate asymmetry that may prevent policymakers from acting in the public interest. On the one hand is the risk of approving a drug that has dramatic side effects, such as the drug thalidomide. On the other hand are the costs of erring on the side of excessive caution by not approving drugs that could have benefited many people. In the second case, the threat of media exposure is smaller, since it is impossible to interview patients who died because a drug was not approved in time.<sup>26</sup> This asymmetry is likely to lead policymakers to be more cautious than is socially desirable when allowing new drugs to be marketed.

As developing countries consider consumer protection issues, they would do well to learn from other countries' mistakes. One such mistake was made in the United States in the mid-1980s. After scientists concluded that asbestos exposure contributed to thousands of cancer

deaths each year, the public outcry led the U.S. Congress to pass a law in 1985 requiring city and state governments to remove asbestos from public buildings at a total cost of \$15 to \$20 billion.<sup>27</sup> The Environmental Protection Agency concluded in 1990 that by releasing asbestos particles into the air, removal efforts had actually increased the health risk.<sup>28</sup> This example illustrates the danger of an overly reactive policy agenda.

### **Consumer protection in transition economies**

Because most people in transition economies lack the basic skills needed to operate in a market economy, they are particularly vulnerable to fraudulent schemes. This section describes one such scheme, the MMM financial pyramid in Russia, and draws some conclusions for consumer protection policy.<sup>29</sup>

#### *MMM and its millions of shareholders*

Millions of Russians believed that MMM, a little-known investment company that offered a 3,000 percent return on investment, could make their dreams come true. In early 1994, within only a few months, MMM sold millions of shares through 136 offices in 50 Russian cities (60 offices in Moscow alone). The price of MMM stock skyrocketed from 1,600 rubles (US\$1) in February to 115,000 rubles (US\$55) in late July.

The MMM investment fund was a classic financial pyramid scheme. Initially, MMM was able to sell a rapidly increasing number of shares and use part of the proceeds to buy back shares at a much higher price. Those who sold their MMM shares during this period received handsome returns, and these success stories attracted wide media attention. There was no warning that, as in any pyramid scheme, the throng of latecomer shareholders would be ruined when the pyramid collapsed.

On July 26 the Russian government disclosed that MMM's president, Sergei Mavrodi, was suspected of having violated tax laws. At month's end shares were quoted at only 1,000 rubles. All but one percent of the stock's value disappeared in a few days, and all but the Moscow office of MMM closed. By early August Mavrodi was in jail following a dramatic raid on his apartment by the tax police. Shortly thereafter, hundreds of thousands of panicking shareholders assembled in front of MMM offices in a frantic attempt to sell their shares. On August 4, MMM announced the temporary closing of all its offices.

The anticipated next chapter in the unfolding drama would seem to be that shareholders, realizing that they had fallen for a scam, would turn against MMM and Mavrodi, seeking restitution from a variety of legal and

other channels. What happened instead illustrates the different perspective required in addressing consumer protection issues in transition economies.

Mavrodi went on the offensive, accusing authorities of wanting to destroy millions of small capitalist shareholders. (Mavrodi claimed 10 million shareholders; outside estimates set the number at 2 to 5 million.) On July 29 he announced publicly, "We have been stopped on the eve of a super breakthrough, after which Russia could have become the richest country in the world, and Russians—MMM shareholders—wealthy people." Mavrodi succeeded in convincing MMM shareholders that the government was to blame, and a drive began to collect the requisite one million signatures to call a referendum to unseat the Russian president.

On August 22, MMM offices reopened and trading resumed. Incredibly, tens of thousands came not to sell their shares but to buy more. To bypass the legal requirement to register new shares, and to deceive consumers once again, MMM sold "tickets" rather than shares. The tickets were each sold for 1,515 rubles, on a vague promise that MMM would exchange 100 tickets for one share at an unspecified future date. In addition to trusting that this exchange would take place, ticket buyers were gambling that MMM shares would rise by a factor of more than 100, to a value exceeding 151,500 rubles—almost 40 percent higher than the peak value before the collapse.

Mavrodi's success in turning many shareholders against the state put the government on the defensive. Faced with shareholders' growing demands for restitution, Prime Minister Viktor Chernomydrin stated on July 30 that there would be no compensation to shareholders because any compensation would be at the expense of Russians who had not purchased MMM shares. Chernomydrin also acknowledged that the government had failed to enact appropriate legislation to prevent the fraud and placed the blame on the Ministry of Finance.

The ministry quickly drafted legislation to give the tax police regulatory control over all stock market transactions and allow the ministry to discriminate against foreign investment banks and institutions. Russian stockbrokers organized to block the legislation, which they feared would jeopardize market development. And the state privatization committee declared the proposed legislation an effort to turn off the taps of commerce and control the financial markets.

Mavrodi was elected to a vacant seat in the lower house of Parliament in a special election in October 1994. He thus gained immunity from prosecution on tax evasion charges. Although he had promised during his cam-

paign that he would save MMM if elected, he announced to the 3,000 supporters who gathered to celebrate his electoral victory the suspension of MMM shares. Shareholders thus had no hope of selling their securities.

*Reasons for MMM's success.* Today, several years after the dissolution of the Soviet Union, the state's desire and ability to protect individual economic rights and consumer rights is still viewed with suspicion.<sup>30</sup> Mavrodi cleverly capitalized on both this suspicion and the government's role as the social guarantor when he accused authorities of denying the Russian people the opportunity to become rich.<sup>31</sup>

The credibility gap faced by governments in transition economies puts them in a particularly difficult dilemma. On the one hand, governments must tread more carefully into consumer regulation than those in industrial countries. On the other hand, the state is still expected to come to the rescue of individuals in the name of social protection—even dishonest individuals like Mavrodi.

Another reason Mavrodi was able to bilk the Russian people is their market and economic illiteracy.<sup>32</sup> Although some MMM shareholders understood that they were participating in a lottery, and that winning or losing depended on their sense of timing, the vast majority simply did not question whether there was a risk. Because MMM was the first pyramid scheme in Russia, Mavrodi, a brilliant mathematician and early learner of financial instruments, had more (and more accurate) information than buyers. Mavrodi knew that consumers would discover the quality of the good only when it was too late.<sup>33</sup>

The worst threat Mavrodi perceived was limited bankruptcy (and perhaps a few weeks in jail), while he could still amass (and possibly hide) a small fortune. His potential gain far exceeded the risk. He may have even bet that the government would be forced to give the firm and its shareholders special treatment once the pyramid's fragility was exposed.<sup>34</sup>

Mavrodi was also successful in reducing cognitive dissonance on a massive scale, that is, making shareholders disregard any negative views about MMM.<sup>35</sup> Shareholders wanted to believe that they were buying a ticket to wealth (see Akerlof and Dickens 1982). The belief was systematically nourished, night after night, by persuasive television ads. Once the pyramid began to crumble, shareholders quickly bought the argument that state intervention was to blame. By believing Mavrodi, they justified their decision to buy shares in the first place. Cognitive dissonance and consumer irrationality conspired in late August, when MMM tickets went on sale. One Russian, oblivious to the distinction between a share and a ticket, said while queu-

ing in the rain, "I believe in MMM and will buy five more shares. The instability is only temporary" (*International Herald Tribune*, August 23, 1994).<sup>36</sup>

The combination of a credibility gap vis-à-vis the state, lack of consumer and economic education, and massive cognitive dissonance among an impoverished population dreaming of getting rich quickly only partly explains the success of the MMM scheme. Because Russia's market economy developed faster than the state's regulatory infrastructure, consumers were left unprotected and fell victim to Mavrodi's scheme.

*The case against the "rational" incentive argument.* One might argue that it was in the rational interests of shareholders to ally themselves with Mavrodi and demand restitution from the government, since their chances of recouping their investment from the second were greater than from the first once the pyramid collapsed. A rational incentive argument does not hold for two reasons. First, although shareholders might have been rational in turning against the government, doing so did not require them to ally themselves with Mavrodi. In fact such an alliance might have weakened government sympathy for shareholders.

Second, while Mavrodi was forging an alliance with shareholders as MMM was collapsing, the government was issuing warnings about MMM as well as a less-known but similar pyramid scheme, RDS. RDS executives were much more passive than Mavrodi had been in marketing MMM shares, and the RDS shareholders turned against the company. Mavrodi's personality seems to have made the difference.

#### *Policy implications*

The magnitude of the MMM scandal underlines the importance of putting in place an appropriate regulatory system. The challenge for the countries of the former Soviet Union is to do so without hampering their transition to a market economy. As they work to establish effective regulatory systems, countries must guard against overregulation. Clear guidelines on entry to the financial sector may well be all that is needed to safeguard consumers against future MMM-like schemes. Such guidelines would discourage unscrupulous operators and force legitimate entrepreneurs to provide full and complete information to the public regarding their products. Overzealous regulation, such as outright bans, may choke off bona fide operations. Pyramid schemes might be allowed so long as the public is aware that buying a share in such schemes carries significant risk. Furthermore, there is the clear need to enact and enforce advertising laws. Such laws would

require a scheme such as MMM to note in its ads that it is not an investment firm and that significant risk is involved.

The more general case of Ukraine's financial sector is also illustrative in this context. About 200 licensed banks operate in Ukraine, ranging from solvent Western-style banks to "treasury arms" of state enterprises to fly-by-night risk venture "capitalists." The central bank should clearly separate "real banks" from other financial institutions. The selected few core banks would have a specific set of duties and be subject to prudential regulation. They would receive a well-advertised seal of approval from central bank authorities. All others would be labeled as more risky financial institutions.

In addition, the creation of nongovernmental consumer advocacy organizations should be encouraged, since consumer groups can be very effective in increasing the responsiveness of the legal system to consumer problems. Consumer unions are becoming increasingly common in the countries of the former Soviet Union. In late 1993 the Consumer Society of Ekaterinburg sued Aeroflot after passengers encountered a 21-hour delay in a freezing Moscow departure lounge. Although no compensation was forthcoming as a result of the lawsuit, the mere fact that Aeroflot could be taken to court by consumers must be regarded as a landmark event.

#### **Topics for future research**

The development of effective consumer promotion programs and policies would be facilitated by research in several areas. First, studies of consumer complaints would assist governments in designing effective redress mechanisms. Such studies might address the following questions: What fraction of consumers express discontent with the products they purchase? What fraction of dissatisfied consumers complain to the seller? What fraction of complaints are resolved satisfactorily at this stage? How successful are consumers who take their cases to court or to an alternative dispute resolution mechanism? Research comparing the effectiveness of various legal redress mechanisms in promoting consumer protection is also needed.

A study of how consumer issues emerge in developing countries would also be useful. It might address the following questions: What is the role of the media in promoting consumer issues? To what extent are consumer issues promoted by policymakers? How do consumer advocacy groups emerge?

Most developing countries do not have a tradition of evaluating public policies and debating the resulting conclusions. When various policies are considered to address

consumer issues, there is rarely even a preliminary quantification of costs and benefits of the alternatives. Available studies have been often sponsored by institutions with a vested interest in the issue at stake.

It is important to educate the public on the fact that consumer protection policies involve costs and benefits, and that their effectiveness should be ascertained. Just as indices have been developed to measure the effectiveness of antitrust policy, for example, it would be useful to develop indices that quantify the effectiveness of consumer protection policies in selected countries.

In addition, a study of public and private consumer organizations is needed to answer the following questions: What are the critical elements for a successful organization? What organizational structure is most effective? What relationship should such organizations forge with government agencies and business organizations?

A final research question is, at what stage during the transition to a market system should consumer protection become a priority? Documenting the emergence of consumer issues and groups during the reform in the countries of the former Soviet Union would be instructive.

## Conclusions

The degree to which consumers are protected depends on the combined effect of policies and institutions. A given degree of protection can be achieved at similar costs with very different combinations of policy instruments. It follows that a holistic approach is called for when designing consumer policies.<sup>37</sup>

In most developing countries, consumers' access to redress is severely limited. Not only is obtaining redress through the legal system expensive; the system often functions poorly or is even corrupt. Any strategy aimed at improving consumers' access to redress should be three-pronged: First, it should include vehicles to facilitate consumers' access to legal redress, such as small claims courts, and mechanisms for collective action, such as class action suits; second, it should include alternative dispute resolution systems, such as arbitration by public consumer organizations; and third, it should seek to increase consumer understanding of available redress by promoting consumer education in primary and secondary schools.

Laws that make using seatbelts mandatory, require cooling-off periods for door-to-door sales, and mandate health warnings on cigarette packages all would be unnecessary if consumers were "rational." Policies that rely on consumer rationality may be expected to be less effective in practice than those that begin with the understanding that consumers often do not act rationally.

In an open economy, producers also stand to gain from strong consumer protection policies. As consumer protection develops, consumers become more assertive in voicing complaints about low-quality goods and services.

Consumer complaints are a source of information that businesses can use as a valuable competitive tool. More demanding consumers provide a market where potential exports to industrial economies can be tested, thereby guaranteeing a minimum return on new, risky investments.

Consumer organizations can play an important role in developing and industrial countries alike. They can educate consumers about products in the marketplace, monitor the safety and effectiveness of products, act as arbitrators, provide quality certification, conduct product testing, and represent consumers before government agencies and legislative bodies. Consumer education, in particular, plays a central role in any strategy to protect consumers. Knowing how markets work, how to use resources efficiently, and how to obtain redress helps citizens achieve higher living standards and helps markets function more efficiently. If a consumer organization is created as a government agency, as in most developing countries, it should be designed to foster rather than hinder the emergence of private alternatives.

## Notes

The section of this chapter on transition economies is an abridged version of a forthcoming article by Daniel Kaufmann, the World Bank's chief of mission in Ukraine. The author is grateful for this contribution. The author also thanks Peter Diamond, Mark Dutz, Tim Ennis, Claudio Frischtak, Jorge Quiroz, Ennio Stacchetti, and in particular, Richard Zeckhauser for their helpful comments on this chapter. The outstanding research assistance of Alexis Camhi and Alejandro Micco also deserves special note.

1. Of course, information on health and safety is beneficial to consumers in both rural and urban societies.
2. Some tentative explanations of why this has not happened in most industrializing countries is offered later.
3. For example, the French government allocates 24 minutes of prime-time television a week to its consumer organization.
4. See Baecher 1988 for details.
5. The following discussion is based on Adler and Pittle 1984.
6. Australia was the first country to enact a mandatory seatbelt law in 1972. New Zealand, France, Puerto Rico, Sweden, Belgium, the Netherlands, Israel, Finland, and Norway followed in less than four years. See Mayer 1989.
7. The tension between fairness and cost-effectiveness is present here. Insuring consumers against the entire set of risks they face achieves the purpose of compensation in an efficient manner but

breaks the link between compensation and seller responsibility. Ultimately, it must be recognized that one of the objectives of consumer policy, especially in developing countries that have recently undergone market-oriented reforms, is that citizens perceive the market system as fair; direct redress mechanisms are more effective in achieving this objective than is insurance.

8. The following countries were included in the study: Canada, Denmark, England, France, Germany, Netherlands, Norway, and Sweden. See Pirie 1987, Goldberg, Green, and Sander 1985, and Graver 1987. A study by Vargas (1989) in Tijuana, Mexico, had similar results.

9. Vidmar (1988) found that the likelihood that a consumer will pursue a problem grows with the monetary value involved. This does not contradict another finding by the same author that since producers also may be expected to spend more resources on winning their case as the monetary value involved grows, the likelihood that a consumer obtains redress decreases with the monetary value involved.

10. The discussion that follows draws on Forte 1991 and Graver 1987.

11. Although private organizations are better suited for this task, they often do not exist in developing countries.

12. For a discussion of search and experience goods, see chapter 6 of this volume.

13. Growing provincial involvement in the field of consumer affairs contributed to the program's demise in the early 1980s.

14. In the 1930s, fearful that it would offend its advertisers, the *New York Times* refused to carry the Consumer Union's paid advertisements. See Morse 1981.

15. Of course Brazilian taxpayers ultimately bear these costs. But given the public good nature of the collective action being induced, this may be precisely what is called for. The Brazilian law also allows the government to bring a case to court when collective or diffuse interests are at stake.

16. It is remarkable to find common law principles in a country with a legal tradition rooted in civil law principles. This finding may be an indication that common law is better suited to handling consumer grievances than civil law.

17. The General Agreement on Tariffs and Trade (GATT) approved an agreement advocating international standardization as part of the Tokyo Round.

18. See International Standards Organization, *ISO Bulletin*, February 1993.

19. The latter no longer holds when countries differ significantly in their quality and safety requirements, which often is the situation for countries with large income differences. In this case increasing the standards in the less-developed country makes local goods better but more expensive.

20. The discussion that follows draws on Reich 1988.

21. Even if government agencies lack the administrative skills to protect consumers, this does not necessarily justify banning imports of certain dangerous goods. Such bans are justified only if market

forces provide more incentives for local producers to consider safety issues than are provided to foreign producers.

22. The market failure underlying manufacturers' lack of interest in providing safety was market power.

23. See "Cualquiera... No da lo mismo!," SERNAC 1992.

24. What follows is based on Mayer 1991, which applies issue emergence analysis to consumer issues.

25. See Rowse 1967.

26. In recent years certain groups of patients, such as those with HIV-AIDS, have organized to lower the requirements for the release of new drugs. The high costs of organizing such a lobby justify describing the situation as asymmetric, even in this case.

27. See Breyer 1993 and the series of articles published in the *New York Times* between March 21, 1993, and March 26, 1993.

28. In 1991 asbestos-related cases represented more than half of all cases before U.S. federal courts (see Viscusi 1991, p. 7), contributing to the near-bankruptcy of Lloyds of London.

29. This section relies on *the Economist*, *the Financial Times*, and *the International Herald Tribune* (various issues, July 25–August 23, 1994). Conversations with Charles Blitzer of the World Bank's Moscow office and Chrystia Freeland of the *Financial Times* are also gratefully acknowledged.

30. One of the most experienced foreign journalists on the scene, John Lloyd, began an article in the *Financial Times* (July 26, 1994) by stating, "The threatened collapse [of MMM] is pitting the newly awakened forces of consumerism against the widely distrusted Russian authorities. . . ."

31. *The Economist* (July 30, 1994) wrote that Mavrodi was trying to blackmail the government into bailing out shareholders. In a letter to the tax agency, Mavrodi warned that his paying the 50 billion rubles in unpaid taxes and fines that the agency claimed he owed would destroy MMM: "I will not forecast what shape the anger of the robbed people will take: a revolution, civil war or something else."

32. During the Soviet era there was an informal, illicit economy in which many ingenious transactions took place. Those "entrepreneurs" have had a head start in the official transition to a market economy and have rapidly amassed wealth through their relative monopoly on "market education."

33. See chapter 6 for a discussion of credence goods, which consumers must buy on pure faith.

34. There existed a literal as well as a figurative moral hazard problem: The extent to which the costs of Mavrodi's recklessness would be borne by others became larger the more widespread the deceit, since this increased his perceived blackmailing power.

35. See chapter 6 for a discussion of cognitive dissonance.

36. Also suggestive of the existence and extent of cognitive dissonance are the disparate beliefs of shareholders and nonshareholders. Whereas shareholders by and large blamed the government for what happened, a poll of Muscovites by the Institute of Sociology revealed that 25 percent blamed MMM, 21 percent blamed the

“gullibility and naiveté” of shareholders, and only 13 percent blamed the government and the taxation service.

37. The Japanese consumer policy framework is famous for following such an approach. See McGregor 1991.

## References

- Adler, R. S., and R. D. Pittle. 1984. “Cajolery or Command: Are Education Campaigns an Adequate Substitute for Regulation?” *Yale Journal of Regulation* 1:59–93.
- Agins, T. 1990. “Customer Service and Challenge for the ‘90s.” *Wall Street Journal*, November 20.
- Akerlof, G. A., and W. T. Dickens. 1982. “The Economic Consequences of Cognitive Dissonance.” *American Economic Review* 72: 307–19.
- Arrow, K. J. 1970. “Political and Economic Evaluation of Social Effects and Externalities.” In J. Margolis, ed., *The Analysis of Public Output*. New York: National Bureau of Economic Research.
- Baecher, C. 1988. “The Role of Consumers Union.” In E. Scott Maynes, ed., *The Frontier of Research in the Consumer Interest*. Columbia, Mo.: American Council of Consumer Interests.
- Blanlot, V. 1993. “Regulation of the Electric Sector” (in Spanish). In O. Muñoz, ed., *Después de las Privatizaciones: Hacia el Estado Regulador*. Santiago: CIEPLAN.
- Breyer, S. 1993. *Breaking the Vicious Circle: Toward Effective Risk Reduction*. Cambridge, Mass.: Harvard University Press.
- Creighton, L. B. 1976. *Pretenders to the Throne*. Lexington, Mass.: Lexington Books.
- Dardis, R. 1988. “International Trade: The Consumer’s Stake.” In E. Scott Maynes, ed., *The Frontier of Research in the Consumer Interest*. Columbia, Mo.: American Council of Consumer Interests.
- Forte, A. 1991. “Business and Government Redress Mechanisms.” Working Paper. Consumer Policy Framework Secretariat, Canada.
- Goldberg, E. D., F. Green, and E. A. Sander. 1985. *Dispute Resolution*. Boston: Little, Brown and Co.
- Graver, K. 1987. *Consumer Redress Systems in Seven European Countries*. The Hague: SWOKA.
- Green, D. H. 1988. “The Role of Secondary Schools.” In E. Scott Maynes, ed., *The Frontier of Research in the Consumer Interest*. Columbia, Mo.: American Council of Consumer Interests.
- Hirschman, A. O. 1970. *Exit, Voice, and Loyalty*. Cambridge, Mass.: Harvard University Press.
- Kaufmann, D. Forthcoming. “The MMM Financial Pyramid.” in *Políticas Públicas Latinoamericanas*.
- Kayak, R. K. 1987. “Consumer Protection Act, 1986: Law and Policy in India.” *Journal of Consumer Policy* 10: 417–23.
- Kinsey, J. 1988. “International Trade and Trade-Offs for Third World Consumers: A Matter of Entitlements.” In E. Scott Maynes, ed., *The Frontier of Research in the Consumer Interest*. Columbia, Mo.: American Council of Consumer Interests.
- Krugman, P. 1994. *Peddling Prosperity*. New York: W. W. Norton.
- Linke, E. 1988. “International Trade and the Consumer: Report on an OECD Symposium.” In E. Scott Maynes, ed., *The Frontier of Research in the Consumer Interest*. Columbia, Mo.: American Council of Consumer Interests.
- Mayer, R. N. 1989. *The Consumer Movement: Guardians of the Marketplace*. Boston: Twayne Publishers.
- . 1991. “Gone Yesterday, Here Today: Consumer Issues in the Agenda-Setting Process.” *Journal of Social Issues* 47: 21–39.
- McGregor, S.L.T. 1991. “International Consumer Policy.” Mount Allison University.
- Micklitz, H. W. 1988. “EC Regulation of the Export of Dangerous Pharmaceuticals to Third World Countries: Some Prospects.” *Journal of Consumer Policy* 11: 29–53.
- Miller, J., and A. Parasuraman. 1974. “Advising Consumers on Safer Product Use: The Information Role of the New Consumer Product Safety Commission.” *American Marketing Association Proceedings* 36: 372–76.
- Morse, R. D. 1981. “The Consumer Movement: A Middle-Class Movement.” In C. B. Meeks, ed., *Proceedings of the 27th Annual Conference*. Columbia, Mo.: American Council of Consumer Interests.
- Pirie, A. J. 1987. “Dispute Resolution in Canada: Present State, Future Direction.” Consultants report to the Law Reform Commission of Canada, Victoria, B. C.
- Reich, M. R. 1988. “International Trade and Trade-Offs for Third World Consumers.” In E. Scott Maynes, ed., *Proceedings of the International Conference in the Consumer Interest*. Columbia, Mo.: American Council of Consumer Interests.
- Robinson, J. R. 1988. “The Content of a College-University Course in Consumer Education.” In E. Scott Maynes, ed., *Proceedings of the International Conference in the Consumer Interest*. Columbia, Mo.: American Council of Consumer Interests.
- Rowse, A.E. 1967. “Consumer News: A Mixed Report.” *Columbia Journalism Review* 6:27-33.
- Salgado, L. H., J. A. Kelly, O. A. Martínez, and G. Pais. 1994. “Competition and Consumer Protection Policy in Mercosur” (in Portuguese). Inter-American Development Bank, Washington, D.C.
- Vargas, J. A. 1989. “An Overview of Consumer Transactions Law in Mexico: Substantive and Procedural Aspects.” *New York Law School Journal of International and Comparative Law* 10: 345–82.
- Vidmar, N. 1988. “Seeking Justice: An Empirical Map of Consumer Problems and Consumer Responses in Canada.” *Osgoode Hall Law Journal* 26.
- Viscusi, W. K. 1991. *Reforming Products Liability*. Cambridge, Mass.: Harvard University Press.
- World Bank. 1984. *World Bank Tobacco Financing: The Environmental/Health Case: Background for Policy Formulation*. W0020/0087W/C2404 Washington, D.C.: World Bank Office of Environmental and Scientific Affairs, Projects Policy Department.

# Bankruptcy policy

Izak Atiyas

This chapter examines the role of bankruptcy policies in market economies. Two somewhat different perspectives on the proper role of bankruptcy are presented. The first emphasizes the role of bankruptcy as a means to enforce debt contracts. Bankruptcy policy is primarily seen as a set of rules and institutions designed to address situations in which a debtor fails to meet its contractual obligations to creditors. Bankruptcy policy supplements other debt collection rules, addressing specific problems, or market failures, that nonbankruptcy debt collection rules fail to address. I call this the debt collection view of bankruptcy.

The second perspective focuses on bankruptcy's role as a mechanism of restructuring and exit. More generally, bankruptcy procedures can be construed as a component of policies for industrial restructuring. Many industrial and developing countries have specific policies designed to address the special problems of declining industries and to reduce barriers to factor mobility, capacity reduction, and exit.<sup>1</sup> These policies deal with firms in need of restructuring. Such firms often are overindebted, because of financial policies that encourage excessive debt accumulation, earnings shocks that generate losses and reduce equity capital, or simply bad financial management. A reorganization of firms' liabilities, or more generally, a change in their ownership structure, may improve their performance. Typically, such reorganization entails the exchange of debt for equity, the extension of maturity, and reductions in principal and interest. Improving efficiency and profitability may also require asset restructuring: Companies may need to divest their unproductive units, eliminate unprofitable product lines, introduce new managerial practices, change their marketing orientations, and adopt more appropriate production technologies. Whenever a company is likely to regain viability through a restructuring of its assets and liabilities, the presence of a legal framework may facilitate renegotiations between the company's claimholders and

increase the chances of an efficient restructuring. Bankruptcy laws may provide such a legal framework.

If a company cannot regain profitability even under the most efficient restructuring scheme, economic logic dictates that it should go out of business. Again, bankruptcy provides a mechanism through which such winding up can take place. Through liquidations, bankruptcy law is believed to enhance economic efficiency by allowing the timely exit of unproductive economic units and by promoting the transfer of the ownership of productive assets to entrepreneurs or managers who can make better use of them. By facilitating the exit of inefficient firms, and thereby reducing excess capacity, bankruptcy policies also help eliminate an important potential barrier to entry.

In practice, most bankruptcy laws prescribe variants of two procedures. The first is bankruptcy liquidation, in which the debtor's assets are sold and the proceeds divided between creditors according to some priority rules determined by law. The legislation in some countries also allows for bankruptcy reorganization, a process under court supervision in which the claimholders of the debtor firm negotiate on whether and how to restructure the debtor's liabilities and assets, possibly with the objective of maintaining the company as a going concern.

The task of bankruptcy law is to formulate actual rules and procedures to be followed in practice. Bankruptcy legislation and policy both exhibit considerable variation across countries. To provide a common framework of analysis, the first section of this chapter discusses why bankruptcy policy is necessary in the first place, from both a debt collection and a restructuring perspective, and derives the objectives of bankruptcy policy. It is argued that, in principle, these two perspectives are not necessarily in conflict with each other. Indeed, in a relatively flawless world, procedures that promote the most efficient restructuring of the debtor would also best serve the creditors' interests by maximizing repayments to the claims on the company—whether immediately (in the

case of liquidation) or after the company, and claims attached to it, are reorganized. However, frictions do exist in the real world: Information is imperfect and often asymmetric, bargaining is often costly, and writing and enforcing detailed contracts are expensive if not impossible tasks. These imperfections, combined with conflicts of interest between the different types of claimholders, encourage uncooperative strategic behavior, with unintended and suboptimal outcomes. These problems create a tension between the debt collection and restructuring roles of bankruptcy policy. Unavoidably, actual bankruptcy policies try to strike a balance between these two roles.

The best way to illustrate these problems, and discuss possible solutions, is to examine the actual experience with bankruptcy laws and identify how different rules affect the strategic behavior of the different parties. The second section reviews the bankruptcy law and practice in a few industrial countries and presents the available empirical evidence on outcomes to which they lead. The section focuses on bankruptcy policy in the United States, the United Kingdom, and France, three fairly different approaches. This section also discusses the role of out-of-court workouts and their relation to formal bankruptcy. It argues that none of the three cases provides a model that satisfies the debt collection aspects of bankruptcy without sacrificing efficiency in restructuring. The section concludes with a discussion of the components of a possible improved model.

The third section concentrates on bankruptcy policy in developing countries. Shifting the analysis to developing countries requires an expansion of the chapter's focus to include additional problems such as outdated legislation, inadequate skills, lack of processing capacity in the court system, inadequate supervision and regulation of the banking system, and constraints that arise from the regulatory environment. It is argued that bankruptcy reform in most developing countries would be ineffective unless undertaken as part of comprehensive regulatory reform.

Before moving to the main discussion, it is useful to make a few points about the chapter's limitations and focus. First, even though the economic literature on bankruptcy dates back to the late 1970s, a systematic theoretical and empirical analysis of bankruptcy models is quite recent compared to other areas of economic policy. Comparative, cross-country analysis is quite limited, and most efforts have concentrated on the problems of bankruptcy in the United States.<sup>2</sup> Information on bankruptcy in industrializing countries is even more scarce. The chapter unavoidably reflects these limitations.

Second, the reader will notice that the bankruptcy procedures of industrial countries are discussed in greater detail than those of industrializing countries. The practical reason for this focus lies in the fact that both the analytical and empirical literature on bankruptcy in industrial countries is more advanced, allowing a more detailed analysis of the implications of different bankruptcy rules. Nevertheless, lessons drawn from the theory and practice of bankruptcy in industrial countries are relevant for industrializing countries that are ready to embark on reform, if only as a caution against making similar mistakes. This is especially so since industrial countries provide some of the basic models of bankruptcy, which industrializing countries often adopt in modified form.

Third, the chapter does not address the problems of bankruptcy policy in formerly socialist economies. These countries face qualitatively different problems associated with bankruptcy and restructuring, including complications that arise because of ambiguities in and the transitional nature of ownership rights and the massive need for industrial restructuring (see Atiyas 1994; Mitchell 1990, 1993; and van Wijnbergen 1992).

Finally, bankruptcy itself is a topic with many elements. The chapter focuses mainly on corporate, rather than personal, insolvencies. In addition, the coverage of the chapter is also influenced by a presumption that a need for industrial restructuring represents an important policy agenda in most developing countries. A concern as to whether bankruptcy can play a positive role in restructuring without jeopardizing the interests of creditors (or the enforceability of debt contracts) is implicit throughout the chapter. Because of this concern, a more detailed and comprehensive treatment of bankruptcy reorganizations relative to liquidations is provided.

### **The role of bankruptcy policy**

This section summarizes two perspectives on the fundamental role of bankruptcy policy in market economies. The first emphasizes the role of bankruptcy in resolving collective action problems in debt collection. The second focuses on collective action problems in debt recontracting and restructuring. The implications of these two perspectives for the objective of bankruptcy policy are then discussed.

#### *A debt collection perspective*

Credit is one of the basic pillars of modern market economies. It is a mechanism through which surplus funds can be allocated to agents that need additional financial resources to realize their optimal consumption

or investment plans. As with other types of contracts, the widespread use of credit contracts is predicated on their enforcement by the state. When a loan contract is breached, the creditor must be assured of recourse to legal remedies.

Outside bankruptcy, these remedies are enumerated in debt collection laws.<sup>3</sup> These laws both prescribe a procedure whereby creditors can enforce their claims against a debtor and define the boundaries of creditors' rights. In terms of procedure, unless a creditor is secured by collateral that supports the loan, the rules of debt collection generally require the creditor to sue the debtor. If the creditor prevails in the lawsuit, the creditor may enforce the claim by foreclosing on real property or by physically seizing personal property. In some cases, the claim may also be satisfied by requiring a third party on which the debtor has a claim to make payments directly to the creditor; in the United States, for example, creditors may be able to seize part of a debtor's wages. In terms of creditors' rights, debt collection laws delineate which of a debtor's property or income can be acquired by creditors. For example, in the United States, the rules in most states limit the extent to which creditors can lay claims on a debtor's wages or the types of assets that creditors can seize. Laws often put "tools of trade" out of the reach of creditors as well.

At the same time, debt collection laws establish a priority ordering among different claims. In general, this priority ordering is formed on a first-come, first-served basis: A creditor that is first to acquire an interest in a particular asset generally has a right to be paid first out of that asset. For that reason, creditor remedies outside bankruptcy have been characterized as a kind of "grab law" (Jackson 1986). Of course, a creditor may also gain priority over an asset if the creditor and debtor agree in a debt contract that the creditor will acquire that asset in the event of a default. Securing a loan by a collateral, then, is tantamount to securing a higher place in the line of claimants to be paid by that collateral in the event of a default.

When a debtor has sufficient assets to pay all its creditors, debt collection laws provide an efficient mechanism to satisfy claims. They cease to be efficient, however, when the value of the debtor's assets is less than the face value of the creditors' claims. Given the first-come, first-served nature of debt collection, every creditor has an incentive to grab a piece of those assets before another creditor does so. In the absence of a mechanism whereby each creditor could be credibly committed to do otherwise, asset grabbing through individual remedies is the most likely outcome of a default.

There are several reasons why this "run to the courthouse" may yield inefficient outcomes, especially when the number of creditors is large. The most important problem has to do with the fact that asset grabbing by individual creditors results in the fragmentation of these assets or in their sale in a piecemeal fashion. If the assets are sold individually, their total value may be lower than if they had been valued collectively or in bundles. As a result, the total value available to the creditors as a group may be reduced,<sup>4</sup> a situation that has been called the common pool problem.<sup>5</sup> The individual remedies prescribed in debt collection laws fail to resolve the common pool problem.

However, the existence of a common pool problem by itself does not obviate the necessity of a formal bankruptcy procedure established and implemented by the state. After all, the disposition of the assets of the debtor in the event of a default might have been specified *ex ante* in the debt contract. However, specifying all the division rules for all possible future contingencies would be extremely difficult and costly. It would require, for example, foresight regarding every possible combination of different types of liabilities that the debtor might contract in the future. The combination of common pool problems and costly (incomplete) contracting therefore generates a useful economic role for bankruptcy. Bankruptcy resolves the common pool problem by preventing asset grabbing, binding the creditors to a collective mechanism of debt collection, and allowing for an orderly disposal of assets to repay creditors' claims. Some scholars, representing what may be called the "minimalist view" of bankruptcy, have argued that this should be the *only* principle guiding the design of bankruptcy law (see, for example, Jackson 1986).

An important implication of this view is that, in principle, bankruptcy law should respect prebankruptcy claims. This does not mean that bankruptcy should not impose any restrictions on these claims. Given that bankruptcy is a collective mechanism, and that it should prevent asset grabbing, some restrictions on individual claims are unavoidable. Rather, what is meant by "respect for prebankruptcy claims" is summarized in the concept of a "creditors' bargain": The bankruptcy system "should 'mirror' the agreement one would expect the creditors to form among themselves were they able to negotiate such an agreement from an *ex-ante* position" (Jackson 1982, p. 860; see also Baird 1986).<sup>6</sup> That is, bankruptcy rules that do not violate the creditors' bargain should not create new claims or change the priority ordering of existing claims.

The logic behind the creditors' bargain is that not respecting prebankruptcy claims would allow stakeholders favored by the bankruptcy rules to transfer wealth from those that are disadvantaged, thereby distorting the original intent of bankruptcy (see Jackson 1986). However, the creditors' bargain can also be interpreted from the perspective of the development of the financial sector. Seen in this light, the bargain is a means to maintain creditors' confidence in debt instruments, and it therefore enhances intermediation.

The creditors' bargain would seem to provide a simple principle to guide the design of bankruptcy policy. For example, the company's assets can be sold and creditors repaid according to the original priority of claims. The issue starts getting more complicated, however, once one notes that under some circumstances it would be in the interest of the original claimants to restructure their claims rather than receive immediate payments. Should bankruptcy allow for such restructuring, and if so how? Before this question can be answered, an examination of why such restructuring may be desirable in the first place is necessary. The most relevant set of circumstances is associated with the adverse incentive effects of debt, discussed next.

*A restructuring perspective: agency problems of debt*

Whereas the preceding discussion of the common pool problem underscores the role of bankruptcy in resolving conflicts of interest among creditors, the agency problems of debt relate primarily to conflicts of interest between debtors and creditors. In environments in which default is an underlying concern, these conflicts of interest arise because debtors are typically interested in maximizing the equity value rather than the total value of the firm. Ex post, actions that are conducive to that objective typically may reduce the value of debt. Ex ante, they generate welfare losses and increase the cost of debt financing. These welfare losses are called the agency costs of debt.

In principle, these agency problems could be resolved if it were possible to include in the debt contract covenants specifying all state contingent actions that borrowers could undertake. However, creditors often cannot perfectly monitor the actions of debtors after a debt contract is written, either because of imperfect information, costly contract enforcement, or both. It is often not even possible to envisage all the possible future contingencies. And even if it were possible, specifying actions for each contingency would be of no use unless those were costlessly verifiable by third parties, such as courts.

Agency problems associated with debt may cause inefficiencies by distorting the debtor firm's investment deci-

sions. The literature has emphasized two main types of distortions. First, a debt overhang may cause insiders to forgo projects with positive net present values. A distortion arises because the firm undertakes the investment only if the expected returns are higher than the required debt repayment plus the cost of investment, whereas the efficiency rule is that expected returns should be higher than the opportunity cost of investment. Underinvestment is the result. Second, debt can encourage a firm to take excessive risk. To consider an extreme but illustrative case, suppose the value of equity is zero. The firm has the opportunity to invest in only one project, which has an uncertain return and a negative net present value. The firm would still invest in the project since, if it were successful, the return to equity would be positive. If the project was not successful, shareholders would get nothing, and the negative returns would be borne by creditors. While undertaking such investments jeopardizes the interests of debtholders, they may be unable to observe and prevent such investments.<sup>7</sup> Whether a debt overhang results in under- or overinvestment depends on the specific circumstances.

The adverse incentive effects of debt and the associated agency costs are magnified during periods of financial distress and may make recontracting beneficial for both creditors and debtors. There may be circumstances in which a reduction in the face value of the claims on an overindebted firm would ameliorate the adverse incentive effects of debt to such an extent that the efficiency gains—or reduction in agency costs—would outweigh the reduction in the face value of the debt and benefit the creditors as a whole. Recontracting may also entail the conversion of debt into equity. If it also allows creditors to obtain some control rights, or creates mechanisms, even if temporarily, to better monitor the debtor, the adverse incentive effects of debt may be further reduced. As will be discussed below, such recontracting occurs quite frequently without any recourse to formal bankruptcy procedures, especially in industrial countries.

When the number of creditors is large, however, a collective action problem may prevent recontracting even when it is desirable for the creditors as a group. Take, for example, debt reduction, and consider the incentives of an individual creditor. If all other creditors were engaged in reducing the face value of their claims, the individual creditor would prefer to hold out and still reap the benefits of efficiency gains. Therefore, rehabilitation of the debtor through a renegotiation of debt reduction may require claimants to act collectively, in a coordinated fashion. A bankruptcy reorganization procedure may provide a forum for such collective action.

Note that the restructuring of the liabilities of a company through recontracting the original claims can be desirable in nonbankruptcy situations as well. However, in the context of defaults, rehabilitation of the debtor by means of restructuring may present a more efficient alternative to liquidation. This would be the case when, for example, the firm's current financial difficulties are due to a temporary liquidity crisis or when the value of the firm's intangible assets is high.

*Additional stakeholders and conflicts of interest*

The preceding discussion has implied that the only relevant parties to a bankruptcy are lenders and borrowers. It also identified two main types of conflicts of interest: those among lenders, arising due to collective action problems, and those between creditors and debtors. Clearly, this view is an oversimplification. There are other stakeholders whose interests are affected by what happens in bankruptcy, for example, workers, suppliers, customers, and the state.

Some of the interests of these stakeholders are represented by specific financial claims against the debtor company. For example, the company may owe wages to its employees or may have purchased intermediate goods from its suppliers on credit. However, these explicit financial claims often do not capture all of the interests of these stakeholders. For example, employees may have wage agreements that represent claims on future earnings of the company. Consumers may hold warranties for products they have purchased. Finally, the interests of some parties are not represented by any type of explicit contracts, even though they are affected by what happens in bankruptcy. Examples are the disruption of implicit long-term contracts with employees and, in the event the firm liquidates, costs that would be borne by consumers in locating and purchasing from another company the spare parts for the products purchased from the bankrupt firm. Such interests are not represented in bankruptcy decisions even though they suffer real costs.

Often, the interests of these stakeholders are at variance with each other, not only because of collective action problems. The primary concern of lenders is repayment of loans. Workers, by contrast, may be interested not only in realizing their claims that arise due to unpaid wages, but also in maintaining employment. Hence, while creditors may wish to pursue liquidation, workers' interests may call for rehabilitation. And even lenders are not homogeneous in their interests. For example, whereas secured creditors are concerned mainly with taking possession of the collateral, and therefore would not be wor-

ried if a viable firm were liquidated, unsecured creditors may be better served if the firm were maintained as a going concern.

*The tension between the creditors' bargain and restructuring*

The restructuring perspective discussed above suggests a straightforward objective: Bankruptcy laws should strive to maximize the value of assets under bankruptcy, net of various costs incurred as the procedure unfolds (discussed in some detail below). A corollary of this objective is that bankruptcy should promote the liquidation of companies whose (post-restructuring) going-concern value is less than their liquidation value. Conversely, whenever the going-concern value of the company is larger than the liquidation value, the company should be reorganized.<sup>8</sup> Maximization of the value of assets would allow recontracting under bankruptcy, to reap any efficiency gains that may be available, and also would subsume efficiency of investment given agency problems of debt.

In principle, this objective—hereafter called the efficient restructuring rule of bankruptcy—need not contradict the creditors' bargain. After all, creditors as a class would benefit from the maximization of the value of the assets over which they have claims. For example, if it were possible to design a procedure that redistributed claims over the maximized value of assets without violating the pre-existing ranking of priority, such a mechanism would satisfy the creditors' bargain. However, given the various conflicts of interest afflicting bankruptcy, translating this objective into actual procedures, with their associated incentives and mechanisms of control, is a complicated task.

Given a set of rules, each stakeholder will act strategically, to maximize his or her own benefit. Strategic behavior, and imperfect and asymmetric information, often results in unintended outcomes. Bankruptcy rules designed to facilitate recontracting may allow for post-bankruptcy bargaining, which may be costly. Or they may influence the bargaining power of the different parties. In particular, they may grant excessive bargaining power to the debtor and may eventually lead to outcomes that violate the creditors' bargain.<sup>9</sup> Ultimately, such a procedure may act as a barrier to exit and may be used by debtors to defer liquidations. Similarly, bankruptcy rules that are designed primarily to repay creditors may end up liquidating viable firms or privileging some classes of creditors over others. These problems of institutional design create a tension between the objectives of respecting the creditors' bargain and promoting efficient restructuring.

Potentially strategic or disruptive behavior can be checked and kept under control by the court. The motive and behavior of the court is determined partly by the law itself, or by its current interpretation. For example, in cases in which a company wants to use bankruptcy procedures not to resolve financial distress but primarily to transfer wealth from other stakeholders to holders of equity, the law may instruct the judge to reject the petition. The law may prescribe the appointment of a trustee and endow that trustee with substantial decisionmaking authority, which would also limit strategic behavior by stakeholders.

In practice bankruptcy laws try to strike a balance between debt collection and restructuring. They differ in the way they distribute decisionmaking authority among the different stakeholders of the company and the court. As will be discussed below, none of the existing models of bankruptcy provides a perfect solution, and each has its own shortcomings. It is useful to examine actual bankruptcy rules and to identify the types of incentives they provide, the outcomes they produce, and problems they pose. More specifically, it would be useful to answer the following questions: How are control rights and decisionmaking authority distributed among the debtor, the creditors, and the court? How are secured creditors treated? How long do firms remain under bankruptcy? What is the likely efficiency of the bankruptcy outcomes associated with different designs? To what degree are creditors' original bargains protected? What are the roles of the court and the court-appointed trustees? These questions are addressed in the next two sections.

## **BANKRUPTCY POLICIES IN INDUSTRIAL COUNTRIES**

This section reviews and evaluates the bankruptcy codes of the United States, the United Kingdom, and France. These three codes differ substantially in the way they distribute decisionmaking authority among the debtor, the creditors, and the court. While the U.S. policy grants substantial bargaining power to debtors, the U.K. law is more creditor-oriented. The French legislation, like U.S. law, is debtor-oriented; although it is designed to preserve going-concern value, it grants more power to judges than to the debtor firm. These differences notwithstanding, discussions of bankruptcy law reform in the three countries reveal a tendency toward convergence: The U.S. system is criticized for being too lenient toward debtors; the U.K. system is seen as encouraging liquidations too rapidly; and a recent reform of the French code represents an attempt both to further protect creditors' rights

and to increase the power of the court. It is possible to detect a tendency in bankruptcy reform efforts across countries to enhance the possibility of maintaining going-concern values without granting bargaining power to the owners and managers of debtor firms.

### **The U.S. debtor-oriented approach**

The legislation covering liquidations and reorganizations is described first. This discussion is followed by an overview and an evaluation of empirical work on bankruptcy reorganizations and informal workouts.

#### *Liquidation*

Liquidation procedures spelled out in Chapter 7 of the U.S. Bankruptcy Code provide the basic framework for bankruptcy, for both firms that enter reorganization and those that file for liquidation. When a firm files for liquidation, the bankruptcy court appoints a trustee to close down the debtor firm, sell its assets, and deliver the revenues to the court. The court then uses the revenues to pay creditors.

The order in which creditors are paid is determined by the absolute priority rule, which specifies the following order of payment: first, the administrative costs of bankruptcy, including the fees for the trustee, and debts incurred after the filing of bankruptcy; second, claims that receive priority by statute, such as taxes, rents, and unpaid wages; third, unsecured creditor claims including trade credits, utilities, damage claims, and claims of long-term bondholders; and last, equity. Higher-priority claims must be paid in full before any payments are made to lower-priority claims. Hence, shareholders receive no payments unless all other creditors have been paid in full.

The absolute priority rule places secured creditors outside the priority ordering. These creditors have priority over funds received by the liquidation of the assets pledged as collateral. To the extent that these funds are insufficient to cover the entire claim, the balance is owed by the debtor and is considered part of the remaining unsecured claims. In principle, secured creditors may receive a payment even if all other creditors receive nothing.

#### *Reorganization*

The 1978 Bankruptcy Code, which replaced the Chandler Law of 1938, introduced significant changes to reorganization procedures in the United States.<sup>10</sup> The main purpose of these changes was to increase the likelihood that the firm would emerge from bankruptcy as a going concern. Essentially, Chapter 11 of the code allows for the renegotiation of the claims on the debtor firm. It

also prescribes a set of rules, including procedural rules, that govern negotiations.

A Chapter 11 case can be initiated voluntarily by a debtor or involuntarily by three or more creditors. The debtor need not be insolvent. Asset grabbing is prevented by an automatic stay invoked as soon as the bankruptcy petition is filed. The stay bars any judicial or administrative actions against the debtor and suspends all principal and interest payments. In particular, secured creditors lose their rights to seize or foreclose on the debtor's property. A key characteristic of Chapter 11 is that the current management of the company remains in control (as debtor-in-possession) until a reorganization plan is approved by the court (after it is negotiated with creditors).<sup>11</sup> The firm's management conducts the business of the firm but is monitored by the court. The debtor-in-possession also takes on fiduciary responsibilities; it has obligations to both shareholders and creditors. A creditors' committee representing the interests of unsecured creditors is formed to oversee the procedure. Although interest accruals on unsecured debt cease, secured debt continues to accumulate interest.

During the first 120 days of the filing, the debtor has the exclusive right to propose a reorganization plan. This exclusive period can be, and often is, extended by the court. The plan may envisage the continuation or the liquidation of the debtor firm. It separates creditors into classes and specifies how the claims will be repaid or reorganized. An accompanying disclosure statement provides information that creditors need to make an informed judgment on the plan. Secured creditors are treated individually.

The code specifies two procedures for the adoption of the plan. The unanimous consent procedure requires the approval of each class of creditors, by two-thirds of the face value of the claims of that class and one-half of the number of creditors. Under the unanimous consent procedure, the plan may reduce the claims of secured creditors, but in that case secured creditors have a right to vote on the plan. Under the cramdown procedure, the court may approve a plan even if some classes of creditors object. In that case, the dissenting class must be treated "fairly and equitably." For secured creditors, this means that they retain their liens and receive periodic cash payments equal to the depreciation of the value of the collateral. For unsecured creditors, fair and equitable treatment requires that they be paid an amount equivalent to what they would have received under liquidation according to the absolute priority rule. Since this payment requires a valuation of the assets, cramdowns are more

costly than unanimous consent procedures. Under both procedures, the plan is binding for all creditors once it is approved.

The code also has provisions to assist the firm in obtaining financial resources to maintain operations. Under the terms of debtor-in-possession financing, the debtor can raise unsecured loans as an administrative expense of bankruptcy, which has a high priority. If the debtor cannot obtain a loan at the administrative expense priority, the court may allow the firm to obtain a loan that ranks higher than all other administrative expenses or allow the debtor to raise secured loans.

The code grants the debtor-in-possession significant powers to recover certain prepetition transfers of the debtor's property, called avoidance powers. Their purpose is to prevent or reverse transfers that would enable some creditors to obtain more than their fair share of the debtor's assets—thereby violating the collective nature of bankruptcy—simply because these creditors either were able to move with greater speed or had more leverage to exact concessions from the debtor.<sup>12</sup>

#### *Empirical characteristics of Chapter 11 procedures*

What types of outcomes do these rules generate in practice? A few indicators were compiled in several recent empirical studies on bankruptcy reorganization in the United States:

*Costs of bankruptcy.* The costs associated with bankruptcy are often classified as either direct costs or indirect costs. Studies have shown that direct costs, such as fees for lawyers, the trustee, and investment banking services, range between 2.8 and 7.5 percent of the book value of the assets of the debtor company. Indirect costs arise from suboptimal actions associated with financial distress and bankruptcy. Some of these costs to the firm result from the higher transactions costs associated with bankruptcy status.<sup>13</sup> Others result from the strategic behavior of the firm while under bankruptcy and include agency costs and the cost of suboptimal investment decisions. The time spent in dealing with creditors and the bankruptcy court is another indirect cost to the firm. If Chapter 11 proceedings involve asset sales, and if assets are specific, the firm's going-concern value will decrease. The indirect costs of bankruptcy are believed to be larger than the direct costs.

*Violations of the absolute priority rule and the bargaining power of the debtor.* Widespread violations of the absolute priority rule have been presented as evidence of the inability of the Chapter 11 system to safeguard the creditors' bargain. In many Chapter 11 cases, the absolute pri-

ority rule is violated because shareholders retain some claims in the reorganized enterprise even though more senior claim holders are not paid in full. In Franks and Torous's (1989) sample of 30 firms, the absolute priority rule was violated in 21 cases; in 18 cases, the agreements awarded some payments to stockholder. Eberhart, Moore, and Roenfelt (1990) found that the mean percentage deviation from the absolute priority rule in terms of excess payments received by shareholders amounted to 7.6 percent of the total value paid to all claimants, ranging between 0 and 35 percent. In a sample of 30 cases, they found that 23 violated the absolute priority rule. In his study of 37 firms that had filed for bankruptcy, Weiss (1990) found that the absolute priority rule was violated in 29 cases. Priority was rarely violated for secured creditors. Shareholders retained some ownership or received cash payments in 30 cases, 28 of which were in violation of the absolute priority rule. As for unsecured creditors, priority was violated among different classes; for example, general creditors received some payments before senior bondholders had been paid in full.

There are several possible explanations for deviations from the absolute priority rule, not all of which reflect violations of the creditors' bargain (see Baird and Jackson 1988). If the firm is worth less than the amount owed the senior creditor, the senior creditor may wish to retain or recombine with the current owners of the firm because they have specialized skills that increase the value of the assets. As a result, while intermediate creditors are paid little or nothing, the current owners retain some stake in the reorganized firm. In this case, deviations from the absolute priority rule do not violate the creditors' bargain. The second explanation for a deviation from the absolute priority rule is that increasing the owners' stake in the financially distressed firm reduces their incentives to invest in excessively risky projects (Eberhart and Senbet 1993; White 1989). In this situation a violation of the absolute priority rule is seen as a measure that ameliorates overinvestment problems and one that need not violate the creditors' bargain. If, however, a violation of the absolute priority rule reflects bargaining power that the debtor gained from the renegotiation rules, it is likely that the creditors' bargain will be violated.

Several aspects of Chapter 11 grant the debtor substantial bargaining power. First, the debtor retains control over the firm. Second, the exclusive period gives the debtor a first-mover advantage in making proposals for an agreement. Third, automatic stay and the consequent cessation of interest accrual on the claims of unsecured creditors make this group of creditors more willing to accept

plans that dictate only partial and low repayment rates on their claims. Fourth, even though costly cramdown procedures are rarely used in Chapter 11 proceedings and cramdowns do not always favor debtors, the threat of their use may be an effective instrument to convince creditors to accept a particular reorganization plan. Finally, the debtor has substantial ability to delay the bankruptcy proceedings. If delays hurt creditors but not the debtor, the debtor gains added leverage over creditors.

*Delays in Chapter 11 proceedings.* Chapter 11 cases last a long time. In the sample of 30 firms that Franks and Torous (1989) examined, the period varied from 37 days to 13.3 years and averaged 4 years. In the study by Eberhart, Moore, and Roenfelt (1990) the time between the filing of a bankruptcy petition and plan confirmation varied from 10 months to more than 6 years, with an average of 2.1 years. In White's (1989) sample of 26 firms, the average case took 17 months. Weiss (1990) calculated an average of 2.5 years, with the range from 8 months to more than 8 years.

Delays can be caused by the existence of a large number of creditors or inadequate financial records, both of which can generate time-consuming disputes. Delays can also be caused by the strategic actions of debtors. Often, debtors bring lawsuits against creditors, question the validity of claims, or even defy court orders. If the value of the assets under bankruptcy decreases over time, even if due only to mounting indirect costs, and if the value of shareholders' claims is close to zero, the debtor may reduce the value of creditors' claims by delaying the proceedings. This provides the shareholders with a credible threat: By delaying the process, the shareholders lose nothing, but they can generate substantial losses on creditors. This threat increases shareholders' bargaining power and allows shareholders to dictate plans that violate the absolute priority rule.<sup>14</sup> Eberhart, Moore, and Roenfelt (1990) found a positive correlation between delays in proceedings and violations of the absolute priority rule, suggesting that such delays indeed reflect the debtor's bargaining power.

*Change in ownership and control.* Evidence on management turnover and changes in ownership under bankruptcy reorganization is important for several reasons. First, in cases in which the firm's poor performance is caused by inept management, a change in management may be an important component of restructuring. Second, if creditors can increase their ability to control and monitor the actions of the debtor, agency costs may be reduced. Third, low managerial turnover itself may reflect the debtor's bargaining power.

The general presumption is that Chapter 11 grants substantial control to the incumbent owners and managers of the debtor enterprise because management retains control unless a trustee is appointed by the judge (LoPucki 1983a, 1983b). Even though the legislation allows for the appointment of a creditors' committee, LoPucki reported that committees were appointed in only 40 percent of the 57 cases she studied and that most were ineffective.

In a more recent study, Gilson (1990) examined 111 publicly traded companies that had experienced severe financial distress. Sixty-one had filed for bankruptcy under Chapter 11, and 50 had restructured their debt privately. He concluded that financial distress generates significant changes in management and ownership. In 75 percent of the cases, banks and other creditors received significant blocs of voting power in the restructured firms. For those firms that restructured their debts, banks received stock in the restructured company in 47 percent of the cases. Their share in ownership averaged 37 percent. Creditors acquired ownership in 75 percent of the Chapter 11 cases and collectively retained about 79 percent of the bankrupt firms' equity.<sup>15</sup> In the 12 cases in which creditors controlled seats on the board of directors, they averaged 38 percent of the seats. Gilson also found evidence of a significant shift of control from the incumbent management and board of directors to nonmanagement bondholders and creditors. On average, only 46 percent of incumbent directors and 43 percent of the chief executive officers were still with their firms when the bankruptcies or debt restructurings concluded. About 16 percent of chief executives leave each year. These results are consistent with those of an earlier study by Gilson (1989) in which he documented an annual turnover in top management of 52 percent following financial distress compared with an annual turnover of 12 percent for a random sample of firms.

These findings underscore substantial changes in the management of bankrupt firms, certainly more than one would expect from the LoPucki study. They may also reflect means to reduce the agency problems of debt. But a question still remains: Do these remedies transfer sufficient power to creditors to enable them to influence decisions, particularly about issues where the interests of owners and creditors differ. According to Gilson (1989, 1990) some 50 to 60 percent of managers still remain in control at the end of one year of renegotiations, and stockholders continue to control a nontrivial number of seats on the board.

*Success of Chapter 11 filings.* Most firms involved in Chapter 11 procedures end up in liquidation. According

to a study by the Administrative Office of the Courts (cited in Westbrook 1993), a confirmed agreement is reached in about 25 to 30 percent of Chapter 11 cases. Even then, one-fourth of these plans envisage the liquidation of the companies. The ratio of successful cases has been increasing from a low of 13 percent in 1982, perhaps suggesting the presence of a learning process, that is, an increased capability in the industry to structure successful agreements. Data also show that larger firms are more likely than smaller firms to conclude a plan to reorganize.

*Postbankruptcy performance.* Hotchkiss (1992) studied the postbankruptcy performance of firms that had successfully completed a Chapter 11 procedure, with a confirmed reorganization plan. In her sample of 197 companies, more than 40 percent continued to experience operating losses in the three years following bankruptcy, and 32 percent filed for Chapter 11 or went through an informal workout for a second time. Hotchkiss found a close association between the continued involvement of incumbent management and poor postbankruptcy performance. She also discovered that the postbankruptcy performance of firms was worse than the earnings forecasts that management had presented to the court and creditors as part of the reorganization plan. Overall, the evidence suggests that Chapter 11 is biased toward the continuation of firms that should be liquidated.

#### *Chapter 11 versus informal workouts*

Informal, out-of-court workouts, an alternative to formal Chapter 11 proceedings for the renegotiation of the claims on a debtor company, offer several advantages. First, it has been argued that the transactions costs associated with informal workouts are lower than those incurred in bankruptcy. Gilson, John, and Lang (1990) found that the average time spent under Chapter 11 is 20 months, versus 15 months under an informal workout. In Franks and Torous's (1993) study, the differential is even greater: 27 months versus 17 months. These findings suggest that indirect costs may be lower in informal workouts. Gilson, John, and Lang estimated the direct costs of informal workouts to be less than 1 percent of the book value of assets, whereas estimates of the direct costs of Chapter 11 proceedings range from 2.8 to 7.5 percent.

Another advantage of informal workouts is that in principle only claims that are experiencing repayment difficulties need be restructured. If renegotiation is costly, then this also leads to cost savings over Chapter 11. For example, Gilson, John, and Lang found that only 70 percent of firms that undertook informal workouts and that

had publicly traded debt outstanding actually restructured such debt.<sup>16</sup>

If it is true that informal workouts are less costly, and hence provide a larger value of assets to be renegotiated, then one would expect that most recontracting of claims would take place out of court. In fact, as Gilson, John, and Lang argue, the larger the difference in the costs associated with each of the systems, the more likely it is that firms will prefer informal to formal restructurings.

However, informal workouts are vulnerable to the hold-out problem mentioned in the earlier discussion of the agency problems of debt. Their success often requires the unanimous agreement of creditors whose claims are in default. Such unanimity may be impossible if individual creditors hold out in order to free-ride on the benefits of debt restructuring or to obtain more favorable treatment.<sup>17</sup> A dissenting creditor excluded from the restructuring plan can resort to individual remedies or force the debtor into an involuntary bankruptcy. The voting rules of Chapter 11 alleviate the hold-out problem. When a reorganization plan has the required minimum number of votes in each class of creditors, it is confirmed and becomes binding on all the creditors. Moreover, dissenting classes can be forced to comply with the plan through a cramdown procedure.<sup>18</sup>

The hold-out problem in informal workouts is likely to be less severe when the claims on the firm are privately held, and among a few creditors, as in the case of bank loans. The problem is typically more severe when creditors are unsophisticated or diffuse, as in the case of trade creditors or publicly traded bonds. The situation is even more complicated because the Trust Indenture Act of 1939 requires the consent of every bondholder in order to change the principal amount, the interest rate, or the maturity of a bond, all of which would be essential in a restructuring.<sup>19</sup>

Empirical evidence generally confirms the importance of hold-out problems and indirect costs. Gilson, John, and Lang (1990) examined 189 financially distressed firms, of which 80 had successfully restructured their debts and 89 had failed and filed for bankruptcy reorganization. Firms that had successfully concluded their debt restructurings had relatively more bank debt. Bank debt is hypothesized to be easier to negotiate because banks are more sophisticated and less numerous than other kinds of creditors, resulting in fewer hold-outs. Similarly, successful informal workouts are also characterized by a lower number of debt contracts per unit value of book liabilities, a variable that again captures creditors' incentives to hold out. These firms also have higher market

value-replacement cost ratios. This ratio is an indicator of the going-concern value that might be lost in a formal reorganization, if reorganization resulted in higher sale of assets.<sup>20</sup>

Additional evidence on indirect costs is provided by Franks and Torous (1993), who compared 37 firms that had filed for Chapter 11 with 45 firms that had successfully completed an informal workout. (About half of the firms that petitioned for Chapter 11 had done so after attempting and failing in an informal workout.) Their study revealed that recovery rates for creditors' claims are higher in informal workouts (80 percent) than in Chapter 11 reorganizations (51 percent). Regression analysis revealed that although the recovery rates were not related to the firms' performance, they were significantly negatively related to higher asset sales. Following a suggestion from Shleifer and Vishny (1993), Franks and Torous interpret asset sales as an indicator of the indirect costs of financial distress arising from distressed or "fire" sales. Distressed sales of assets generate revenues that are lower than their long-run equilibrium values, which, all else constant, decreases the financial resources available to repay creditors. Interestingly, asset sales have a larger negative effect on recovery rates for firms under Chapter 11, providing additional evidence that Chapter 11 suffers from higher indirect costs.

Finally, Franks and Torous (1993) found that deviations from the absolute priority rule for equity are higher in informal workouts (9.5 percent) than in Chapter 11 reorganizations (2.5 percent). Equity deviations are correlated with two characteristics. The first is the insiders' option to delay the renegotiation process. In informal workouts this takes the form of a threat to file for Chapter 11; in Chapter 11 it reflects the threat to delay the firm's emergence from reorganization. Note that the value of this option is highest when the value of equity is close to zero or when the value of the firm is close to the face value of debt. Deviations from the absolute priority rule for equity are positively correlated with the value of the option.

The second characteristic that is correlated with equity deviations is the complexity of the firm's capital structure, which is captured by size. It is hypothesized that the larger the firm, the more difficult it would be for creditors to act cooperatively, and thus the easier for equity holders to gain concessions. Indeed, Franks and Torous found the deviations to be positively correlated with size. Interestingly, the impact of size is smaller for firms under Chapter 11, suggesting that relative to informal workouts, those aspects of equity bargaining power related to com-

plex capital structures may be diminished under Chapter 11. This result is probably associated with creditors' greater ability to act cooperatively. In general, only equity holders gain in informal workouts, whereas junior debt holders also gain under bankruptcy.

#### *An evaluation*

Empirical evidence seems to suggest that renegotiations through informal workouts entail smaller transactions costs than Chapter 11 proceedings. However, informal workouts are more vulnerable to hold-out problems.<sup>21</sup> Firms with simpler capital structures or fewer creditors are more likely to be successful in informal workouts.

The evidence also reveals the interdependency between informal workouts and Chapter 11 reorganizations,<sup>22</sup> and is consistent with the hypothesis that formal reorganizations are subject to higher transactions costs. In cases where most of these costs are likely to be borne by creditors (that is, when the value of equity is close to zero), the procedural rules of Chapter 11 grant significant bargaining power to debtors by providing them with an option to delay the procedures. (This is partly compensated by increasing the ability of creditors to act in a more coordinated manner.) *Ex ante*, shareholders' ability to threaten creditors to effectively decrease the value of the firm, and consequently the value of the creditors' claims, grants the shareholders bargaining power in informal workouts as well.

What are the implications for the creditors' bargain and restructuring? Since the bargaining power of the debtor derives primarily from the rules that govern the reorganization process, rather than the original contracts, granting such power violates the creditors' bargain. Considering investment efficiency, the important provisions of Chapter 11—such as automatic stay, equity violations of the absolute priority rule, and debtor-in-possession financing—generally increase incentives to invest. Hence, the net effect on efficiency depends on the nature of the agency problem. If the firm suffers from underinvestment, then reorganization under Chapter 11 may improve efficiency. However, in the case of overinvestment, these provisions of Chapter 11 exacerbate the problem (Gertner and Scharfstein 1991). The empirical evidence summarized above suggests that the latter is most frequently the case. Overall, Chapter 11 does not seem to promote the maximization of the value of assets under bankruptcy. Rather, it seems to encourage the rehabilitation of firms whose liquidation value is larger than their going-concern value.

Another problem with granting excessive bargaining power to debtors, as emphasized by the debt collection

view of bankruptcy, is that the procedure becomes quite vulnerable to abuse by firms that utilize it for purposes other than managing insolvency. Chapter 11 is especially susceptible to that problem because eligibility does not depend on the firm's being in a state of insolvency or even illiquidity. An example is the 1983 bankruptcy filing of Continental Airlines, which was motivated not by a concern about illiquidity or insolvency, but by the desire of the airline's management to renegotiate with labor.<sup>23</sup>

#### **The U.K. creditor-oriented approach**

Whereas the main purpose of Chapter 11 of the U.S. Bankruptcy Code is to maintain the debtor firm as a going concern, the primary objective of the United Kingdom's 1986 Insolvency Act is to encourage the repayment of creditors' claims.<sup>24</sup> Before enactment of the 1986 law,<sup>25</sup> the predominant insolvency procedure was receivership. However, receivership was believed to lead to the liquidation of companies that could be reorganized. To counter this weakness, the 1986 law introduced an alternative called administration. Although the administration procedure has been labeled by some as the U.K. equivalent of Chapter 11, it prescribes a fundamentally different set of rules and procedures for the treatment of insolvent debtors. Most important, all procedures envisaged in the act have a common feature: On their initiation, managers and owners lose their control over the debtor company.

An interesting provision of the U.K. law requires managers to declare insolvency as soon as a reasonable prospect for avoiding a default ceases to exist. Managers who fail to do so can be disqualified from holding a position on the board of any company for as long as 15 years.<sup>26</sup> The intention of the provision is to encourage an earlier declaration of insolvency to avoid value losses.

The three options under the U.K. system—liquidation, administrative receivership, and administration—are discussed below.

#### *Liquidation*

A creditor or the company itself can request the appointment of a liquidator. The role of the liquidator is similar to that in the United States: to sell enough assets to satisfy all creditors' claims in accordance with their respective legal rights.

#### *Administrative receivership*

A receiver is an individual appointed by a secured creditor (the appointor) to enforce its security. Whereas a receiver is appointed over a particular asset, an adminis-

trative receiver is appointed over the entire company's assets by the holder of a "floating charge," which in most cases is a bank.<sup>27</sup> The receiver decides whether the company should be maintained as a going concern. If the receiver decides not to do so, he sells the assets to pay the claims of the appointor. The balance is then passed to a liquidator. If the company has positive cash flow, it is often possible to sell the business; if cash flow is negative, additional financing is required. Such financing is often secured from the appointor. Going-concern sales are often made to incumbent management, possibly due to management's superior knowledge about the true state of the business (Franks and Torous 1992).

The receiver is mainly responsible to the appointor. By contrast, the administrative receiver is responsible to the preferential creditors, to the appointor, and, to a lesser degree, to junior creditors. To protect the interests of other creditors, the law imposes restrictions on the behavior of the receiver; for example, he is required to sell the assets for a full price. Franks and Torous (1992) reported a case in which a receiver was successfully sued for failing to advertise properly the sale of an asset.

Nevertheless, the incentive of the administrative receiver to realize the going-concern value of a debtor firm is relatively small. When a conflict of interest is likely to arise between a secured creditor and junior creditors, the administrative receiver is likely to decide in favor of the secured creditor who appointed him. The absence of an automatic stay also limits the behavior of the administrative receiver. The appointment of a liquidator usually prevents the administrative receiver from managing the firm as a going concern. It is also unlikely that an administrative receiver would decide to rehabilitate a company if faced with opposition from the appointor. According to Clarke (1993), the main factor that discourages administrative receivers from rehabilitating a company in such circumstances is that they are professionally dependent on a small group of financial institutions that makes most appointments. She also indicates, however, that once an administrative receiver decides to rehabilitate a company, with the consent of the appointor, administrative receivership is probably the procedure most likely to succeed.

#### *Administration*

Either the debtor company or a creditor can request the appointment of an administrator to represent all creditors' claims. The court will appoint an administrator only if at least one of the specified purposes can be achieved, namely, to maintain the company as a going concern, to

secure a more advantageous realization of the company's assets, or to come to an arrangement with creditors. Once an administrator has been appointed, an automatic stay is in force over all proceedings and actions against the company, and a liquidator cannot be appointed.

The administrator takes over the management and control of the company and has extensive powers, including the power to remove the company's management. He is required to produce formal proposals for an arrangement within three months of his appointment. The proposal is submitted to a creditors' meeting, where it is voted on. Confirmation requires an affirmative vote by creditors representing at least 50 percent of the outstanding claims.

The administration procedure, which is much more conducive to maintaining the company as a going concern than is administrative receivership, is nonetheless rarely used. Creditors secured by a floating charge may prevent the granting of an administration order by appointing a receiver before the court rules on the request for administration. Furthermore, an administrator can be appointed only if the receiver resigns his office. This provision limits the use of administration. Generally, secured creditors rarely have incentives to relinquish control to an administrator, since a receiver better serves their interests.<sup>28</sup> The administration procedure is most often used in cases where no creditor is secured by a floating charge.

As noted by Franks and Torous (1992), the administration procedure seems to suffer from a fundamental inconsistency. The appointment of an administrator is likely to be most warranted when the conflict of interest between secured and unsecured creditors is acute. This would happen, for example, when the liquidation value of a company is less than its (uncertain) value as a going concern yet covers the face value of the secured creditor's claim. In such cases the secured creditor is likely to prefer liquidation, whereas the interests of unsecured creditors, as well as the restructuring criterion, favor continuation. Although an administrator could in principle prevent a premature liquidation and maintain the going-concern value, it is under these exact circumstances that the secured creditor is likely to preempt an administration order.

#### *Creditors' bargain or premature liquidations?*

The shareholders or incumbent management of the debtor are afforded smaller, if any, bargaining power by the U.K. insolvency procedures than they would be by the U.S. bankruptcy reorganization procedure. Control in the U.K. system is exercised by the receiver, administra-

tor, or liquidator, all of whom are certified insolvency practitioners (and often accountants). The ability of secured creditors to block an administration procedure has been interpreted as a guarantee that the Insolvency Act obeys the creditors' bargain (Webb 1991). Indeed, although detailed statistical evidence is not available, there exists a general presumption that in most cases absolute priority among different classes of creditors is honored. Nonetheless, by favoring secured creditors, the U.K. system may jeopardize the interest of junior creditors. The U.K. code thus can be said to respect the bargain of the secured creditors rather than that of all creditors.

The U.K. procedures are also believed to entail lower transactions costs. Shareholders have no power to delay the process, receivership results in a speedy settlement of claims, and the receiver can act without having to report back to the creditors or the court on a day-to-day basis. Perhaps most important, none of the procedures involves costly and convoluted bargaining.

However, whereas the U.S. system creates strong incentives to maintain a company as a going concern even when it is worth more in liquidation, the U.K. system may do just the opposite. By emphasizing the rights of creditors, and in many cases giving priority to secured creditors, the system may result in premature liquidations. Although it may be too early to judge the impact of the administration procedure, the small number of administration cases—perhaps a few hundred compared with thousands of receiverships—suggests that administration has not produced a radical change in the U.K. insolvency system.

As for investment efficiency, the U.K. code is more likely than its U.S. counterpart to exacerbate problems of underinvestment. If maintaining the debtor firm as a going concern requires new financing, it is more difficult to raise such financing in the U.K. system because no automatic priority is granted to such financing as it is in the U.S. system. Acquiring such priority for new financing would require the consent of existing creditors.

To summarize, the U.K. system avoids some of the main problems of the U.S. reorganization procedures and more closely respects the creditors' bargain. However, these advantages are achieved possibly at the cost of premature liquidations and underinvestment. The introduction of the administration procedure does not seem to have compensated for these shortcomings. The main problem seems to be that the creditors' bargain is interpreted too narrowly, in a way that favors one group of creditors at the expense of others. One may hypothesize

that a system that represents all creditors, and that attempts to resolve conflicts of interest between different classes by promoting a course of action that maximizes repayment to all creditors, may help prevent premature liquidations. To achieve such a system may require weakening the ability of secured creditors to veto the appointment of the administrator.

### **The French court-controlled system**

The French Insolvency Act, enacted in 1985 and reformed in 1994, has three stated objectives: maintain the firm in operation, preserve employment, and enforce credit contracts. The law prescribes a single process for all cases of insolvency covering both reorganization and liquidation. There are two procedures, one ("simplified procedure") for small firms (those with less than 50 employees or sales less than 20 million francs) and another ("general procedure") for large firms.

An interesting aspect of the French legal framework regulating situations of financial distress is that it is directed partially at preventing bankruptcies.<sup>29</sup> The Bankruptcy Prevention Act of 1984 prescribes several preventive measures aimed at assisting the debtor in reestablishing its financial health before defaulting or becoming obligated to file for bankruptcy. The purpose of the act is to provide a framework for negotiations between a company and its principal creditors and the expert assistance to help the company resolve its financial difficulties, thereby promoting informal workouts.

A company seeking relief under the 1984 act petitions a commercial court. If the court is convinced that bankruptcy is inevitable, it appoints a conciliator under whose supervision the company and its creditors negotiate an agreement, which they file with the relevant agencies. The agreement is treated as confidential. Failure of the debtor to comply with the terms of the agreement is deemed an act of bankruptcy.

Under the general procedure, the court appoints an administrator and a representative of the creditors. The commencement of bankruptcy also initiates a six-month observation period (which may be extended for an additional six months), during which payments to creditors are halted.<sup>30</sup> Any financing secured during the observation period is treated as a priority claim. Control of the debtor may remain with the current management, under the supervision of the administrator, or the court may order the administrator to assume effective control. The observation period ends with a judge's decision on whether the company will continue in the same legal form, be sold to third parties, or be liquidated.

The administrator, having decided that the company has a chance of survival, prepares a plan of reorganization. The plan may include debt write-offs, sale or shut-down, or even the addition of certain lines of business. It may envisage continuation or the partial or total sale of the company. Under a continuation plan, the owners of the company remain the same, although equity may be restructured. The court also may impose the replacement of the managers. Once it receives the plan, the court schedules a hearing at which all relevant parties, including workers' representatives, can voice their views. But they can neither vote on the plan or veto it. The decision on whether to adopt the plan rests with the court.<sup>31</sup>

A reorganization plan also may envisage the sale of the company. The Insolvency Act emphasizes and facilitates the sale of the business, in part or as a whole, as a solution to the company's problems. Potential purchasers may bid for its sale as soon as bankruptcy starts. Purchase offers must provide details on how future activities will be financed and on future levels of employment. The law instructs the judge to choose the sale that ensures the highest level of employment and of payment to creditors.

If the court deems that the company cannot be rehabilitated or sold, it can order the company's liquidation. In that case, a liquidator is appointed, and the assets of the company are sold to satisfy creditors' claims.

According to a recent study, about 94 percent of all bankruptcy cases end in liquidation (Biais 1994). For cases involving large companies, this rate drops to 40 percent. According to evidence from the Toulouse region, 80 percent of reorganizations end with continuations when the firms are owned by managers (private proprietorships), whereas 80 percent of larger, publicly held companies end up in sales. The observation period is about one month in cases that end in liquidations and seven months in those that end in reorganizations.

Based on a sample of 1,000 firms (and 1,200 loans), Malecot (1992) reported that the average recovery rate for bank loans is 69 percent under plans that envisage continuation and 55 percent under liquidations. Biais (1994) cited evidence that repayment rates for creditors are below 35 percent in sale reorganizations in the Toulouse region. Also, in 99 percent of cases that end in continuation, incumbent managers remain in charge.

The main distinguishing feature of the French system is the exclusive power of the court to determine the course and outcome of the bankruptcy process. Because the three objectives identified in the law—maintaining the firm as a going concern, preserving employment, and satisfying creditors' claims—are potentially contradictory,

the court often must strike a balance between them. Most judges who administer bankruptcy proceedings are businessmen, which possibly encourages economic reasoning in the resolution of bankruptcy. A study reported by Biais (1994) found that when the court was faced with a variety of offers, it acted to preserve employment in 33 percent of the cases, to maintain economically viable firms in 26 percent of the cases, and to pay back creditors in 24 percent of the cases. In addition, the work force and management agreed with the decision of the court in 80 percent of the cases, whereas the creditors agreed in only 23 percent of the cases.

Despite the limited empirical evidence, it is safe to conclude that the French bankruptcy system was not designed to satisfy the creditors' bargain. Regarding investment efficiency, given the law's concern with employment and maintaining the debtor as an operating unit, underinvestment or premature liquidations are not likely to pose a significant problem. In fact, if anything, the law may result in overinvestment and deferred liquidations. Even though there may be a bias toward maintaining as going-concerns even firms that are worth more under liquidation, debtors are not likely to benefit from excessive bargaining power either, for two reasons. The first reason is that the law limits the duration of the observation period. French debtor firms typically spend less time under bankruptcy reorganization than their U.S. counterparts. Second, debtors have little control over the company during the observation period, and they have no right to propose a plan (although they can influence its design).

Amendments to the French Insolvency Act—a new law adopted in 1994—introduced some significant changes. First, bankruptcy prevention mechanisms were strengthened by requiring the social security agency to warn the court when a company fails to pay its contributions. Second, the court can grant an automatic stay or even impose a reorganization plan on dissenting minority creditors in the prebankruptcy stage. Both these changes increase the role of the court in resolving financial distress. As for the formal bankruptcy procedure, the changes introduce more protection for creditors' rights. Most important, the new law requires the purchaser to repay secured creditors in full.

### **In search of a better model**

Bankruptcy law in industrial countries has so far been unable to satisfy the debt-collection aspect of bankruptcy without sacrificing efficiency in restructuring. The problem can be stated as follows: How can control rights and

decisionmaking authority be distributed in bankruptcy reorganization so that the emerging institutional structure encourages outcomes that maximize the total value of the firm, rather than specific claims on it, without violating the creditors' bargain?

The U.S. code—if not directly, through granting bargaining power—grants the debtor-in-possession substantial control rights and decisionmaking authority, which results in delayed liquidations. Although the U.K. code gives a leading role to secured creditors, premature liquidations become the main problem. The French system grants the judge power to impose solutions on all parties. But without a personal stake in the whole process, this third party is not compelled to work to maximize the value of the company subject to the creditors' bargain.

The review in the preceding sections helps in identifying some elements of an improved reorganization procedure. The U.S. model suggests that an improved model would reduce the bargaining power of the debtor. Two modifications are suggested. The first is to introduce tighter limits on the reorganization process, which judges are obligated to enforce. Doing so would decrease debtors' ability to threaten creditors by delaying the procedure and would curtail debtors' bargaining power in informal workouts. The second, more important change would be to curtail the control rights of the debtor once a firm is in bankruptcy. This can be achieved by having the judge automatically appoint and grant substantial managerial authority to a trustee, who might be monitored by the court and the creditors.<sup>32</sup>

The discussion on the U.K. system suggests that to prevent premature liquidations, the power of secured creditors to veto the appointment of an administrator should be reduced. Unsecured creditors thus would not be marginalized in the decisionmaking process. The administrator would still have as a mandate the maximization of the value of the firm.

With these changes, the U.S. and U.K. systems would be much more similar to each other than they are now. In both systems, substantial responsibility and decisionmaking authority would be given to third parties, such as trustees and administrators. To preserve the creditors' bargain, any reorganization plan prepared by the administrator would be subject to the creditors' approval. With the threat of delay removed, the outcome of the voting system would more closely reflect the creditors' preferences, rather than the bargaining power of the debtor.

The main difference between the modified U.S. and U.K. systems proposed here and the French system is that the creditors would retain voting power. Whether the

administrator would be appointed by the court or by the creditors, the boundaries of his authority would have to be worked out. In addition, the law would have to give the administrator and the court an economically meaningful mandate. Commercialization of administrators' services, and perhaps linking their rewards to the bankruptcy outcome, may increase market discipline on the whole process.<sup>33</sup>

A very different solution, one that simulates the market mechanism more closely, was proposed by Aghion, Hart, and Moore (1993). Once a firm files for bankruptcy, all debts would be canceled, and ownership rights would be allocated according to absolute priority. Senior creditors would receive actual shares; junior creditors would receive options to buy shares at a price equal to the claims of the senior creditors. Shareholders would have an option to buy shares at a price equal to the face value of the claims of all creditors. This mechanism aligns incentives among creditors, and between creditors and shareholders (and hence encourages the realization of maximum value for the firm), and preserves the absolute priority rule at the same time. Nevertheless, it also requires a well-functioning financial market so the parties can raise the cash necessary to exercise their options. If such financing is not available, ownership of the firm is effectively transferred to senior creditors (or, more correctly, to the particular class of creditors ranking above the class that cannot raise financing). Efficient,<sup>34</sup> but nevertheless unfair,<sup>35</sup> outcomes are likely to be the result.

## **BANKRUPTCY POLICIES IN INDUSTRIALIZING COUNTRIES**

As noted in the introduction, bankruptcy policy in industrializing countries suffers from a set of fundamental shortcomings not (or no longer) encountered in the industrial world. Some of these pertain to bankruptcy laws, but more concern the general institutional and regulatory environment.

An overview of these shortcomings is provided here, followed by a discussion of corporate reorganization procedures in India to illustrate some of these problems in some detail. Bankruptcy reform in Colombia is then reviewed, along with some of the factors that explain its relative success. Finally, implications for bankruptcy policy are suggested.

### **Main problems**

The main problems of bankruptcy policy in industrializing countries include outdated legislation, lack of differ-

entiation between enterprises and their owners and managers, an inadequate institutional structure, poor supervision and regulation of the banking system, and regulatory barriers to the mobility of labor and capital. These are discussed below.

#### *Outdated legislation*

Bankruptcy laws in many industrializing countries are outdated. In Venezuela, for example, rules governing bankruptcy procedures were established in the commercial code. The rules were inspired by the French and Italian legislation in the late nineteenth century. The Venezuelan code took its current form in a reform undertaken in 1919.

The Turkish code was inspired by Swiss legislation dating to the late nineteenth century. Although several amendments to the code were enacted throughout the 1980s, the core of the code was not substantially changed.

The age of a piece of legislation, of course, is not in itself evidence of inadequacy. The problem is that the framework of bankruptcy law in industrializing countries in general is ill suited to address the insolvency problems of modern corporations. For the most part, the law has not been informed by the substantial learning in this area over the past two decades, especially in industrial economies.

#### *Lack of differentiation between enterprises and their owners and managers*

Bankruptcy laws often fail to make a clear distinction between a company as a productive unit and its owners and managers.<sup>36</sup> A direct consequence is that the rehabilitation of a debtor company means that its owners must be "saved" as well. Reorganization procedures are typically perceived as a mechanism for providing the owner a "breathing space," rather than one that aims at preserving or enhancing the going-concern value of the enterprise.

This view has several implications. First, in cases in which failure of the enterprise is primarily due to mismanagement by owners and managers, the probability of inefficient outcomes is increased: Firms that might be viable under more able management may be liquidated or, conversely, firms may be rehabilitated without fixing the root causes of their failure. Second, striking a balance between adhering to the creditors' bargain and restructuring is more difficult. A legal framework that emphasizes rehabilitation ends up providing excessive bargaining power to the owners and managers of the debtor enterprise and grossly violates the creditors' bargain. A legal framework that aims at protecting creditors'

rights, by contrast, is likely to encourage substantial losses in going-concern values.

An important version of this problem occurs when the code associates default (and bankruptcy) with fraudulent behavior. The reorganization procedure in the Turkish code restricts eligibility to "honest" debtors. In Venezuela the establishment of bad faith on the part of the debtor results in the termination of the reorganization procedures and initiates a liquidation procedure. Fraudulent behavior by management may therefore result in liquidation of a firm that has higher value under continuation. In Colombia the initiation of a liquidation procedure triggers a criminal investigation of the debtor. The emotional aggravation as well as the possible loss in reputation associated with such an investigation might well deter the debtor from making a bankruptcy filing. In addition, pre-bankruptcy agency problems that encourage excessive risk taking might also be exacerbated.

#### *Inadequate judicial and financial infrastructure*

In many countries the processing capacity of the court system is severely limited. Courts are underfinanced and underendowed with staff and equipment, and record keeping is poor. As a result of these deficiencies, bankruptcy procedures are extremely lengthy, even absent delays due to strategic behavior of interested parties. The situation in Colombia prior to a reform that revamped the bankruptcy system is a case in point. The Colombian reorganization procedure required court approval of the agreement reached by a debtor and its creditors. Of a sample of 19 cases awaiting such approval in 1989, nine had been waiting for more than a year. When limited processing capacity combines with strategic behavior, delays of course increase further.

The skills needed to ensure successful reorganizations are also relatively scarce. Because judges typically are poorly informed about corporate finance, the proceedings seldom benefit from economic reasoning. One of the major deficiencies of reorganization procedures in India is inadequate project appraisal skills (discussed below). The extent to which judges can make a valuable contribution to the recontracting process thus is severely limited. By contrast, one of the major reasons for the apparent success of bankruptcy reform in Colombia is that the competent authority is well endowed with the necessary technical and financial skills (see below).

Trustees appointed by the court typically are not capable of handling the complicated financial transactions that may be required for successful reorganizations, nor do they have the skills to run a company on even a tem-

porary basis. The Venezuelan law, for example, only requires a trustee to be 21 years old, to be a businessman or a lawyer, and not to have declared bankruptcy. Most trustees are lawyers, yet a common concern among professionals who take part in bankruptcy procedures in Venezuela is that trustees lack the necessary skills to handle bankruptcy cases.

Another problem common to industrializing countries is deficient legal documentation. Before reforms in Colombia, inadequately prepared loan documents made validation of debt claims one of the major causes of delays in reorganization procedures. Debtors that wanted to prolong the proceedings could contest the validity of debt claims; their objections had to be resolved in ancillary lawsuits, which dragged on because of the limited capacity of the court system.

Inadequate documentation also facilitates asset stripping. Even when assets are pledged as collateral, owners and managers can remove them from the enterprise or transfer them preferentially to friends or family members. In the latter case, insufficient documentation makes it more difficult for judges or trustees to exercise avoidance powers and recover the transferred assets.

Problems with legal registries are also common. In Jamaica a mortgage can be registered either under the Companies Act with the registrar of the companies or under the Registration of Titles Act. In the event of liquidation, which is governed by the Companies Act, mortgages not registered under the Companies Act can be treated as invalid. The flow of information between legal registries is very poor in many countries, allowing fraudulent transfers and registration of more than one claim against a single collateral. The main effects of these shortcomings is to reduce financial intermediation. To the extent that lending does take place, however, these problems make it more difficult to resolve conflicts in situations of insolvency.

Finally, inadequate accounting and disclosure rules increase information costs associated with financial distress and bankruptcy. Absence of these rules makes it more difficult for stakeholders to assess the value of the various options available to them. In particular, it complicates a correct appraisal of the going-concern value of the firm and its liquidation value. The degree of asymmetry of information between owners and other stakeholders also may increase. When accounting rules are lax or not standardized, and a company's books do not convey credible information about the company's true financial situation, insiders typically possess more information than outsiders. This situation aggravates agency problems

associated with debt financing, expands the scope for disruptive behavior, and increases the cost to reach an agreement under bankruptcy reorganization.

#### *Inadequate supervision and regulation of the banking system*

Bankruptcy law attempts to resolve collective action problems in debt collection and therefore presumes that creditors will actively seek repayment. Its effectiveness in inducing efficient restructuring and exit is similarly predicated on such behavior. In industrializing countries, where capital markets are typically underdeveloped, the banking system is the major creditor of the corporate sector, especially for large firms. Absence of adequate supervision or prudential regulation may under certain circumstances diminish banks' incentives to behave as aggressive creditors.

Banking systems typically suffer from a variety of market failures and policy-induced imperfections. One widespread example is explicit or implicit deposit insurance.<sup>37</sup> Another is imperfect or asymmetric information about the quality of the assets of the bank. These imperfections may create incentives for banks to take on excessive risks at the expense of depositors and the state, especially when financial distress among borrowers creates distress in the bank as well. This may in turn discourage active debt collection, or recourse to bankruptcy, and may even induce banks to refinance bad loans.

In industrial countries, supervision and regulation of the banking system curtails such behavior by forcing banks to maintain and disclose accurate assessments of the riskiness of their assets, discouraging excessive risk taking (by requiring banks to maintain a minimum level of equity capital), and in extreme cases penalizing poor performance or excessive risk taking. These in turn increase incentives for debt collection and recourse to bankruptcy. In many industrializing countries, however, especially those that have not embarked on financial sector reform, banking regulation and supervision are weak or ineffective. As a result, main creditors in the financial system have no incentives to behave as bankruptcy law typically presumes they will.<sup>38</sup>

Perhaps a more fundamental problem exists when the creditor banks are owned by the state. In such cases, the problem is not inadequate supervision or regulation but the fact that their behavior is influenced by political considerations, or that their economic mandate requires them to deviate from sound commercial practices. Under political influence, banks may make loans without regard to some firms' creditworthiness. If such loans are not

repaid, pressure can be put on banks to delay or even forgo attempts at debt collection. In many cases, the economic mandate of such banks has been to allocate loans on the basis of industrial or social policy criteria, rather than economic viability. Or banks have been asked to act as the disbursement agency of the government's economic ministries or the central bank. In these cases as well, banks have no incentive either to collect debts or to initiate bankruptcy.<sup>39</sup>

#### *Regulatory barriers to the mobility of labor and capital*

The degree to which bankruptcy can successfully carry out its debt-collection and restructuring roles critically depends on the absence of regulatory barriers to the mobility of labor and capital. While the mobility of these two factors can be taken for granted in most industrial countries, there are significant restrictions in many industrializing ones.

Constraints on labor redeployment or retrenchment eliminate one of the most crucial potential sources of productivity increases and reduce the attractiveness of restructuring. Even though the main objective of these restrictions in many countries is stated as the protection of labor, labor ultimately carries the main cost of immobility. Inability to undertake corrective measures often encourages managers and owners to salvage whatever they can from the firm through asset stripping. Company revenues dry up, production stops, and wage claims are rarely repaid.

Restrictions on capital can also hamper restructuring. Restrictions on sales of assets or land eliminate one of the most valuable sources of finance for corporate reorganizations.

#### **Corporate reorganizations in India**

Reorganization procedures in India are illustrative of many of the problems just discussed. The barriers to industrial and corporate restructuring in India involve the legal framework for corporate reorganization and liquidation as well as regulations that constrain the mobility of labor and capital.<sup>40</sup>

The legal framework for formal corporate reorganizations in India is laid out in the Sick Industrial Companies (Special Provisions) Act of 1985, designed to expedite the rehabilitation of faltering industrial firms. One of the main shortcomings of the act is the criteria it establishes for eligibility: To benefit from the act, a company's cumulative losses must be larger than its net worth. Designing a viable rehabilitation plan for a company with such a poor financial structure would be quite difficult.

Nonetheless, qualifying companies are referred to the Board for Industrial and Financial Reconstruction (BIFR). If the company chooses to propose a rehabilitation scheme, it must be accepted by all involved parties (creditors, labor, state and central governments) to receive the board's approval. If the company offers no proposal, and the board decides that it is in the public interest to rehabilitate the company (which it always does), it appoints an operating agency (usually a financial institution) to examine the company's potential for rehabilitation. The agency's report to the board may include a proposal for rehabilitation or a recommendation for liquidation. If the agency proposes that the company be liquidated and one or more parties disagree, the board can either refer the case to the high court for liquidation or sell the assets. The board remits the sale proceeds to the high court for distribution to claim holders.

The board's procedures are lengthy. The mean duration of the cases handled between 1987 and 1992 was slightly more than one year. Over 19 percent of the cases were resolved after three years. Procedural rules are a principal cause of delays. The procedures require unanimous consent of all parties at almost all stages, which is to say that all parties have veto powers and can delay the proceedings as a matter of bargaining strategy. Goswami and others (1993) noted that:

Promoters veto original OA [operating agency] reports on the ground that they have better schemes; three months later they present something that is unviable and unacceptable to the BIFR. Consultants prepare estimates of productivity and profitability that often exceed those of the best firms in the industry (p. 20).

Procedures can also be stopped by appeals. Although there are cutoff dates for claim holders to take action in appeal cases, there are no limits on the time taken by the board or the appellate authority.

A second cause of delay is the board's preference for exhausting all possibilities of rehabilitation, even though, by the very nature of the eligibility criteria, most cases involve firms that are not viable. This preference precludes the board's using the threat of winding up (liquidation) to encourage consensus. Even when the company fails to carry out a sanctioned scheme, the case is referred back to the board rather than sent directly to be wound up.

Insufficient staff also contributes to delay. In 1993 the board consisted of a chairman and six members, to whom

1,010 cases were referred between 1987 and 1992. This staff shortage also caused cases to be determined without sufficient analysis.

Another deficiency of the Indian system is that the criteria used to determine the feasibility of rehabilitation schemes are flawed. Goswami and others (1993) found that very few of the board-sanctioned rehabilitation schemes satisfied the minimal viability criteria: The return on new loans and the return on equity, properly discounted, should be positive. As a result, firms that should have been liquidated were maintained as going concerns, often with vast injections of new, subsidized financial resources. Not only were the criteria established to assess the viability of rehabilitation schemes based on undiscounted financial flows; in addition, as is banking practice in the public sector, banks' success was judged by the volume of deposits and loans rather than portfolio quality or loan recovery. Banks thus had incentives to treat bad accounts as operationally sound, since recognizing a bad loan as such would have ultimately reduced the size of the bank's loan portfolio (Anant and others 1994).

According to Goswami and others (1993), a large percentage of the companies that were put under a rehabilitation plan failed to improve their performance. In 1991, 39 of the 164 schemes sanctioned that year had failed, and 64 of the companies continued to incur losses. Most rehabilitation plans actually made recovery in performance difficult by further increasing the debt burden of the company through injection of new subsidized loans. The conversion of debt to equity was disallowed by the Reserve Bank of India until 1992. Even after conversions were allowed, however, they were used only infrequently because of the tax disadvantages.

Finally, as in the United States, the incumbent management retains control under Indian bankruptcy law. The Companies Bill allows creditors to request a change in management if they can prove improper behavior or if management has suspended payments to creditors. But improper behavior is hard to prove. And as for defaults, Indian courts have taken the view that defaults of limited liability companies are not covered by the Companies Bill.

Cases referred to the high court for liquidation take at least 10 years to conclude, and creditors rarely recover any of their claims. Although the time to conclude a case could be significantly shortened if the board sold the company's assets, the board has rarely chosen this option.

In most cases both the creditors' bargain and the restructuring criterion of bankruptcy are grossly violated. However, even if the provisions of the bankruptcy law

were vastly improved, new and more efficient principles were identified to guide the behavior of the Board, and its technical and processing capacity was increased, several aspects of India's economic and regulatory environment would still limit efficient corporate reorganization. First, the sale of surplus land is tightly restricted by both the Urban Land (Ceiling and Regulation) Act and local governments. Second, the Industrial Disputes Act stipulates that state governments must approve labor cutbacks; state governments have consistently refused to grant such permission. These regulations severely constrain the flexibility of enterprises in responding to financial distress and prevent them from raising finance.

### **Bankruptcy reform in Colombia**

The rules governing insolvency procedures in Colombia underwent significant reform in 1989. The experience in Colombia is interesting for two reasons: the prereform bankruptcy system provides a vivid example of the problems associated with an extremely debtor-oriented code, and the reform introduced innovative institutional solutions to these problems.

Before 1989 there were two insolvency procedures in Colombia. The first was basically a reorganization procedure, designed to conclude a conciliatory agreement between a company experiencing financial difficulties and its creditors with the purpose of rehabilitating the debtor's business. The second was a liquidation procedure. Reorganization procedures were of two types: The first procedure was optional and used mainly by small and medium-size companies. The competent authority was the district judge. Firms under the supervision of the Superintendency of Companies that experienced insolvency problems, by contrast, had to go through a mandatory reorganization procedure.<sup>41</sup> These firms could not petition for liquidation before reorganization was attempted.

Mandatory reorganizations, which were overseen by the superintendency, were widespread in Colombia. In 1986 and 1987, for example, 60 of some 1,000 manufacturing companies supervised by the superintendency were under reorganization. According to data from the superintendency, the value of their assets was about 12 percent of total assets, reaching 20 percent in some sectors such as textiles.

Under the mandatory procedure, the debtor retained management of the company unless fraud was established. The superintendency validated the list of claims, and any objections raised by the involved parties were resolved. The parties were then convened in a hearing,

where they voted on a proposal for agreement. If an agreement was reached, it was sent to the district judge for confirmation. If no agreement was concluded, a liquidation procedure was initiated.

The procedures were typically lengthy. As of May 1989, of 82 cases in which an agreement had not been reached, 54 had been ongoing for more than two years and 36 for more than three years. The stage that caused most of the delays was the validation of claims. Debtors used their option to delay by raising objections, which had to be resolved by the district judge. The fact that many loan documents, especially those involving small creditors, had been inadequately prepared helped the debtors in that respect. The debtors had other means to delay the process as well. For example, by not attending the hearing, they could ensure postponement. The creditors had no remedies against such behavior. In addition, the parties could appeal almost every decision of the judge or the superintendent. Objections or delays sometimes were filed by creditors "friendly" to the debtor. Finally, the limited processing capacity of the court system also contributed to the delays. In May 1989 there were 16 cases in which an agreement had been reached and awaited confirmation by the judge; of these, nine had been languishing for more than a year.

Decree 350, enacted in 1989, reformed the mandatory reorganization procedures. The most significant change was designation of the superintendency as the sole competent authority for mandatory cases. The superintendency was thus endowed with authority to decide on matters that had previously required the intervention of the district judge. Most important, the superintendency was granted the authority to resolve disputes arising from objections raised during the validation of credits. The superintendency was also authorized to confirm agreements reached between parties.<sup>42</sup>

Decree 350 also introduced other significant changes. For example, tight time limits were assigned to the various stages of reorganization, and many decisions of the competent authority were rendered nonappealable. The decree also allows creditors to establish various mechanisms of control and monitoring during the procedure and requires the formation of a committee of creditors consisting of representatives of all classes of creditors (including public agencies, workers, and financial and nonfinancial creditors). The appointment of an examiner of the property, credits, and affairs of the debtor is also mandated. Both the committee and the examiner are given the right to request that the competent authority remove the firm's owner from the management team.

A comprehensive evaluation of the impact of the reform is premature. However, evidence to date on mandatory procedures suggests that these reforms have led to substantial improvement (table 8.1). As can be seen, the percentage of cases in which an agreement was reached within a year increased from 18 percent to 32 percent after enactment of Decree 350. Similarly, whereas prior to reform only 52 percent of the cases resulted in an agreement within two years, this ratio increased to 60 percent under the new law.

Although no comparable data exist for the optional cases that still take place in the court system, there is consensus among professionals, lawyers, and the business community that, even though the two procedures are governed by very similar legislation, it takes much longer to reach an agreement in the optional system under the court system than in the mandatory system under the superintendency. The reasons are the following. First, relative to judges, the superintendency has more flexibility to resolve disputes. Second, the staff of the superintendency is likely to focus more on economic solutions to the problems posed by cases rather than on strict procedural problems. Third, and perhaps most important, the superintendency is being increasingly equipped with the commercial, financial, economic, and technical know-how, both to handle the various restructuring issues raised during reorganizations, and to mediate between conflicting parties to forge mutually beneficial and acceptable solutions.

### Bankruptcy policy reform in industrializing countries

The tension between satisfying the creditors' bargain and preserving going-concern value gains added significance in the context of bankruptcy reform in industrializing

TABLE 8.1  
Time to conclude a mandatory reorganization procedure under Colombia's old and new bankruptcy laws

Duration	Cases initiated under the commercial code <sup>a</sup> (percentage of total)	Cases initiated under Decree 350 <sup>b</sup> (percentage of total)
12 months or less	18	32
13-24 months	34	28
25-48 months	29	8
49 months or more	11	0
Not concluded	8	32
Total	100	100

a. Cases initiated before May 1989.

b. Cases initiated between May 1989 and December 1991.

Source: Superintendency of Companies, Bogotá.

countries. Inadequate enforcement of credit contracts, and the consequent insufficient protection of creditors' rights, is one of the important causes of the underdevelopment of financial markets. Lenders try to compensate for these deficiencies by imposing high collateral requirements or by rationing credit away from large portions of potential borrowers. Hence, reforming bankruptcy laws so as to uphold the creditors' bargain would have the beneficial effects of increasing financial intermediation and encouraging the development of financial instruments. In an environment where weaknesses in contract enforcement mechanisms already hinder the development of the financial sector, a bankruptcy system that gave excessive bargaining power to debtors would seriously jeopardize willingness to lend and exacerbate credit rationing.

However, an approach to bankruptcy that emphasizes debt collection may also lead to substantial losses in the going-concern value of debtor firms. Such losses may be especially high in countries going through significant changes in policy regimes, and where the business sector requires substantial restructuring to regain competitiveness, as in the aftermath of a trade reform. Under these circumstances bankruptcy policy should encourage reorganizations. Either way, the opportunity cost of establishing an inefficient bankruptcy system is probably higher in industrializing countries, as they face sharper trade-offs between debt collection and restructuring. The situation is further aggravated by insufficient capacity in the court system and the relative lack of workout skills, since these make it more difficult to monitor and control strategic behavior under bankruptcy.

Given these constraints, how should industrializing countries approach bankruptcy reform? First of all, bankruptcy reform should be carried out in parallel with reform of other types of regulatory policy that have an impact on bankruptcy outcomes. As suggested by the previous discussion, the supervision and regulation of the financial system, accounting and disclosure rules, and regulations that restrict the mobility of labor and capital should occupy top positions on the reform agenda. A general improvement in loan documentation and legal registries is also key. Without an improvement in the relevant regulatory environment, bankruptcy reform is bound to have only a limited effect.

Second, correcting some of the fundamental shortcomings of bankruptcy legislation is likely to improve outcomes significantly. Bankruptcy laws should distinguish clearly between the productive enterprise and its owners and managers, and decouple the fate of the first from that of the second. This step would introduce significant flexibility and

increase the number of options available to interested parties. In addition, default should be divorced from any association with fraudulent behavior. Decisions about whether to liquidate a firm or maintain it as a going concern could then be made based on economic criteria rather than on the ethical standards of its owner or manager.

As for a bankruptcy model, the question is, how can bankruptcy policy encourage reorganizations without jeopardizing debt collection? The starting point could be that, whereas the law should explicitly provide for reorganization, it should not grant the debtor excessive bargaining power. This objective would require that owners and managers relinquish control rights over the firm once it is in bankruptcy reorganization or, in a less extreme option, that creditors be allowed to request the removal of management on the basis of protecting their economic interest.

Where should decisionmaking authority be placed?<sup>43</sup> The low level of processing capacity and skills in the court system warns against a model that requires the active participation and decisionmaking power of judges. A quasi-judicial body that focused exclusively on bankruptcy reorganizations might help to expedite these procedures. As discussed above, this option was tried in Colombia and India. While reorganization procedures in India still present serious problems, Colombia seems to have benefited significantly from reform.

In addition to the fact that the regulatory environment in Colombia is much less restrictive, the essential differences between the two countries' systems—which can account for the difference in success—are the following. First, India's Board for Industrial and Financial Reconstruction makes every effort to save a company from failure. Rather than a mandate of law, this seems to be a political mandate. By contrast, the Superintendency of Companies in Colombia takes a more business-like perspective. Second, the superintendency is increasingly staffed with accountants and financial analysts, as well as lawyers. Moreover, being the supervisor of the companies, it already has detailed accounting and financial information on their performance. In India most members of the reconstruction board are career bureaucrats who have no expertise in financial matters. Third, interested parties have less veto or other power to delay the procedures in Colombia; the superintendency is quite powerful. In India, by contrast, unanimous consent is almost always required. Another important attribute of the superintendency in Colombia is that it is widely recognized, by both the banking system and industry, as an impartial and competent institution.

Similar arrangements can be made in countries in which these conditions are likely to hold and the independence of the quasi-judicial agency from political influence can be guaranteed. Representation of the private sector in the agency may help to preserve independence.

Another possibility is to establish specialized commercial or even bankruptcy courts. But this may be a luxury for many countries, especially those whose judicial systems are already strained and underfinanced. Overall judicial reform is likely to be necessary in such countries, and bankruptcy may have a low priority.

In cases in which administrative or quasi-judicial solutions are not feasible, it would be preferable to grant greater control powers to banks, or creditors in general, rather than to debtors for two main reasons. First, if judicial capacity is low, it would be difficult for judges to control the strategic behavior of debtors with substantial bargaining power. As a result, procedures might be delayed for a long time, as the prereform experience in Colombia shows. As discussed earlier, such delay is likely to lead to substantial losses in the value of assets under bankruptcy and to result in outcomes that are inefficient from both debt collection and restructuring points of view. Second, however underdeveloped, banks are likely to possess the strongest appraisal and workout skills in the economy.

Finally, private banks in many industrializing countries are owned by financial-industrial business groups. This raises the possibility that in addition to debt collection, banks might also be motivated to use bankruptcy for anticompetitive purposes—for example, to drive from the market competitors of industrial members of their business group. Prevention of such predatory behavior would require an increase in competition in the banking system as well as effective application of competition policies.

### Topics for future research

Research on the efficiency properties of different bankruptcy models is still in its infancy, and more analytical and empirical work is required to better inform policymaking in this area. Research on the bankruptcy laws of industrializing countries that have not yet undertaken reforms would provide further insights on the main problems of the legislation, in particular how private agents have responded to the shortcomings of the legislation. Are informal workouts widespread? How common is private arbitration? Or is the main consequence of the inadequacies of the legislation a low level of financial intermediation?

Analyses of bankruptcies in industrializing countries that have undertaken reforms are needed in order to understand which models work better under developing-country conditions. In addition to analyses of the legislation, detailed studies of individual cases, as well as statistical analyses of the characteristics of firms that enter bankruptcy, would be useful. In this regard, the pattern of recovery rates across various stakeholders, time spent under bankruptcy, and firm performance following reorganization are empirical areas that deserve special attention.

Regarding industrial countries, empirical work—which to date has concentrated mainly on the United States—should be expanded to cover other bankruptcy models, especially those in Europe. For the United Kingdom, it would be interesting to examine companies under administrative receiverships and compare cases where the company is preserved as a going concern or sold with little fragmentation of its assets with those where the company is sold piecemeal. Is it the conflict of interest between secured creditors and other stakeholders (especially unsecured creditors) that accounts for the differences? Further work on administrations would also be helpful. Are most administrations preempted by the actions of secured creditors? Finally, in-depth studies of large cases may reveal whether liquidations do indeed result in a loss of going-concern value. More information on the extent of informal workouts in countries other than the United States would be valuable.

For all industrial countries, more information on the postbankruptcy performance of companies under different bankruptcy models is needed. Assuming that a higher rate of failure after bankruptcy reorganization reflects deferred liquidations, a comparison of postbankruptcy survival and failure rates would provide an important indicator of the relative efficiency of different bankruptcy laws.

Little attention has been devoted to whether early detection of financial distress can help prevent further financial deterioration and thus avoid costly bankruptcy. More detailed examination of the French and U.K. laws, and their early-warning mechanisms, would be very informative in that respect. If early detection is desirable, what type of rules would best encourage voluntary cooperative behavior to resolve financial distress before bankruptcy becomes inevitable? Does public policy have a useful role in early detection?

Recourse to bankruptcy is much less frequent in Japan and Germany than in the United Kingdom and the United States, especially for large firms. It is generally

believed that in these countries, the banking system takes a much more active role in recontracting firms out of financial distress and acquires significant control rights over firms' operations during the process. Moreover, banks also provide considerable expertise to help a turnaround. It would be interesting to know, first of all, how these workouts compare to those in the United States. Do the debtors have more or less bargaining power (for example, are violations of the absolute priority rule more or less widespread)? How do the incumbent management and nonbank creditors of the companies fare in these workouts?

In general, the role and fate of incumbent management in bankruptcies and workouts is another area that deserves investigation. There seems to be a consensus, at least in the popular press, that the removal of incumbent management at the time of a bankruptcy filing should be facilitated in the United States. The argument is that if incumbent management had something valuable to contribute to the firm, it would be hired by the creditors. There is some evidence on management turnover during periods of financial distress and under bankruptcy in the United States. But the high turnover rates found so far contradict the popular and academic presumption, as well as the empirical evidence, that debtors have substantial bargaining power in bankruptcy reorganizations. More empirical work is thus warranted for the United States and other countries.

### Notes

1. For a review of these policies in a sample of OECD countries, see Atiyas and others 1992.
2. The only exceptions are Franks and Torous 1992, Franks, Nyborg, and Torous forthcoming, Mitchell 1990, and White n.d.
3. For a brief review of debt collection outside bankruptcy, see Baird and Jackson 1985.
4. Note that the issue here is not only that the going-concern value of the firm may be higher than its liquidation value. Even in cases in which liquidation is the most efficient action, assets may still be worth more when sold in packages rather than individually.
5. Jackson (1986) mentioned two additional problems associated with individual mechanisms of debt collection. Knowing that to be paid in full necessitates acting before other creditors, each creditor is likely to spend more resources monitoring the debtor and other creditors. Presumably, some of this expenditure is wasteful and could be avoided if debt collection were a collective process. The second problem has to do with attitudes toward risk. Suppose all creditors are unsecured. Then one can think of cases in which the expected value of what each creditor would receive under a first-come, first-served mechanism of debt collection would be equal to

what each would receive under a collective mechanism that divided the assets available for distribution in proportion to each creditor's claims. If lenders are risk-averse, however, then the expected utility of (certain) payments under the collective mechanism would be higher than that of (uncertain) payments under the individual mechanism.

6. Jackson (1986) described this as the requirement that bankruptcy should preserve the *relative*, as opposed to the *absolute*, values of claims.

7. The classic expositions of the problems of under- and overinvestment are found in Myers 1977 and Jensen and Meckling 1976, respectively.

8. This is the criterion most widely used to assess the efficiency of bankruptcy rules. See, for example, White 1989.

9. Note that in the context of recontracting, the creditors' bargain would require that the rules governing renegotiations not grant any bargaining power to any of the claim holders that was not envisaged in the original distribution of claims.

10. For a brief history, see *International Financial Law Review* 1990. White 1984 presents a detailed comparison of the two laws.

11. If the incumbent management engages in fraud or is deemed incompetent, the court may appoint a trustee to assume management. Such appointments are rare, however, and not easily granted by courts (*International Financial Law Review* 1990, p. 54).

12. The transfers that can be voided include those made in preference to some creditors within 60 days prior to the filing of the bankruptcy petition (that is, if the transfer enables the creditor to receive more than it would have if the debtors' estate were being liquidated) and fraudulent transfers occurring within one year prior to the petition in which the debtor receives less than equivalent value in exchange for such transfer and was either insolvent or rendered insolvent by that transaction.

13. For example, if entering bankruptcy damages the firm's reputation, the firm will have to spend more resources to convince trading partners to continue their business, and trading partners will require more advantageous terms. Similarly, the loss of consumer confidence may require management to sell products and services at a lower price.

14. In a slightly different interpretation, Franks and Torous (1989) suggested that violations of the absolute priority rule may reflect the creditors' purchase of the shareholders' option to delay the proceedings.

15. In the rest of the cases, the firms either were liquidated or merged into other firms.

16. Two features of Chapter 11 may partly compensate for its cost disadvantages. The first is the automatic stay, which prevents costly fragmentation of the firm's assets. The second is availability of debtor-in-possession financing, which may help preserve the firm's value during negotiations. However, the net effect of these provisions depends on the nature of agency problems, as discussed below.

17. See Gertner and Scharfstein 1991 for a theoretical treatment of this problem in informal workouts.
18. Gilson, John, and Lang (1990) mention asymmetric information as a second problem that may cause the failure of informal workouts. Insiders normally have better information about the true value of the firm than creditors. Shareholders or management have an incentive to use this advantage to misrepresent the value of the firm and thus gain more favorable terms in the restructuring. Since rational creditors are aware of insiders' incentives, the asymmetry in information may cause renegotiations to fail. Gilson, John, and Lang argue that the insiders' information advantage in Chapter 11 is much smaller due to the disclosure requirement.
19. As a result, restructurings of publicly held bonds are often done through exchange offers, in which the firm offers cash and a package of debt and equity securities in exchange for the existing bonds. To avoid a hold-out problem and encourage bondholders to participate in the exchange, the firm often includes more senior bonds in the package. Moreover, bondholders are asked to eliminate the protective covenants of the old bond; hence, the offer is often contingent on acceptance by a specified majority of bondholders.
20. This going-concern value is probably due to intangible assets. The presence of intangibles, which may be lost in a Chapter 11 proceeding, may also discourage junior creditors from holding out.
21. The informal workout of Donald Trump provides a recent example. According to news articles, the creditors were convinced by their counsel that a Chapter 11 case would be a drawn out, conflictual, costly process. Under the threat of that prospect, Trump was able to get an extremely favorable agreement in the informal workout (*Washington Post*, November 29, 1992).
22. See Gertner and Scharfstein 1991 for a formalization of this linkage between informal workouts and formal organizations.
23. See Graham 1992 and Kallen 1991 for more examples.
24. This section relies on Clarke 1993, Franks and Torous 1992, Rajak 1988, Webb 1991, and the *International Financial Law Review* 1990.
25. The U.K act was based on the recommendations of the Report of the Review Committee on Insolvency Law and Practice (1982).
26. Franks and Torous (1992) reported that as of September 1991 more than 1,000 directors had been disqualified.
27. A fixed charge is a security over a particular asset, whereas a floating charge gives security over a pool of assets or over the whole company. Under a floating charge, assets of the company may be used freely by the managers unless and until the charge "crystallizes" due to, for example, a default. On crystallization, it becomes a fixed charge over the relevant assets.
28. The benefit derived from the automatic stay provision of administration is one incentive for a creditor to opt for this alternative rather than receivership.
29. This section draws primarily on Biais 1994, *International Financial Law Review* 1990, Malecot 1992, Mitchell 1990, and Simeon and others 1990.
30. Under the simplified procedure, the observation period is shorter and the appointment of an administrator is not mandatory.
31. Under the previous law, a certain percentage of creditors had to approve a rehabilitation plan. Under the current law, a court may adopt a plan even if all the parties object to it.
32. More specific measures of a similar spirit have been proposed by LoPucki (1993). He proposes that (a) the judge eliminate the interests of *insolvent* shareholders, and (b) the judges' discretion to extend the exclusive period be curtailed.
33. *The Economist* ("When firms go bust," August 1, 1992, p. 63-65) calls this approach to reform "beefing up the bureaucracy."
34. For example, transfer of ownership to secured creditors with floating charges would presumably make them interested not in realizing their security, but in maximizing the value of the firm. In this way, asset fragmentation would be avoided and going-concern value would be preserved.
35. Aghion, Hart, and Moore (1993) note that this transfer of wealth from junior to senior debt would be compensated ex ante by more favorable pricing of junior debt.
36. In many countries, the distinction between owners and managers is not pronounced. Most companies are family-owned and managed. Diffuse ownership is rare. Even large companies, which may employ professional staff in higher-level management positions, are tightly controlled by owners.
37. Deposit insurance is implicit whenever depositors expect to be bailed out by the state in the event of bank failure, even if no explicit legal scheme protects depositors' claims on the bank.
38. Mitchell 1993 provides a comprehensive discussion of various factors that may induce "creditor passivity" in the context of the formerly socialist economies of Europe.
39. Goswami and others 1993 and Anant and others 1994 offer vivid descriptions of the implications of these problems for reorganizations in India.
40. This discussion draws principally on Goswami and others 1993.
41. These were firms that employed more than 100 permanent workers, firms whose foreign liabilities exceeded one-third of the value of assets, or those in which the government had a majority interest.
42. Granting judicial powers to an administrative authority created a controversy that was resolved with the adoption of a new constitution enabling an administrative authority to assume the functions of a judge.
43. The term "decisionmaking authority" refers to the authority to decide whether the company can be rehabilitated, to prepare a reorganization plan, and to impose a plan despite dissenters. A judge can exercise this authority directly, or indirectly by appointing and monitoring an administrator or a trustee.

## References

- Aghion, Phillipe, Oliver Hart, and John Moore. 1993. "The Economics of Bankruptcy Reform." *Journal of Law, Economics, and Organization* 8(3): 523–46.
- Anant, T. C. A., Tamal Dhatta Chaudhury, Shubhashis Gangopadhyay, and Omakar Goswami. 1994. "Industrial Sickness in India: Institutional Responses and Issues in Restructuring." Indian Statistical Institute, New Delhi.
- Atiyas, Izak. Forthcoming. "Exit and Restructuring Policies in Turkey." In R. Erzan, ed., *Competition Policies for Turkey*. New York: Macmillan Press.
- . 1994. "Restructuring Programs in Transitional Economies." In Vedat Milor, ed., *Changing Political Economies: Privatization in Post-Communist and Reforming Communist States*. Boulder and London: Lynne Rienner Publishers.
- Atiyas, Izak, Mark Dutz, and Claudio Frischtak, with B. Hadjimichael. 1992. "Fundamental Issues and Policy Approaches in Industrial Restructuring." *Industry Series Paper* 56. World Bank, Industry and Energy Department, Washington, D.C.
- Baird, Douglas. 1986. "The Uneasy Case for Corporate Reorganizations." *Journal of Legal Studies* 15(1): 127–47.
- Baird, Douglas G., and Thomas H. Jackson. 1985. *Cases, Problems, and Materials on Bankruptcy*. Boston: Little, Brown and Company.
- . 1988. "Bargaining After the Fall and the Contours of the Absolute Priority Rule." *University of Chicago Law Review* 55(3): 738–89.
- Biais, B. 1994. "An Economic Analysis of the French Bankruptcy Law." Private Sector Development Occasional Paper, Vice Presidency of Finance and Private Sector Development, World Bank, Washington, D.C.
- Clarke, Allison. 1993. "Company Rehabilitation and Reorganization in the United Kingdom after the 1986 Insolvency Act." In A. C. Budak, ed., *Symposium on Corporate Financial Distress and Relation with Banks in the Turkish, UK, and United States Law*. Istanbul: Istanbul Chamber of Industry.
- Eberhart, Allan C., and Lemma Senbet. 1993. "Absolute Priority Rule Violations and Risk Incentives for Financially Distressed Firms." *Financial Management* 22(1): 101–16.
- Eberhart, Allan C., William T. Moore, and Rodney L. Roenfeldt. 1990. "Security Pricing and Deviations from the Absolute Priority Rule in Bankruptcy Proceedings." *Journal of Finance* 45(5): 1457–69.
- Franks, Julian R., and Walter Torous. 1989. "An Empirical Investigation of U.S. Firms in Chapter 11 Reorganization." *Journal of Finance* 44(3): 747–67.
- . 1992. "Lessons from a Comparison of the United States and UK Insolvency Codes." *Oxford Review of Economic Policy* 8(3): 70–81.
- . 1993. "A Comparison of Financial Recontracting in Distressed Exchanges and Chapter 11 Reorganizations." University of California-Los Angeles, Department of Economics.
- Franks, Julian, Kjell Nyborg, and Walter Torous. Forthcoming. "A Comparison of US, UK, and German Insolvency Codes." Private Sector Development Occasional Paper, Vice Presidency of Finance and Private Sector Development, World Bank, Washington, D.C.
- Gertner, Robert, and David Scharfstein. 1991. "A Theory of Workouts and the Effects of Reorganization Law." *Journal of Finance* 46(4): 1189–1222.
- Gilson, Stuart C. 1989. "Management Turnover and Financial Distress." *Journal of Financial Economics* 25(2): 241–62.
- . 1990. "Bankruptcy, Boards, Banks, and Blockholders." *Journal of Financial Economics* 26(2): 355–87.
- Gilson, Stuart C., Kose John, and Larry Lang. 1990. "Troubled Debt Restructurings: An Empirical Study of Private Reorganization of Firms in Default." *Journal of Financial Economics* 26(2): 315–53.
- Goswami, Okmar, and others. 1993. "Report of the Committee on Industrial Sickness and Corporate Restructuring: Submitted to the Union Minister of Finance, Government of India." Indian Statistical Institute, New Delhi.
- Graham, Mary. 1992. "Business: Bankrupt and Bullish." *Atlantic Monthly* 269(3): 24.
- Hotchkiss, Edith. 1992. "The Post-Bankruptcy Performance of Firms Emerging from Chapter 11." New York University, Department of Economics, New York.
- International Financial Law Review*. 1990. "Solving the Insoluble: A Legal Guide to Insolvency Regulations around the World." Special Supplement. June.
- Jackson, Thomas H. 1982. "Bankruptcy, Non-Bankruptcy Entitlements, and the Creditors' Bargain." *Yale Law Journal* 91(5): 857–907.
- . 1986. *The Logic and Limits of Bankruptcy Law*. Cambridge, Mass.: Harvard University Press.
- Jensen, Michael C., and W. Meckling. 1976. "Theory of the Firm: Managerial Behavior, Agency Costs and Capital Structure." *Journal of Financial Economics* 3(3): 305–60.
- John, Kose. 1993. "Managing Financial Distress and Valuing Distressed Securities: A Survey and Research Agenda." *Financial Management* 2(1): 60–78.
- Kallen, Laurence H. 1991. *Corporate Welfare: The Megabankruptcies of the 80s and 90s*. Seacaucus, N.J.: Carol Publishing Group.
- LoPucki, Lynn M. 1983a. "The Debtor in Full Control—Systems Failure Under Chapter 11 of the Bankruptcy Code: First Installment." *American Bankruptcy Law Journal* 57(1): 99–126.
- LoPucki, Lynn M. 1983b. "The Debtor in Full Control—Systems Failure Under Chapter 11 of the Bankruptcy Code: Second Installment." *American Bankruptcy Law Journal* 57(2): 247–73.

- LoPucki, Lynn M. 1993. "The Trouble with Chapter 11." *Wisconsin Law Review* 4(3): 729–60.
- Malecot, J. F. 1992. "The Over- and Under-Investment Incentives Under French 1985 Bankruptcy Law." University of Paris-Nanterre.
- Mitchell, Janet. 1990. "The Economics of Bankruptcy in Reforming Socialist Economies." Report to the National Council for Soviet and East European Research. Cornell University, Department of Economics, Ithaca.
- . 1993. "Creditor Passivity and Bankruptcy: Implications for Economic Reform." In Colin Mayer and Xavier Vives, eds., *Financial Intermediation in the Construction of Eastern Europe*. Cambridge: Cambridge University Press.
- Myers, S. 1977. "The Determinants of Corporate Borrowing." *Journal of Financial Economics* 14(1): 147–75.
- Rajak, Harry. 1988. *Company Liquidations*. Oxfordshire: CCH Editions.
- Shleifer, Andrei, and Robert Vishny. 1993. "Liquidation Values and Debt Capacity: A Market Equilibrium Approach." *Journal of Finance* 47(4): 1343–66.
- Simeon, Moquet Borde, and others. 1990. *Doing Business in France*. Vol. 2. New York: Matthew Bender, Times Mirror Books.
- van Wijnbergen, S. 1992. "Economic Aspects of Enterprise Reform in Eastern Europe." World Bank, Central Europe Department, Washington, D.C.
- Webb, David C. 1991. "An Economic Evaluation of Insolvency Procedures in the United Kingdom: Does the 1986 Insolvency Act Satisfy the Creditors' Bargain?" *Oxford Economic Papers* 43(1): 139–57.
- Weiss, Laurence A. 1990. "Bankruptcy Resolution: Direct Costs and Violation of Priority of Claims." *Journal of Financial Economics* 27(2): 285–314.
- Westbrook, J. L. 1993. "Chapter 11 Reorganizations in the United States." In A. C. Budak, ed., *Symposium on Corporate Financial Distress and Relation with Banks in the Turkish, UK and United States Law*. Istanbul: Istanbul Chamber of Industry.
- White, Michelle J. 1989. "The Corporate Bankruptcy Decision." *Journal of Economic Perspectives* 3(1): 129–62.
- . 1983. "Bankruptcy Costs and the New Bankruptcy Code." *Journal of Finance* 38(3): 477–87.
- . 1984. "Bankruptcy Liquidation and Reorganization." In D. Logue, ed., *Handbook of Modern Finance*. Boston: Warren, Gorham and Lamont.
- . (n.d.) "The Costs of Corporate Bankruptcy: A U.S.–European Comparison." Law and Economics Working Paper Series no. 1. University of Michigan, Department of Economics, Ann Arbor.

# Labor policies and regulatory regimes

Zafiris Tzannatos

Labor regulations can change the price of labor (by imposing minimum wages or subsidized employment schemes), its quantity (by prohibiting child labor or setting restrictions on cross-country worker mobility), and its quality (by establishing occupational certification and licensing or industrial health and safety standards). In addition, regulation can affect the way labor is exchanged in the market, through job security rules, trade union laws, and dispute resolution provisions. Labor exchange is also affected by other, non-labor market policies, such as social assistance that provides income support to the nonworking poor. Such policies alter the willingness of workers to supply labor by changing their reservation wage and can increase employer costs (depending on the way they are financed). Finally, industrial and trade policy affects the allocation of labor and wages both in specifically targeted and other sectors of the economy.

The objective of this chapter is to discuss some of the main issues of labor regulation. The next section shows why labor is not a conventional commodity, how the unfettered workings of the labor market may fail to produce efficient outcomes, and why governments may wish to intervene to improve the allocative and distributive results. It also shows the different labor market outcomes that arise from freely negotiated contracts compared with regulated ones. The third section provides a typology of labor market interventions and examines specific labor market regulations in areas such as trade unions, dispute resolution, unemployment insurance, minimum wages, and job security. The fourth section examines the case of the dock industry in Britain as an illustration of the impact of regulation and deregulation on labor markets, discussed from a partial equilibrium perspective. The fifth section looks at economywide effects of labor market deregulation in the United Kingdom, the industrial country that experienced the most radical reform in industrial relations in recent years, and in developing countries that went through adjustment and trade liber-

alization in the 1980s. The final section summarizes the discussion.

The chapter's two principal messages are, first, that there are neither easy answers nor hard-and-fast rules relating to the overall coverage and optimal degree of labor market regulation (that is, policy *design*). Similarly, it is far from trivial to assess the desirability and effects of economywide labor market deregulation. The range of potential labor market policies is wide and their interactions with other markets complex. It is therefore difficult to anticipate or even assess the effects of regulation or deregulation amidst other interventions or market failures.

Second, the effects of policy *reform* in specific areas (in terms of industries, occupations, and so on) are more predictable because they can be assessed through partial equilibrium analysis. The change in the U.K. dock industry is a case in point. However, a partial equilibrium perspective may lead to outcomes that fall short of a global optimum.

For reasons of brevity the chapter has been written mainly from the perspective of regulation—what, when, and how much to regulate. In short, it argues that there are conditions under which regulation is socially beneficial. Had it been written from the perspective of deregulation, the message would have been the same: When the conditions that originally justified an intervention either have changed or no longer apply, regulation should be amended or repealed.

## Rationale for labor market interventions

Government interventions in labor markets are precipitated by dissatisfaction with the technical, allocative, or distributional consequences of their unrestricted operation. Technical inefficiency arises when the economy is not at its production frontier (for example, because workers are underutilized in their current jobs due to information failure about more suitable jobs). Allocative

inefficiency arises when the product mix (quantities) does not reflect societal preferences (relative prices), as in the case of monopoly or monopsony. A third kind of inefficiency, usually missing from the debate, arises when the initial distribution of endowments prevents the economy from maximizing social welfare. Thus, a reallocation of resources can improve society's welfare, as in the case of diverting resources to educate the poor.

The precise motivation behind, and objectives of, particular labor market regulation (and deregulation) episodes are in general complex because they reflect not only economic factors but also historical, political, social, and cultural elements. Governments may attempt to deregulate if the beneficiary effects of enhancing competition (such as structural adjustment policies) are hindered by imperfections in the labor market. Governments may introduce regulations as a means of achieving a more equitable pattern of distribution between individual members of, or groups within, society. Minimum wage legislation is perhaps the most widely implemented regulatory policy under this heading. The following situations, among others, may give rise to, and provide a rationale for, labor market intervention.

- *Regulation can rest on the belief that the market for labor, if left alone, will fail to function efficiently because of imperfections on either the demand or the supply side.* First, labor is a factor of production, and its demand is covered by Marshall's laws of derived demand (Marshall 1890). This means that labor's price (wage) and utilization (employment) depend on output demand (elasticity of demand for the final product), production technology (ease of substitution between labor and other factors), the availability of factors other than labor (their elasticity of supply), and the structure of costs (in particular, the share of labor in total costs). Thus, even competitive labor markets may not result in an efficient allocation of labor—the prime role they are envisaged to play—if there are imperfections elsewhere in the economy: The economy would not necessarily achieve higher allocative efficiency if interventions corrected market failures in one market but not in others (see Lipsey and Lancaster 1956 on the theory of the second best).

Second, on the labor supply side and again with reference to Marshall, labor has four distinct peculiarities: Labor is not sold but only hired (the property right remains with the seller); its suppliers are at a disadvantage in bargaining; it cannot be stored; and investments in labor have a long gestation period. These peculiarities can lead to inefficient "spot" equilibria in the labor market due to the presence of monopoly or monopsony

power, myopia (the inability of individuals to see far in the future), "irrationality" (short-run considerations may lead to discounting future costs and benefits at different rates than those that would maximize welfare in the long run), and credit market failure.

Expressed differently, labor markets work efficiently only when product markets, from which demand for labor is derived, operate in a competitive fashion (there are no monopolies); when employers do not collude (there is no monopsony power); when workers are concerned with themselves only (there are no unions); if the economy is an open one (a price taker in the world market); and if there is no economic rent to be appropriated by employers' coalitions or trade unions. Finally, labor is the most politicized factor of production, and labor transactions are more relational and conflictual than are voluntary exchanges between equals.

- *Interventions can also be introduced if the level of unemployment consistent with demand–supply equilibrium in the labor market is judged to be unacceptably high, and income support and re-employment assistance to the unemployed is deemed to be desirable.* The precise definition of what constitutes an "unacceptably high" level of unemployment varies and is partly determined by the socially acceptable living standards gap between the unemployed and the employed, and the willingness to publicly finance unemployment insurance. Interventions may also be desirable (for example, in the form of subsidized retraining) to ensure that society's human capital stock suffers minimum depreciation during the down-phase of the cycle and until recovery resumes. However, it is relevant to recognize here that labor market regulation is not the only (or necessarily the most efficient) means of achieving these particular objectives. When left to their own devices, employers and employees can arrive at implicit contracts addressing the problem of cyclical fluctuations in employment (Manning 1990). In essence, this solution involves the employer's taking on the role of an insurance company that provides the worker with coverage against the possibility of becoming unemployed at some future date, with the premium for this insurance taking the form of a somewhat lower wage than the employee would otherwise have earned. The important point here is that implicit contracts, although hard to be confirmed in practice due to their nature, represent a mechanism whereby labor market actors arrive themselves at a nonregulatory solution.

- *Missing markets provide a further reason for labor market regulation.* The fact that the market for information may be incomplete frequently prompts governments to pro-

vide both employers and workers with information, in many cases at little or no cost. An obvious example is the provision of information relating to job vacancies as a means of improving the efficiency of job searches.

- *In the presence of asymmetric information, insurance companies are unable to distinguish between good and bad risks.* The state can use its power to require all workers to buy unemployment, accident, and health insurance with a specific level of coverage (what is commonly described as a “pooling equilibrium”; Rothschild and Stiglitz 1976). This action can lead to the adoption of a premium rate that is the average of the fair rates for the different risk groups and can prevent the good risks from being driven out by the bad ones. Whether such a pooling equilibrium is better for the good risks than a separating equilibrium will depend on the circumstances. The good risks will pay a higher premium rate than is actuarially fair for them, but they will be able to buy more coverage in a pooling equilibrium than in a separating equilibrium. If bad risks are sufficiently risk-averse to care more about the extent of coverage than the higher premium rate, by contrast, they will pay a lower premium rate than is actuarially fair for them, and they will be unambiguously better off. The existence of adverse selection necessarily leads to some labor market inefficiency. The enforced pooling equilibrium implies that high-risk employment is being subsidized at the expense of low-risk employment. Also, the fixed level of coverage implies that some workers will have more insurance coverage than they want. Whether the effect of those with more coverage than they want (who will be attracted into riskier occupations) balances the effect of those with less coverage than they want (who will be attracted into less risky occupations) depends on the level of fixed coverage chosen.

- *Paternalism can correct for irrationality.* Some people do not make adequate provision for their retirement or for labor market risks such as unemployment and industrial injury. The state, therefore, may adopt the paternalistic policy of requiring individuals to purchase more insurance coverage than they would otherwise have chosen. This argument therefore complements that derived from considerations of adverse selection in supporting a compulsory system of labor market insurance. People who are forced to buy more insurance than they want, either because of paternalism or forced pooling equilibrium, will treat the additional insurance contributions as a tax and can therefore be expected to reduce their labor supply. This means that the advantage of financing benefits from compulsory earmarked contributions is reduced by the presence of involuntary participants.

- *Labor market regulation may be a necessary component for the achievement of wider social objectives.* The Charter of the Fundamental Social Rights of Workers (the “Social Charter” adopted in December 1989 by all members of the European Union except the United Kingdom) is a case in point. The implementation of this charter was seen by many as a logical extension of the European Single Market initiative, designed to prevent unfair competition and “social dumping” between member countries and to promote social integration throughout the European Union. Although concerned with wider issues relating to social policy, this charter lays down regulations covering a host of issues including workplace health and safety, maximum hours of work, the rights of part-time and subcontracted workers, training entitlements, and consultation requirements.

By their very nature, social contracts such as the charter (especially if the benefits are heavily weighted to those with full-time or permanent employment) raise potential conflicts between the gainers in society and those who either gain less (for example, part-time workers and the unemployed) or those who meet the costs of the provisions, the taxpayers at large.

- *Individually negotiated contracts can lead to inefficiency compared with regulated contracts.* The difference between individual and regulated contracts can be illustrated by the right of employers to terminate freely a worker’s employment even without cause—a practice that in the United States has been labeled “employment at will.” According to this principle, the firm could at any time and at will dismiss a worker “for good cause, for bad cause, or even for cause morally wrong” (Dertouzos and Karoly 1992). The freedom of U.S. employers to dismiss workers has been gradually curtailed since the 1980s with the introduction of restrictions on employers’ ability to dismiss without showing just cause. In theory, the shift away from the employment-at-will principle should reduce flexibility and induce inefficiency. However, employment-at-will can be efficient only if the following conditions hold: agents are infinitely rational; wages are not affected by norms of fairness; contracting costs are low; dismissal policies of one firm do not have an effect on other firms; there are no other externalities of dismissals (for example, in the form of unemployment benefits paid by the rest of the society); and possible injustices involved in dismissals are of no concern to society. If these conditions do not hold, then it is employment-at-will, not an appropriately designed just-cause policy, that would lead to inefficiency (Levine 1991).

Labor regulation that restricts individual contracting can also increase efficiency by reducing uncertainty. For

example, in the case of employment-at-will, regulation can reduce the employer's ability to engage in capricious behavior, enabling management and labor to develop a long-run employment relationship and reap the mutual benefits of this cooperation. Against this benefit, one must examine the costs of reducing employers' ability to freely adjust their work force.

- *Regulation can create equal opportunities, ensure equal treatment, and reduce discrimination, thereby increasing the allocative and distributional efficiency of markets.* Regulation can prescribe not only "general" rights for individual workers but also "specific" rights for certain groups. The case of general versus specific rights can be clarified with reference to female workers. Many countries have laws that establish the general right of women to be treated in the same way as equally qualified men. The objective of such legislation is to remove discrimination on the demand side of the labor market (Zabalza and Tzannatos 1985). However, granting women the same general rights as men may not be sufficient to ensure that women attain the same labor market outcomes as men. Women face the additional sex-specific constraints on the labor supply side arising from childbearing and, more generally, reproduction. Thus, women are granted (to various degrees) special rights through maternity legislation—such as leave during pregnancy and around the time of birth, which can be accompanied by guaranteed reentry to their previous job. In some cases legislation may not only afford women equality in opportunity but also aim at achieving equal (or more equal) outcomes through affirmative action provisions (Faundez 1994).

### **A typology of labor market interventions**

There is no generally agreed categorization of labor market interventions. Some interventions affect the institutional process of labor exchange, whereas others more directly affect labor demand or labor supply. Another way to look at interventions is by examining their impact on efficiency or equity. Finally, a popular distinction of labor market policies has been between "active" and "passive" policies (sometimes called "proactive" or "reactive"; OECD 1992). Although there are no precise criteria for including a particular policy under either of these descriptive titles, active policies typically include measures that enhance the allocative role of the labor market (such as information gathering and dissemination, retraining, placement of retrenched workers, and so on); passive policies aim primarily at reducing hardship to workers (such as the payment of unemployment benefits). However, the distinction is not at all clear. For example,

benefit payments can increase efficiency to the extent that the unemployed can afford to search longer for better-paying (higher-productivity) jobs, and public works combine income support with the creation of useful infrastructure.

The following list, though neither exhaustive nor mutually exclusive, indicates the main areas that have been subject to regulatory control:

- *Labor rights and standards.* The asymmetry in bargaining between buyers and sellers has been addressed by governments through establishing some minimum protection for individual or groups of workers. Such asymmetry is typically seen in terms of an unequal balance of bargaining power between the buyers and sellers of labor. To redress these asymmetries, legislation can establish minimum labor rights with respect to the freedom of association, organization, and collective bargaining. In many instances, national legislation is supplemented by international (International Labor Organization) conventions on labor standards such as the minimum age for work, especially in certain activities; prohibition of forced labor; maximum hours of work in a day or week; minimum health and safety standards, and so on. The extent of rights and level of standards are not rigidly defined in the international conventions, whereas the relevant provisions and enforcement vary between countries and also within countries at different points in time. For example, international conventions do not make clear whether the right to organize refers to secret ballots or secondary picketing, whether the requirement for majority-based decisions means the total work force or simply the workers who happened to vote in that particular occasion, or whether the cutoff age for child labor is 10 or 15 years of age.

- *Conflict resolution, intermediation, and information.* The government can create and enforce rules within which conflicts between employers and workers should be solved by direct mediation in labor disputes or by providing services to improve information flow between the affected parties. It can also offer job placement services for workers. Of course, there are equally valid counterarguments to such activities: Each can be undertaken voluntarily, market failures can have less damaging effects than government failures, and there is no reason why information is more efficiently or better supplied by the government than by the affected agents and parties.

- *Direct wage and benefit interventions.* Minimum wages, wage subsidies, and payroll taxes are prime examples of wage interventions. Benefits can include bonuses, supplements, family or housing allowances, and training and education entitlements.

- *Job security legislation.* This refers to a host of provisions that protect the worker's employment status either directly (for example, by prohibiting termination of employment or reassignment) or indirectly (by imposing penalties on employers that increase the costs of termination and thus reduce the "demand" for employment changes). Regulation includes measures such as severance pay and advance notice for dismissal. Again, the range of job security legislation varies, from the constitutional provision of the right to work in the formerly planned economies to (employers') employment at will in the United States.

- *Social insurance and assistance.* This includes insurance for old age (workers' pensions), disability (temporary and permanent), unemployment (benefits paid while out of work), workers' accidents (compensation for industrial injuries and occupational hazards), health (to workers and their families), and maternity.

Underlying these areas of regulation is a wide range of specific policies and provisions. For example, with reference to social insurance and assistance policies, workers' pensions can be individually or pay-as-you-go funded, privately or publicly managed, benefit- or contribution-defined, awarded at the age of 55 or 65, and so on (World Bank 1994). Which "regulation" is best is hard to tell without taking into account prevailing norms (should all people above some threshold age be entitled to income support irrespective of whether they have paid any contributions?), social factors (are families altruistic or, in fact, relevant for intrahousehold transfers to any significant extent?), and political considerations.

The remaining part of this section examines selected areas of labor market regulation that include trade unions, whose objective to secure (additional) pay and employment can benefit their members but affects the general allocation of labor in the economy; dispute resolution, which affects unions as well as labor market outcomes; and wage coordination mechanisms. The discussion then moves to unemployment benefits, which operate through the supply side of the labor market; minimum wages, which affect labor demand as well as pay and employment outcomes for workers at the low end of the skills distribution and thus have a bearing on social considerations; and job security regulations, which affect labor demand and have elements of protecting the segment of the labor force that is prone to employment fluctuations. The concluding section draws attention to the fact that regulation, whose enforcement is typically confined within national boundaries, can increasingly become outdated with globalization and should be either amended to take into account inter-

national aspects of production (such as imports or migration) or coordinated at the international level.

### *Unions*

Unions arise from the asymmetry in contracting between individual workers and employers, the concern for basic labor rights, and the different perceptions about the merits of employment relations governed by individual contracts or collective agreements. The basic question to be answered is: Are unions good or bad, and should they be encouraged or discouraged?

Economic analysis usually assumes that the alternative to a unionized labor market is one characterized by the atomistic, perfectly competitive structure that ensures that markets clear, with individual workers choosing whether to work by comparing the given, perfectly competitive wage with the marginal utility of leisure (nonmarket activity). However, the removal of unions may reveal market imperfections on the labor demand side in the form of monopsony. Hence, policy decisions that have as their central objective the "return" to a perfectly competitive labor market (with all its well-known potential benefits) can succeed only if they are accompanied by policies designed to free up the demand side of the market. Indeed, the presence of unions in such circumstances may offer a second-best alternative to free competition. In this case, the countervailing influence of unions may result in a set of outcomes closer to the competitive equilibrium than that offered by competition on the supply side of the labor market and monopsony on the demand side.

The potential benefits associated with the presence of unions in the form of "voice" (empowerment) as opposed to "exit" (separation) effects should be seen against their costs (in the form of welfare losses due to misallocation of resources). In general, these effects have been found to be small and of comparable magnitude to the dead-weight loss arising from monopolies in product markets, which is typically less than 1 percent of total product (Rees 1963; Johnson and Mieszkowski 1970; DeFina 1983; Freeman and Medoff 1984; Pencavel 1991). However, even these low estimates may overstate the allocative loss from unions because they assume that employment is determined by a static demand curve of labor. It is nonetheless possible that labor contracts are not on the demand curve but on the firm's iso-profit curve (as suggested by the efficient bargaining models; McDonald and Solow 1981), and that unions and collective bargaining in general can have beneficial effects on the productivity of their members (the so-called second face of unionism; Freeman and Medoff 1979).

*Dispute resolution*

The breakdown of negotiations between individual workers and their employers can take various forms ranging from poor relations at the workplace (with potential costs including decreased levels of labor productivity) to labor turnover (with the potential loss to the employer of past investments in workers' human capital). At the level of collective contracting, the stakes are arguably much higher for workers, their unions, and employers, with the ultimate cost of a negotiation breakdown being lost incomes to workers and lost profits to employers. Given the potential cost to both of the contracting parties, it is likely that workers and employers have a strong incentive to achieve a solution, in preference to conflict. Like all good threats, the employer's threat of a lockout and the union's of a strike are best if they ensure that an agreement is reached while remaining unused.

In real life, collective bargaining sometimes breaks down, and production, labor earnings, and profits are lost. It is simply not known whether these costs to society are greater or less than those that would arise from breakdowns in individual employer-employee pay negotiations. Indeed, given economies of scale in the production and dissemination of information, it is possible that a system of collective agreements, through its ability to resolve disputes, may be a less costly option from a social point of view than individual contracting.

There is a strong body of empirical evidence in Australia, Canada, Japan, New Zealand, the United Kingdom, and the United States to suggest that the cause of disputes under collective bargaining is asymmetries in the information possessed by the involved parties (Hicks 1932; Walsh 1975; Hazledine and others 1977; Mauro 1982; Hayes 1984; Tracy 1987; Booth and Cressy 1987). A common case is when the trade union misjudges the maximum wage the employer is willing or able to pay. In such circumstances, regulation through its information-gathering and disseminating roles can prove decisive in resolving disputes.

To understand the process, it is important to recognize the distinction between the union proper (sometimes called the official union) and its rank-and-file membership. This distinction results in a tripartite framework where the official union (often a well-informed professional body) acts as an intermediary between the union membership and the employer, reconciling the aspirations of the former against what it judges (on the basis of its more complete knowledge of the overall situation) that the employer would agree to pay. This reconciliation between worker aspirations and labor market realities

may be achieved without either party having to resort to its "no-trade" sanction. Should negotiations break down and a dispute occur, the role of the official union as a purveyor of information continues: The union passes information in both directions about concessions acceptable to each side and any other relevant issues that materialize as the dispute progresses. Information is transmitted until demands fall into balance with offers, at which time a settlement is achieved.

Viewing trade unions in this way—as an information-gathering and spreading body—suggests that policies might be targeted at increasing the efficacy with which unions fill this role. The introduction of cooling-off periods, during which all parties take time out to reassess the situation before implementing no-trade strategies, is one such example. Another example is the requirement that the employer (generally seen as the party in possession of more complete information) divulge to the union and its members certain types of information to minimize the possibility that disputes arise because workers have incorrectly estimated the employer's ability to pay.

Some degree of conflict is inevitable when wages and other employment conditions are set by negotiation (either collective or individual) as opposed to the invisible hand of the market. Recognizing this, there are grounds for believing that a centralized, union-based system of wage bargaining may be less costly to society than an individual-based negotiating system in terms of both total transactions costs and dispute costs. It may therefore be more appropriate to devise policies that seek not to remove unions but rather to increase the efficiency with which they perform these tasks.

*Wage coordination schemes*

Collective bargaining and dispute resolution mechanisms are potentially a powerful means to facilitate wage *coordination*, an influential determinant of labor market and macroeconomic performance. For example, the Japanese system of wage setting is decentralized (firm-based) but coordinated in the sense that it follows company rules based on seniority (hence, they are transparent) rather than individual contracting. Germany and the Netherlands have also coordinated systems through strong employer organizations between large companies and across industries, as well as between unions. Some coordination in France is accomplished through the significant shareholding of government in production in the form of public services, utilities, and large nationalized industries. In Italy there is informal employer coordination through the large firms, regional employers associa-

tions, and union confederations. Finally, Sweden has a centralized employers organization as well as union confederations.

The economies in which these employer-worker coordination mechanisms serve as an alternative to individual contracting have performed satisfactorily. Evidence (OECD 1994) suggests the apparent superior macroeconomic performance of nine countries that have coordinated wage mechanisms—Australia, Belgium, Finland, France, the Netherlands, Norway, Portugal, Spain, and Sweden—which had on average 17 percent cumulative real wage growth and a 7 percent annual inflation rate between 1980 and 1993. Austria, Germany, and Japan ranked at the top of the league, with 31 percent wage growth and inflation of only 3 percent. By contrast, the apparently flexible labor markets in which bargaining is uncoordinated and negotiations take place at the plant or company level (Britain, Canada, New Zealand, Switzerland, and the United States) had 2.5 percent cumulative real wage growth and a 6 percent inflation rate in the same period. Thus, centralized pay setting may not be associated with the typical effects of wage inflexibility that conventional theory postulates. The relationship between the degree of centralization in wage setting and macroeconomic performance may not, however, be linear. Comparative studies across a sample of industrial economies point to a U-shaped relationship: low and high levels of centralization improve macroeconomic outcomes (Calmfors and Driffil 1988).

In most economies, coordination evolved gradually through piecemeal legislation over the course of decades rather than as a massive policy intervention at one point in time. Although some policies may have created insiders and outsiders in the labor market, policies have usually blended the social concerns with the economic realities of the time. Most countries with coordinated systems, especially those in Europe, are in a process of change, partly because of their exposure to external competition and their failure to take account of their countries' international trade performance, and partly because of the declining trend in manufacturing, in which collective bargaining is more common than in white-collar sectors.

#### *Unemployment benefits*

Unemployment deprives workers (and their families) of labor income and decreases their current consumption and welfare. An added concern is that workers seeking to escape unemployment may accept jobs that undercompensate them for their skills and thereby lower their contribution to output.

Four arguments can be given to justify some income support to the unemployed. First, according to the deadweight loss argument, the variability of workers' income and their families' welfare cannot be fully offset by precautionary savings due to reasons of myopia. Benefits are seen as a way to reduce this variability. Second, the poverty argument assumes that the unemployed are among the poorest and that benefits can prevent consumption from dropping below some critical level. However, evidence from industrial countries suggests that unemployment benefits are paid mainly to those at somewhat below the median income, whereas those in the bottom two income deciles get little (Hamermesh 1992). Third, in the efficiency argument, benefits can improve the allocation of labor by enabling the worker to hold out for higher-wage offers (Ehrenberg and Oaxaca 1976). Finally, the political economic argument holds that in the specific case of massive layoffs during adjustment or transition, some form of unemployment benefits can reduce workers' resistance to costs due to restructuring and facilitate transition. However, none of these reasons by itself justifies public involvement in unemployment insurance. An additional argument is required: that private schemes cannot insure against the common risk of widespread recession.

Unemployment benefits are perhaps the most controversial of social policies. Arguments against such benefits include, first, the fact that unemployment is not harmful in itself (compared, for example, with injury or disability). Second, that unemployment is preventable in the sense that particular individuals can avoid it by lowering their reservation wages (that is, the wage below which an individual is unwilling to work). Finally, unemployment benefits may simply crowd out private savings. It can therefore be argued that unemployment benefits make the economy suffer a deadweight loss because taxes must be raised to finance them, while the welfare of workers receiving these benefits is unchanged.

Empirical evidence confirms the positive relationship between the generosity of benefits, in terms of levels and eligibility, and the unemployment rate (Layard, Nickell, and Jackman 1991). This effect comes primarily through an increase in the duration of unemployment rather than the incidence of unemployment. That is, the availability of benefits reduces the pressure for re-employment. However, the adverse effect of unemployment benefits on work incentives, although statistically significant and positive, is usually small and dependent on overall macroeconomic conditions.

In general, the generosity of unemployment benefits has decreased in most countries during the last decade

with no immediate impact on unemployment. In Germany during the 1980s, the unemployment rate was rising at the same time benefits were falling, and the same was true for the United Kingdom. Generous benefits may have a greater effect on search behavior during tight labor market conditions (Burtless 1990). In fact, the rise of unemployment benefits in many countries during the mid- to late 1960s was justified precisely on efficiency grounds: At that time of low unemployment, it was deemed beneficial to the economy to increase benefits for unemployed workers so they could spend more time on job search, and ultimately find more productive, higher-paying jobs than they would have found if they had been pressured by income constraints.

There are several considerations to be taken into account in the design of unemployment insurance. First, although unemployment insurance should help the unemployed and their families to maintain a socially defined level of consumption (the benefit adequacy principle), the level set must not create an undue disincentive to work. This generally requires benefits to be low (for example, below the minimum wage but equal to some agreed poverty line). The correct balance is hard to determine. For example, the United Kingdom is still struggling to find the right balance between wages and benefits and reduce the poverty trap (the rate at which benefits are withdrawn when earned income increases) even after its sizable reforms of the 1980s (see the discussion of the British reforms later in this chapter). The current structure of taxes and benefits in the United Kingdom is such that the net weekly income of a couple with two children would rise by less than £50 if their gross earnings rose from £50 to £300 a week (*The Economist* 1994).

Second, unemployment programs can incorporate longer waiting periods, for example, benefits may not be paid during the first few weeks of unemployment. Third, eligibility can be reduced by including disqualification clauses (for instance, because of employee fault). Similar arguments apply with respect to the duration of benefits which can be limited to, say, six or nine months. Fourth, when the objective of unemployment benefits is to increase the acceptability of reforms by reducing worker resistance to retrenchment in the short run, their costs and benefits should be examined against their fiscal implications and effects on poverty. In this case, alternative policy instruments should be considered (such as severance awards, means-tested social assistance, family and child benefits, and other public social spending on education and health).

Even if the right combination of policy characteristics could be determined, the question of whether an unem-

ployment insurance program should be introduced remains unanswered. In general, informal insurance mechanisms will be relatively more important (compared with government-mandated schemes) in the least-developed countries, where unemployment is hardly a meaningful concept amidst predominantly subsistence activities, the formal sector is small, the tax base is lacking, and "social" security takes the "private" form of extended family networks.

In countries with more mature economies, the following questions must be answered: Should benefits be financed from workers' pay, employers' payroll taxes, or general taxes? Should contributions be set at a flat rate or be earnings-related? Who should be covered? Should unions or the government administer the unemployment insurance scheme? How should the formal and informal sectors be treated? What other mechanisms are in place that affect incentives, welfare, and savings?

#### *Minimum wages*

Minimum wage legislation has been widely adopted at various times by both industrial and developing economies. Although in practice minimum wages are commonly evaded in informal sector activities, especially in developing countries, they have the potential to induce large distortions by holding wages above their market clearing levels. This wage rigidity leads to an excess supply of labor at the prevailing minimum wage, which can manifest itself in the form of (increased) unemployment or lower wages in the unprotected or evading sectors.

A serious methodological issue exists in assessing the effects of minimum wages: There does not exist an indisputable counterfactual against which to judge the consequences of minimum wage legislation. The labor market may not assume a perfectly competitive structure when minimum wage legislation is abandoned. The relevant counterfactual may instead assume an employers' tacit agreement not to raise wages above an agreed level (a fact mentioned, not surprisingly, by Adam Smith). In this case, which can be put more formally in a monopsonistic context, minimum wage legislation can potentially lead to a reduction in distortion relative to the situation that would prevail in its absence. This case has been verified in Morocco, where agricultural workers' restricted geographical mobility across the large estates provides employers with some monopsonistic power (table 9.1).

Empirical evidence on the employment effects of minimum wages is rather diverse. Some studies have indeed produced estimates that appear to be consistent with the prediction that minimum wages lead to increased unem-

TABLE 9.1  
**Effects of minimum wages in selected economies**

<i>Economy (period)</i>	<i>Remarks</i>
Bangladesh (1980s)	Significant negative employment effect on skilled workers and significant positive effect on wages for unskilled labor (Anderson and others 1991). Nonsignificant impact on privately operated industrial sectors, but it could not be determined whether low levels of minimum wages or massive evasion were the cause (Azam 1994c).
Chile (1970s–1980s)	Real industrial wages fell only 20 percent, whereas minimum wages in industry fell 50 percent; minimum wages are still found to affect (“Granger-cause”) real wages (Paldam and Riveros 1986; Azam 1994c).
Colombia (1980s)	Substantial negative effects on employment (Bell 1994).
Costa Rica (1980s)	High noncompliance especially among workers in small firms and unskilled workers (Gindling and Terrell 1993).
Kenya (1960s–1970s)	An increase in the minimum wage following independence induced an expansion of the private demand for education (Collier and Lal 1986; Azam 1994a).
Mexico (1980s)	Significant noncompliance. The real value of minimum wages was eroded by inflation, and even the lowest market wages rose above the minimum. No noticeable impact on employment (Bell 1994).
Morocco (1980s)	Minimum wages have probably reduced employers’ rent in large agricultural holdings and augmented the incomes of the poorest families. No adverse effect on employment (Pascon and Ennaji 1986; Azam 1994b).
Puerto Rico (1960s–1980s)	Enforcement of minimum wages has been rigorous and affected both low wages and the distribution of wages. Negative effects on employment and positive effects on unemployment (Reynolds 1965; Santiago 1989).
<i>General</i>	
Second-order effects	Minimum wages do not generally prevent downward adjustment of real wages because other dynamic adjustments take place in response to macroeconomic shocks (Freeman 1992).
Efficiency-wage considerations	In Africa minimum wages in urban areas may have attracted better migrants from rural areas and increased their productivity (Mazumdar 1989, 1994).
Political considerations	Studies based on simple static examination of minimum wages are uninformative. Minimum wages play different roles in different countries. The political economy aspects of minimum wages on costs and benefits must be identified (Tabellini and Rama 1994; Azam 1994c).

ployment for the United States, although the effect can be small (Ehrenberg and Smith 1991). A more recent study of the United States found an elasticity of employment to minimum wages for teens and young adults of only  $-0.1$  to  $-0.2$  (Neumark and Wascher 1992). Some studies have found that minimum wages have no adverse employment effect, and some have even found positive effects (Wellington 1991; Card, Katz, and Krueger 1993).

With respect to other countries, recent reviews of the employment effects of minimum wage legislation in the United Kingdom and other OECD economies have concluded that this legislation does not result in increased unemployment (Callaghan and Jones 1993). In a study of the minimum wage in France, Fitoussi (1994) concluded that there was no significant evidence to support the cor-

relation between minimum wages and overall unemployment. Evidence on a direct link between minimum wages and unemployment in other European countries is lacking as well (Gregory and Sandoval 1994).

Another effect of minimum wages that may dampen the adverse effect on employment can come from reductions in the duration of job search of low-paid workers. Such workers will accept the first job they are offered to the extent that the minimum wage exceeds their reservation wage. Also, efficiency wage theory suggests the existence of a positive association between the wage paid to workers and the level of labor productivity, with the consequence that minimum wage legislation can, in principle, result in output and welfare gains as opposed to losses (see table 9.1). Recent empirical evidence reported by

Machin and Manning (1992) suggests a positive correlation between minimum wages and employment in Britain, an effect (apparently attributed to monopsonistic behavior) that was particularly strong in the catering sector. Similarly, a recent U.S. study found that fast-food restaurants in New Jersey hired more workers after that state increased its minimum wage by 19 percent in 1992 (Card and Krueger 1993).

The effects of minimum wages on the distribution of earnings are ambiguous. One might think that the imposition of minimum wages would narrow the distribution of earnings by increasing the pay of the mainly unskilled workers who previously were employed at lower wages. However, the pay of semiskilled workers might also increase either through substitution of the unskilled by the semiskilled (the slope of the supply curve of the semiskilled is positive) or because of the ripple effect on the semiskilled, whose pay will increase to preserve their relative status (Grossman 1983). The increase in the pay of the unskilled and semiskilled is unlikely to extend to high-paid supervisory and managerial staff for two reasons. First, they are not substitutes in production to the other two groups. Second, because of higher wages at the lower end of the distribution, the firm will be able to screen and hire more productive unskilled and semiskilled workers requiring less supervision, thereby reducing the number and earnings of supervisors and managers (Calvo and Wellisz 1979).

Other forces are also at work. First, if demand contraction takes the form of retrenchment of workers in the covered sector, labor supply to the uncovered sectors will increase, thereby decreasing the wages of workers in those sectors. Second, the dispersion of earnings could increase if firms reduced the hours worked by minimum wage workers proportionately more than the difference between the competitive and minimum wages. Thus, the net effect of minimum wages on the distribution of earnings, a prime concern of the policy, is not clear beforehand.

Of course, specific cases exist in which minimum wages have rather clear effects. High minimum wages for male workers in the export-processing zones of Mauritius used to discourage male employment, and low minimum wages for women induced an excess demand for female workers. After the minimum wage for male workers was eliminated in December 1984, 95 percent of new recruits for the following year were men who were paid less than the former minimum wage (Robinson 1994).

In conclusion, the effects of minimum wages on employment and the distribution of earnings and poverty are usually found to be country-specific and variable.

Fiscal instruments or a combination of labor market and fiscal policies might better serve the objective of poverty reduction because the low paid are not necessarily the poor. Many low-paid workers are so-called secondary workers (that is, second-income earners within the family, such as adolescents or spouses). Thus, although such workers are paid little, their household incomes are not necessarily below the poverty line. Also, in many countries the poor do not work at all. For example, in Britain only 1 percent of women whose husbands are out of work and who have dependent children are employed themselves, and also, fewer than half of all single mothers work. Finally, many of the working poor are poor because of personal circumstances (for example, a large family or incapacitated members). In such cases it would be preferable to address poverty through the social security system and general taxes rather than through minimum wages.

#### *Job security regulation*

Job security can arise from free contracting between workers and employers. Lifetime employment in Japan, for example, is basically a postwar phenomenon that became relatively widespread, especially after the onset of rapid economic growth when rapid technological change increased training requirements (Taira 1970; Johnson 1982; Seike and Tan 1992). Legislation can extend the benefits of longer-lasting employment relationships to workers and companies unable to strike optimal contracts.

Job security regulations include restrictions on individual dismissals and on collective layoffs, and mandatory severance payments in the case of separation. Job protection is usually not extended to those who perform their duties in a careless way leading to employer losses, pass trade secrets to competitors, are chronically absent from work, suffer from substance abuse, or are convicted of a felony. Regulations also can control collective dismissals when massive layoffs are expected to create negative externalities in workers' communities. Such regulations can take the form of a minimum notice requirement (three or six months' notice is common), prior approval by the government, a ceiling on dismissals (for example, no more than 50 layoffs per month), or a restriction on the percentage of the firm's work force that can be affected (for example, no more than 2 percent of workers per month). Severance payments are usually based on the worker's tenure with the employer.

Job security regulations can affect both the level of employment and the speed and extent of industrial restructuring during an economic downturn, especially a

prolonged recession. Mandated job security legislation may lower the level of employment prior to adjustment by making employers cautious about hiring permanent workers. Sluggish job creation and low employment levels in the formal sector in countries as diverse as India, the United Kingdom, and Zimbabwe have been partly attributed to job security provisions (Nickell 1982; Fallon and Lucas 1991), although there is some evidence to suggest that job security laws had no impact on employment levels in Malaysia (Standing 1989). Fallon and Riveros (1989) estimated that, on average, job security regulations reduced formal employment in 35 Indian industries by 18 percent and in 29 industries in Zimbabwe by 25 percent. However, they concluded that job security did not slow employment adjustment in India or in Zimbabwe.

The case for eliminating job security regulation is far from conclusive, for many reasons. First, when dismissal becomes relatively difficult the firm has an additional incentive to increase or at least prevent a decline in worker productivity. This incentive can result in more training and an increase in productivity both by creating additional skills and by making workers more willing to accept reassignments within the firm because of their employment security (Paredes 1993). In this case the adverse effect of restricting layoffs on flexibility should be considered against the benefits of greater productivity and lower social costs arising from a more gradual restructuring of the labor force through attrition. This argument assumes that worker effort does not drop as a result of job security. If it does, then firms will not provide more training but will increase the capital-labor ratio and reduce employment.

Second, although job security imposes a cost on employers, this cost may eventually be passed to workers. For example, workers may be willing to accept lower wages or greater within-firm flexibility in return for greater job security. Some empirical studies have shown that the costs of mandates are largely shifted to wages with little effect on employment (Gruber and Kruger 1991). For this to happen, wages should be flexible. Other studies have arrived at less clear-cut conclusions, with no evidence that wages per hour are systematically affected by the introduction of job security regulation (Fallon and Lucas 1991). The ambiguity rests on the fact that there are many forces at work whose net effect is unclear a priori. They include compensating differentials that, in a competitive labor market, should reduce wages in protected sectors; increase wage bargaining, which can push up wages; or reduce worker motivation, which can raise wages per unit effort. All this suggests that the behavior

and treatment of other aspects of employment and pay are crucial in determining what the eventual effect of job security legislation can be.

Third, advance notice requirements, one of the main elements of job security regulation, may reduce the duration of unemployment to the extent that they afford workers more time for job search. Consequently, such notices may have both an employment and an efficiency effect. Empirical evidence suggests that job security legislation negatively affects the incidence of unemployment though it does not reduce a spell of unemployment (Podgursky and Swain 1987; Ehrenberg and Jacobson 1988). In terms of efficiency, pay in a new job (replacement earnings) among workers who receive advance notice of termination is higher than that of workers whose employment is terminated suddenly (Ruhm 1992).

This discussion of job security legislation and the preceding section on minimum wages both speak of flexibility in the labor market. Flexibility typically refers to the ease with which employment and wages adjust, usually downward. Numerical flexibility refers to the adjustment in employment and hours of work and financial flexibility to the pay adjustment in response to changes in product markets and profitability. Other types of flexibility can enable the economy to adjust, sometimes more efficiently than through prices (wages) and quantities (employment). Functional flexibility, the reassignment of workers to other jobs within the firm, is common among permanent workers in firms in Germany and Japan. Neither country has much reason to envy the (narrowly defined) more flexible economies of the United States or the United Kingdom. Finally, certain attitudes toward entrepreneurship and management are important determinants of labor market outcomes, giving rise to managerial flexibility and distancing strategies (such as subcontracting) on the employer side. The design of job security regulation should, therefore, take into account aspects of flexibility beyond the conventional "wage cum employment reduction" view.

#### *Labor regulation and international trade*

The discussion so far has focused on labor reforms from a national perspective, bypassing an important consideration: the integration of the world economy through international trade, advances in technology, and the globalization of production. When labor rights and labor standards differ among countries, such differences can give a cost advantage in internationally traded goods to some countries. Along with international trade, technology has enabled labor services to be subcontracted directly to

workers in developing countries. For example, Caribbean workers perform data entry procedures and transmit their work electronically to U.S.-based companies. Skilled Indian engineers receive initial drawings by satellite and send the final products to the United States in the same way. Regulation therefore can no longer be concerned with only the working of the national labor market but must take into account conditions in overseas markets as well.

Two broad views on this issue exist. The first holds that labor regulation reduces economic efficiency and growth, especially in countries with a high incidence of poverty (Herzenberg 1990). The second view holds that differences in labor regulation among countries tend to discriminate against those countries with higher standards and greater respect for workers' rights, typically industrial countries. It is within this latter perspective that the United States considers the violation of basic worker rights and minimum labor standards as unfair trade practices. It has adopted legislation to this effect (such as the Omnibus Trade and Competitiveness Act of 1988) that restricts trade and investment guarantees in countries that either do not enforce or violate labor rights and standards. One policy issue in a world trade context is whether countries can agree on some level of labor rights and standards that would punish neither the more "principled" (industrial) economies nor those with surplus labor (the developing ones).

Labor regulation is also important from another international perspective: migration. Although there is a general consensus that flows of capital and goods should be free, there is no such agreement about the movement of people (Russell and Teitelbaum 1992). In general, labor regulation should make labor exchange possible at least through temporary relocation of labor (Bhagwati 1987). The services that have a strong provider-relocation requirement (such as construction and engineering) call for more permanent provisions (OECD 1989).

The removal of restrictions on migration is necessary for increasing the mobility of workers between countries, but it may not be sufficient. Complementary actions are often required, as illustrated by the European Union's principle on the free movement of citizens of member states. The adoption of this principle required additional "regulation" in the areas of educational qualifications, certification, and licensing. Had national medical councils and engineering associations continued to place restrictions on the employment of doctors and engineers qualified in another member state, they would have violated the principle of free mobility of workers within the

European Union. Thus, regulation (in the form of mutual recognition of credentials) was necessary, as well as the provision entitling workers who can legally practice a profession in one member state to practice that profession in all other member states.

### **A case study: the dock industry in Britain**

This section considers in some detail a case study that illustrates the practical issues involved in labor market regulation and subsequent deregulation. The experience of the international dock industry is interesting for several reasons. First, the characteristics of the industry that gave rise to its regulatory structures were very similar across countries. Second, an array of alternative regulatory solutions to the industry's problems have been put in place. Finally, largely led by the United Kingdom's 1989 deregulation of the dock labor market, many governments are currently either fundamentally reforming or completely dismantling their regulatory structures.

The dock industry has witnessed marked technological changes over the last two decades as a result of the development of lift-on/lift-off containerization and roll-on/roll-off trailers. So rapid was the introduction of the new technology that the labor-intensive processes that had been used for decades were quickly rendered inappropriate. It is no exaggeration to say that the changes in established working practices necessitated by these technological advances were as significant as those required of the manufacturing industry as a whole at the onset of the Industrial Revolution (Turnbull, Woolfson, and Kelly 1992).

#### *Industry labor problems*

In the United Kingdom during the early 1990s, port costs were some 60 percent higher than those in continental European ports. In September 1991 a container ship could be unloaded and reloaded in the port of Tilbury (London) at an average rate of 25 TEU (twenty-foot equivalent units) container boxes per gross crane hour. However, the same ship could be handled at the northern European port of Zeebrugge at a rate of 40 TEUs per hour. Significant productivity differentials are also evident at ports in the newly industrialized countries. For instance, whereas productivity rates of 30 TEUs per gross crane hour are currently achieved at U.S. ports, the rates exceed 40 TEUs per hour at Hong Kong and 50 TEUs per hour at Singapore (Turnbull and Sapsford 1993).

Clearly, these productivity differentials cannot be attributed to technology, because both the vessels and the container boxes are identical. The structure of capital

stock in the ports is also not a factor because even 20-year-old cranes are technically capable of high performance levels. Although a range of factors have been suggested as contributing to such productivity differentials (including the structure of port ownership and administration), the functioning of dock labor markets is now widely recognized as a critical factor.

The docks have been traditionally characterized by one of the most archaic and inefficient of all occupational or industrial labor markets. Irregular demand for labor and associated casual forms of employment made chronic surplus labor commonplace, which in turn nurtured inefficient working practices and frequent strikes. Excess manning, for example, was a common means of sharing available work. Strikes were typically protests against low average earnings, irregular earnings, unequal job opportunities, or irregular work. Conflict and inefficiency in the international dock industry seemed inseparable from the casual nature of employment. But the problem ran much deeper in the sense that pressures of excess labor supply in the docks can generate a vicious spiral of unemployment, low earnings, conflict and, in many ports, bribery and corruption.

Faced with industrial conflict and inefficiency (high labor mobility and excessive variability in wage levels), countries began introducing regulatory controls in the dock industry in the late 1940s and early 1950s. In some countries, employers took actions to foster stability (as in the Netherlands). In others, the union took the lead (as in the West Coast U.S. ports). Usually, however, it was government action that established, and subsequently underwrote, a system of joint regulation by employers and unions (as in Australia, Britain, France, and New Zealand).

#### *Regulation and deregulation on the waterfront*

The National Dock Labor Scheme was established in 1947 to regulate the dock industry in Britain. Its primary objective was to ensure greater regularity of employment for dock workers and an adequate labor supply for port employers. A national board and 22 local dock labor boards were established to administer the scheme, each with an equal number of employer and worker representatives. Both employers and workers were required to register with the scheme. Registered employers were prohibited from hiring nonregistered dock workers. Registered dock workers who were out of work were placed in a pool of reserve labor and paid "attendance money." If the amount of such money, plus any earnings, fell below a guaranteed wage (agreed to by the National

Joint Council for the Port Transport industry), the worker would receive supplementary pay to reach the minimum. In essence, because workers were hired on a shift-by-shift, casual basis, the port operators were merely operational employers, whereas the national board acted as the holding employer.

The persistence of the casual work relationship between employers and workers and its resulting record of poor industrial relations led an investigating committee to recommend that the casual system of employment be abandoned in favor of a system of tenured permanent employment for registered dock workers (Aldington-Jones 1972). These recommendations were accepted by the government in 1972 in the form of legislation providing a "job for life" for each registered dock worker. Whether this legislation provided a true job for life for dock workers is, however, questionable because an increase in the rate of job loss in the industry followed its passage.

The most important reason this system of regulation failed to bring employment security, industrial peace, and rising labor productivity was the sheer pace of technological change faced by the industry. The new cargo handling techniques—such as containers, roll-on/roll-off vessels, and bulk carriers—reduced manpower requirements dramatically and at a speed far greater than the rate at which the industry was able to shed labor. Mechanization was a major contributing factor to the marked increase in the industry's surplus labor rate during the 1970s. Despite an elaborate network of regulation, insecurity of employment remained the hallmark of dock work.

The regulatory framework that had governed the operation of the dock labor market in the United Kingdom for 40 years was dismantled, almost overnight, in April 1989. Deregulation was largely a reaction to the low labor productivity and poor labor relations record of the U.K. port industry. It also reflected the government's general view that regulatory and institutional structures impede the free operation of market forces.

Deregulation of the dock industry produced significant (and most likely sustainable) gains, along with some short-term adjustment costs. Productivity in the U.K. port industry rose sharply in the three years immediately following deregulation. In addition, strike activity dropped significantly. However, strikes are not the only manifestation of industrial conflict, and their reduction is only a partial indicator of improvement (Douglas 1923; Yoder 1940). Absenteeism (and even accidents) at the workplace constitutes an alternative manifestation of industrial conflict (Knowles 1952; Handy 1968). While

strike activity was reduced in the dock industry as a consequence of legislative changes there was an increase in absenteeism and accidents (Turnbull and Sapsford 1993; Evans and others 1993).

The deregulation of the dock industry resembled that of the coal industry in the United Kingdom. Before 1984 the coal industry was also heavily unionized and used old technology, and its workers were paid wages above those that a free market would have enabled them to earn. A nonunionized coal field could have produced nearly one-quarter more output than a unionized one (Pencavel 1977). The deregulation of the coal industry in the early 1980s brought about long-term benefits at the expense of severe industrial conflict in the short run: The miners' nearly year-long strike was one of the most significant events of that decade.

This section showed that in the case of a specific industry, the effects of deregulation can be anticipated. Although the evidence suggests that there may be some short-run costs associated with deregulation (such as increases in absenteeism and accidents), there were clear longer-run benefits. However, reforms of interventions that cut across the labor market (such as economywide minimum wages or trade union regulation, discussed in the next section) are more difficult to assess and their impact harder to predict.

### **Empirical evidence of economywide reforms**

In the case of specific industries, results of reforms are predictable within a partial equilibrium framework. The need for trading off short- with long-run effects does not usually arise, because affected workers are typically a small fraction of the labor force. It is, however, particularly difficult to identify and measure economywide effects when systemic policy reforms are undertaken. The path to a "first-best" equilibrium may generate nontrivial unemployment at the macro level, and the eventual pattern of wages is rather unpredictable in the sense that it depends on the slopes and elasticities of products and inputs, as well as product mix and factor intensities (Edwards and Cox-Edwards 1990). Even if employment and wage effects can be estimated, the costs of reforms may be heavily concentrated in the current time period, whereas the benefits may arise in the future. A clear illustration of this "costs now, benefits later" phenomenon is the restructuring now under way in formerly planned economies. Alternatively said, if one accepts that reforms require (unspecified) long and variable lags for full effects, it is nearly impossible to reject the beneficial value of labor deregulation.

Bearing these difficulties in mind, this section first examines the U.K. labor market, which experienced the most radical change in regulation among the OECD economies in recent years. The behavior of the labor market in a wide range of developing countries that underwent adjustment and trade reform is then discussed.

#### *Labor market reform in the United Kingdom*

The rigidities in the U.K. labor market have been widely regarded as the cause of the country's poor macroeconomic performance since the 1960s. By the end of the 1970s, the United Kingdom's powerful and inflexible trade unions and high unemployment benefits were two areas of significant concern. Labor deregulation in the 1980s significantly curtailed the role of unions and fundamentally altered industrial relations. Closed shops became practically extinct, management prerogatives were restored, job security regulation was eased, collective agreements became more flexible, unemployment benefits were reduced and eligibility rules became tighter, and trade union membership fell by 3 million during the decade. No aspect of the labor market escaped regulatory reform: Labor demand, labor supply, and the institutional framework were all significantly affected.

By most measures, however, the United Kingdom's economy does not seem to have benefited from the massive labor reforms of the last 15 years. Although changes in the industrial relations framework and reductions in benefits and the power of unions have altered the distribution of earnings by increasing inequality, they have not led to low inflation with low unemployment (Barrell 1994).

Why have the U.K. labor reforms had no apparent impact on macroeconomic performance? First, the expected macroeconomic effects may have been neutralized by the move away from employer-union bargaining toward decentralized bargaining, which reduced labor market coordination. However, it is also possible that labor reforms, especially in pay-setting procedures, had weaker effects on economic performance and employment outcomes relative to those of other changes in the economy that emanated from macroeconomic and trade factors (Metcalf 1994).

Second, although labor reforms may have exerted some independent effect on labor productivity, competitiveness is influenced by a range of factors in addition to productivity in the labor market. Specifically, trends in U.K. competitiveness in world markets during the 1980s were heavily influenced by movements in both relative inflation rates and exchange rates (Brown and Wadhvani 1991). Although micro problems may have been sub-

stantially diminished by the supply-side reforms of the 1980s, macro policy was severely constrained by the United Kingdom's participation in the European Union's Exchange Rate Mechanism, which resulted in tight monetary policy (Minford and Riley 1994). In addition, the reduction in worker security, by fostering greater uncertainty for consumers, may have created less stable demand (Anderton and Mayhew 1994).

Finally, the supply response may have been constrained by the limited skills of British workers and the lack of a responsive training system. This deficiency restricted the adjustment during the rapid change away from manufacturing toward services, and the reduction of employment security did not significantly change worker behavior in the expected direction (Anderton and Mayhew 1994).

These explanations suggest that labor market rigidities may be less important than those in other (domestic or international) sectors, and that labor regulation or deregulation alone is not sufficient to change the course of the macroeconomy. The "price theorist's ideal changes" that took place in the United Kingdom during the 1980s may have reduced rigidities and union power and increased mobility, incentives, and the responsiveness of wages and employment at the subsector microlevel. But these changes increased wage inequalities without improving at the macro-level either the response of aggregate real wages to unemployment or the transition out of unemployment. These changes do not seem "to reflect the working of an ideal labor market" (Blanchflower and Freeman 1994). The United Kingdom may have ended up with the worse aspects of two possible worlds: the wage inequality of the decentralized U.S. labor market together with the high and lengthy spells of unemployment that are a characteristic of European labor markets.

#### *Labor regulation, adjustment, and trade in developing economies*

Labor market regulation in developing countries has been assessed from both micro and macroeconomic perspectives. Reviewing the experience of a sample of 23 developing countries, Fallon and Riveros (1989) concluded that labor regulations such as minimum wages and non-wage costs (including social security, and medical and fringe benefits) appeared to have distortionary effects in the countries and sectors in which they were effectively enforced. They provided evidence that job security regulations can have a negative effect on employment levels at a given level of output but less so on the pace of employment adjustment.

The Fallon and Riveros study faced the same methodological and empirical issues as those discussed earlier in the case of the United Kingdom. The link between labor regulation and macroeconomic performance is neither linear nor always clear. For example, while it is always tempting to ask whether adjustment could have been faster in the absence of minimum wage legislation, minimum wages may be in general irrelevant for macroeconomic performance or simply ignored *de facto*. When economic crises occur, either governments tend to let real minimum wages fall or workers simply become willing to accept lower wages when their jobs are at stake. Real minimum wages in Mexico were nearly halved in the 1980s, and 16 percent of male workers and 66 percent of female workers were paid less than the minimum wage by 1988. In Algeria and Jamaica nearly half of the microenterprises did not comply with minimum wage legislation, and evasion in Niger and Swaziland was nearly universal. In Kenya real average wages fell by 23 percent in the first half of the 1980s, but minimum wages had already fallen by 41 percent (World Bank 1995).

Horton, Kanbur, and Mazumdar (1994) examined the interactions between labor markets and adjustment programs in a cross-country context. The countries examined in the study have different (in some cases, widely different) labor market regulatory systems. The study included Argentina, Bolivia, Brazil, Chile, Costa Rica, Côte d'Ivoire, Ghana, Kenya, Malaysia, the Republic of Korea, and Thailand. Still, the authors found that shifts in the intersectoral allocation of labor resources occurred largely in the desired direction—into the tradable goods sector. The role of trade unions in the adjustment emerged as varied and complex (see also Standing 1992). In neither Africa nor Asia did unions constitute a major obstacle to successful adjustment in the aggregate, although the response of unions varied from outright opposition in some cases to active cooperation in others. Unions in Argentina and Brazil attracted some apparent blame for the lack of adjustment, but their presence and activities in Costa Rica did not appear to restrict adjustment in that country. Although much labor market legislation was removed in Bolivia, a significant recovery had yet to emerge when the review was conducted.

On the face of these results it may be tempting to conclude that the presence of unions does not necessarily lead to adverse macro outcomes. However, another explanation may be that unions in most developing countries are relatively unimportant—that is, they are not a serious obstacle to adjustment simply because they operate on only a small part of the labor market and leave the large

informal sector untouched. The conclusion of Horton, Kanbur, and Mazumdar (1994) was that neither aggregate real wage rigidity nor labor market inflexibility appeared to have hindered the process of structural adjustment.

The evidence on the linkages between labor market characteristics and macroeconomic performance is also mixed in the countries that have a successful economic record (Freeman 1993). The East Asian newly industrialized countries have had remarkably high and consistent growth rates in the last two to three decades. Yet union repression in the Republic of Korea and wage repression in Singapore cannot be considered as significant factors in the growth of their economies as compared to the economies of Taiwan (China) and Hong Kong (World Bank 1993). In addition, these four East Asian economies have performed equally well despite significant differences in the industrial composition of output and employment and considerable transformations of their economies over time (Fields 1994). Thus, neither growth nor adjustment seems to relate singularly to specific labor market characteristics. It seems that in the same way that growth is the driving (positive) force of employment and earnings, macroeconomic imbalances are the driving (negative) force of the economy during adjustment. In fact, the relationship between specific policy variables and economic growth is generally weak with the exception of the investment ratio to GDP (Levine and Renelt 1990). Specifically with respect to labor policies, another study of 31 countries estimated that labor market distortions do not account for more than 10 percent in the variation of economic growth (Agarwala 1983).

A further source of insight into the process of trade and labor reform is provided by Papageorgiou, Michaely, and Choksi (1991), who evaluated 36 liberalization episodes in 19 countries over the period 1951–82. No strong evidence emerged to suggest that labor market performance, especially as it relates to the structure of regulation, exerted a marked impact on the likelihood of success in liberalization. One possible exception is the case of Spain. The authors argued that the marked rise in aggregate unemployment that accompanied Spain's liberalization episodes was due to labor market rigidities. Although similar increases in aggregate unemployment accompanied liberalization episodes in Chile, it is interesting to note that the authors attributed the rise in Chile not to labor market rigidities but to problems of exchange rate overshooting. Whether these diagnoses are correct remains an open question: it is by no means an easy task to disentangle the effects of trade liberalization from those of other policy changes (Greenaway 1993).

What this discussion on the evidence suggests is that isolating specific labor regulatory effects and establishing the causality between the labor market and economic performance are complicated tasks. In some cases diagnoses are easy because regulation has reached overtly inefficient or unsustainable levels. For example, employment protection legislation and the resulting severance awards that a worker qualifies for only months after entering into an apparently temporary contract are clearly inefficient and a cost deterrent to employment creation (Cox-Edwards 1993). It is not so much that such provisions defy the arguments for some form of regulation outlined earlier in this chapter. Rather, they simply defy common sense.

Another example of inappropriate regulatory policy is minimum wages that are set below the level of social assistance to which a worker may be entitled. At such a level, it would be as attractive to the worker to remain on benefits as it would be difficult for employers to hire workers. If, by contrast, minimum wages are set much above the competitive wage, they either will be evaded or, when enforced, will be paid to the relatively more skilled workers whose productivity is high enough to justify such costs to their employers. In both of these cases, the objectives of policies, whether social assistance programs or minimum wages, are simply defied by inappropriate policy design and provisions.

Employment and pay practices in the government sector are themselves a source of labor market distortions in many countries. For example, mandatory hiring of university graduates into the civil service is widely practiced in developing countries such as Egypt, Guinea, Mali, Senegal, Sudan, and Togo (Squire and Suthiwart-Narueput 1994). Such practices cannot be justified on either efficiency or equity grounds, and are usually adopted because of problems outside the labor market (such as the overt public subsidization of higher education) or political economy considerations. They also can hinder attempts to provide a viable and efficient civil service. Overstaffing in the Egyptian government sector had already reached 40 percent by the mid-1970s (Gelb, Knight, and Sabot 1991). In other countries the net change in government employment during adjustment has been unclear because the downsizing of the civil service in some areas was offset by the automatic rehiring of graduates of training colleges in sectors exempted from retrenchment (such as security or education workers).

Finally, the deficit-financed growth of the public sector as an employer of last resort is another case in which the arguments made in this chapter do not apply. For

example, the civil service in Tanzania has grown 4.5 percent a year over the last 30 years, while average pay has fallen in real terms by 75 percent since 1972 and the debt situation has deteriorated. Similar changes have been observed in the public sector in other countries (Lindauer, Meesook, and Suebsaeng 1988). The magnitude of these figures suggests that such policies go far beyond what even the most comprehensive demand management recipe would ever prescribe. In such cases the issue is not whether deregulation would be beneficial. The answer is clearly affirmative. Rather, the crucial question is how civil society can align the (apparently) political considerations that created such regulations with the economic reality and concern for the general welfare of the population.

### Summary

The Marshallian laws and peculiarities of labor provide sufficient ground for countries to regulate various aspects of the labor market. Labor market regulation also can be used to address adverse effects caused by imperfections in other markets. However, there are no theoretical reasons to suggest that labor policies can be introduced uncritically. What is appropriate in some circumstances may not necessarily be right in others.

Decisions regarding what to regulate, when to regulate, and by how much to regulate will vary in each particular situation. Where excessive or inappropriate regulation is easily identified, the case for deregulation is clear cut, as the discussion of the British dock industry and of high minimum wages and employment in the government sector illustrated. This does not, however, prove the case for complete deregulation of the labor market, as the discussion on the redefinition of the regulatory framework in the United Kingdom and on developing countries' experience with adjustment revealed.

Labor policies should not be adopted simply because of an abstract belief that markets work better in the absence of regulation. Rather, such policies should be based on a critical evaluation of the evidence and an assessment of the likely effects of deregulation against the intended objectives, whether economic or noneconomic. Indeed, decisions about whether to regulate, deregulate, or let the labor market be guided by its own invisible hand involve economic as well as political judgments. Often policymakers regard it in their personal interest to consider views of interest groups. It is both unrealistic and somewhat naive to expect policymakers to introduce policies today solely because of the benefits they promise for some time in the long run. And even if policies would cre-

ate gains in the short run, the compensation principle of welfare economics (which holds that the winners can compensate the losers) may not be relevant if the benefits are dispersed, and the potential losers can forestall reform and preserve their status by bribing the potential winners. When political considerations become relevant, the issue is not simply more or less regulation but also better or worse governments. In practice, countries have achieved notable and sustained economic growth and social development with regulated labor markets. However, increasing international competition and the globalization of production call for a redefinition of the regulatory framework to enable labor market adjustment and increase labor market flexibility as defined in this chapter.

To improve the understanding of how labor markets and regulation work, future research should integrate the study of various aspects of labor flexibility and institutional aspects in which labor is exchanged within the context of the macro and international economy. Such analysis should take into account labor heterogeneity and norms and resulting social institutions in order to better assess the differential impact of trade and technology on product and labor demand. From this knowledge will come an understanding of the demand for different skills and workers' earnings, the response of the education and training system to changes in the demand for labor and workers' earnings, and the effects of constraints faced by the suppliers of labor with respect to housing and geographical mobility.

The key to the success of future research may be not so much what will be studied but how. This can be highlighted with the quotation of the researcher (Rees 1993) who introduced the subject of labor economics in its modern form to the profession:

the neoclassical theory of wage determination, which I taught for 30 years and have tried to explain in my textbook, has nothing to say about wage fairness. The factors involved in setting wages and salaries in the real world seem to be very different than those specified in the neoclassical theory. . . . [Fairness is] a powerful force in determining wage structure, but does not exclude the ultimate effect of neoclassical wage determinants (p. 243).

Research must move away from the notion that an explicit microfoundation is required for every behavioral assumption. The analytical niceties do not always provide a reliable basis for prediction in complex, continuously evolving economic systems.

**Note**

The author thanks Claudio Frischtak, Jane Armitage, Peter Fallon, Anat Levy, Ana Revenga, Haneen Sayed, and Michael Walton for their comments.

**References**

- Agarwala, Ramgopal. 1983. *Price Distortions and Growth in Developing Countries*. Working Paper 575. Washington, D.C.: World Bank.
- Aldington-Jones. 1972. *Interim Report of the Joint Special Committee on the Ports Industry*. London: HMSO.
- Anderson, K., and others. 1991. "The Effects of Labor Laws and Labor Practices on Employment and Industrialization in Bangladesh." *Bangladesh Development Studies* 19:131-56.
- Anderton, B., and K. Mayhew. 1994. "A Comparative Analysis of the UK Labor Market." In R. Barrell, ed., *The UK Labor Market: Comparative Aspects and Institutional Developments*. Cambridge: Cambridge University Press.
- Azam, Jean-Paul. 1994a. *Contourner l'état: la croissance économique au Kenya, 1964-1990*. Paris: OECD Development Center.
- . 1994b. "Effects of Minimum Wages in Developing Countries: An Exploration." Background paper to *World Development Report 1995*. Washington, D.C.: World Bank.
- . 1994c. "Industrial Wage Dynamics in Bangladesh." Consultant's report. World Bank, Washington, D.C.
- Barrell, R., ed. 1994. *The UK Labor Market: Comparative Aspects and Institutional Developments*. Cambridge: Cambridge University Press.
- Bell, L. A. 1994. "The Impact of Minimum Wages in Mexico and Colombia." Paper presented at the Labor Market Workshop. World Bank, Washington, D.C., July 6-8.
- Bhagwati, J. 1987. "Services." In M. Finger and A. Olechowski, eds., *The Uruguay Round: A Handbook for the Multilateral Trade Negotiations*. Washington, D.C.: World Bank.
- . 1988. *Protectionism*. Cambridge, Mass.: MIT Press.
- Blanchflower, O., and R. Freeman. 1994. "Did the Thatcher Reforms Change British Labor Market Performance?" In R. Barrell, ed., *The UK Labor Market: Comparative Aspects and Institutional Developments*. Cambridge: Cambridge University Press.
- Booth, A., and R. Cressy. 1987. "Strikes with Asymmetric Information: Theory and Evidence." Discussion Paper 178. Australia National University, Centre for Economic Policy Research.
- Brown, W., and S. Wadhvani. 1991. "The Economic Effects of Industrial Relations Legislation Since 1979." *National Institute Economic Review* (April): 57-70.
- Burtless, G. 1990. "Unemployment Insurance and Labor Supply: A Survey." In W. L. Hansen and J. Bayers, eds., *Unemployment Insurance: The Second Half-Century*. Madison: University of Wisconsin Press.
- Callaghan, B., and R. Jones. 1993. "Wages Councils and Abolition: The TUC [Trade Unions Council] Perspective." *International Journal of Manpower* 14: 17-37.
- Calmfors, L., and J. Driffil. 1988. "Centralisation of Wage Bargaining and Macroeconomic Performance." *Economic Policy* (April): 13-61.
- Calvo, G. A., and S. Wellisz. 1979. "Hierarchies, Ability, and Income Distribution." *Journal of Political Economy* 87: 991-1010.
- Card, D., L. F. Katz, and A. B. Krueger. 1993. *An Evaluation of Recent Evidence on the Employment Effects of Minimum and Subminimum Wages*. Working Paper 4528. Cambridge, Mass.: National Bureau of Economic Research.
- Card, D., and Alan B. Krueger. 1993. *Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania*. Working Paper 4509. Cambridge, Mass.: National Bureau of Economic Research.
- Collier, P., and D. Lal. 1986. *Labor and Poverty in Kenya, 1990-1980*. Oxford: Clarendon Press.
- Cox-Edwards, A. 1993. "Labor Market Regulation in Latin America and the Caribbean." Report 31. World Bank, Latin America and Caribbean Technical Department, Regional Studies Program, Washington D.C.
- DeFina, R. H. 1983. "Unions, Relative Wages, and Economic Efficiency" *Journal of Labor Economics* 1: 408-29.
- Dertouzos, J. N., and L. A. Karoly. 1992. *Labor Market Responses to Employer Liability*. Santa Monica, Calif.: RAND Institute for Civil Justice.
- Douglas, P. H. 1923. "An Analysis of Strike Statistics, 1881-1921" *Journal of the American Statistical Association* 18: 866-77.
- The Economist*. 1994. "Caught in the Trap." December 3, p. 72.
- Edwards, S., and A. Cox-Edwards. 1990. *Labor Market Distortions and Structural Adjustment in Developing Countries'* Working Paper 3346. Cambridge, Mass.: National Bureau of Economic Research.
- Ehrenberg, R. G., and G. H. Jacobson. 1988. *Advance Notice Provisions in Plant Closing Legislation*. Kalamazoo, Mich.: W. E. Upjohn Institute.
- Ehrenberg, R., and R. Oaxaca. 1976. "Unemployment Insurance, Duration of Unemployment, and Subsequent Wage Gain." *American Economic Review* 66: 754-66.
- Ehrenberg, R., and R. Smith 1991. *Modern Labor Economics*. 4th ed. New York: Harper-Collins.
- Evans, N., D. MacKay, M. Garret, and P. Sutcliffe. 1993. *The Abolition of the Dock Labor Scheme*. Research Series 14. Sheffield, England: Department of Employment.
- Fallon, P. R., and R. Lucas. 1991. "The Impact of Changes in Job Security Regulations in India and Zimbabwe." *World Bank Economic Review* 5: 395-413.

- Fallon, P. R., and L. A. Riveros. 1989. "Adjustment and the Labor Market." Policy Research Working Paper 214. World Bank, Country Economics Department, Washington D.C.
- Faundez, J. 1994. *Affirmative Action: International Perspectives*. Geneva: International Labor Office.
- Fields, Gary S. 1994. "Changing Labor Market Conditions and Economic Development in Hong Kong, the Republic of Korea, Singapore, and Taiwan (China)." *World Bank Economic Review* 8: 395-414.
- Fitoussi, J. P. 1994. "Wage Distribution and Unemployment: The French Experience." *American Economic Review* 84: 59-64.
- Freeman, R. 1992. "Labor Market Institutions and Policies: Help or Hindrance to Economic Development?" Paper presented at the annual conference for development economics, World Bank, Washington, D.C., April.
- . 1993. "Does Suppression of Labor Contribute to Economic Success? Labor Relations and Markets in East Asia." World Bank, Washington, D.C.
- Freeman, R., and J. L. Medoff, 1979. "The Two Faces of Unionism." *The Public Interest* 57: 69-73.
- . 1984. *What Do Unions Do?* New York: Basic Books.
- Gelb, A, J. B. Knight, and R. Sabot. 1991. "Public Sector Employment, Rent Seeking and Economic Growth." *Economic Journal* 101: 1186-99.
- Gindling, T. H., and K. Terrell. 1993. "Who Earns Less than the Minimum Wage in Costa Rica?" Paper presented at the conference on economic analysis of low wages and the effects of the minimum wage. University of Aix-en-Provence, Arles, France.
- Greenaway, D. 1993. "Liberalising Foreign Trade Through Rose Tinted Glasses." *Economic Journal* 103: 208-22.
- Gregory, M., and V. Sandoval. 1994. "Low Pay and Minimum Wage Protection in Britain and the EC." In R. Barrell, ed., *The UK Labor Market: Comparative Aspects and Institutional Developments*. Cambridge: Cambridge University Press.
- Grossman, J. B. 1983. "The Impact of Minimum Wage on Other Wages." *Journal of Human Resources* 18: 359-78.
- Gruber, Jonathan, and Alan B. Kruger. 1991. "The Incidence of Mandated Employer-Provided Insurance: Lessons from Workers." In David Bradford, ed., *Tax Policy and the Economy*. Cambridge, Mass: MIT Press.
- Hamermesh, D. 1992. "Unemployment Insurance for Developing Countries." Policy Research Working Paper 897. World Bank, Population and Human Resources Department, Washington, D.C.
- Handy, L. J. 1968. "Absenteeism and Attendance in the British Coal-Mining Industry: An Examination of Post-War Trends." *British Journal of Industrial Relations* 6: 27-50.
- Hayes, B. 1984. "Unions and Strikes with Asymmetric Information." *Journal of Labor Economics* 2: 57-83.
- Hazledine, T., and others 1977. "Strike Incidence and Economic Activity: Some Further Evidence." *New Zealand Economic Papers* 11: 91-105.
- Herzenberg, S. A. 1990. "Introduction." In *Labor Standards and Development in the Global Economy: Papers Presented at the Symposium on Labor Standards and Development held in Washington D.C., December 12-13, 1990*. Washington, D.C.: U.S. Department of Labor, Bureau of International Labor Affairs.
- Hicks, J. 1932. *Theory of Wages*. London: Macmillan.
- Horton, S., R. Kanbur, and D. Mazumdar. 1994. *Labor Markets in an Era of Adjustment*. 2 vols. World Bank: Washington, D.C.
- Johnson, C. 1982. *MITI and the Japanese Miracle*. Stanford: Stanford University Press.
- Johnson, H. G., and P. Mieszkowski. 1970. "The Effects of Unionization on the Distribution of Income: A General Equilibrium Approach." *Quarterly Journal of Economics* 84: 539-61.
- Knowles, K. G. 1952. *Strikes: A Study in Industrial Conflict*. Oxford: Blackwell.
- Layard, R., S. Nickel, and R. Jackman. 1991. *Unemployment, Macroeconomic Performance, and the Labor Market*. Oxford: Oxford University Press.
- Levine, D. I. 1991. "Just-Cause Employment Policies in the Presence of Worker Adverse Selection." *Journal of Labor Economics* 9: 294-305.
- Levine, Ross, and David Renelt. 1990. "A Sensitivity Analysis of Cross-Country Growth Regressions." World Bank and Harvard University.
- Lindauer, D., O. Meesook, and P. Suebsaeng. 1988. "Government Wage Policy in Africa: Some Findings and Policy Issues." *World Bank Research Observer* 3: 1-25.
- Lipsey, R., and K. Lancaster. 1956. "The General Theory of the Second Best." *Review of Economic Studies* 24:11-32.
- Machin, S., and A. Manning. 1992. *Minimum Wages, Wage Dispersion and Employment: Evidence for the UK Wages Councils*. Discussion Paper 80. London: London School of Economics, Centre for Economic Performance.
- Manning, A. 1990. "Implicit Contract Theory." In D. Sapsford and Z. Tzannatos, eds., *Current Issues in Labor Economics*. Houndmills: Macmillan.
- Marshall, A. 1890. *Principles of Economics*. London: Macmillan, (8th ed., 1966).
- Mauro, M. J. 1982. "Strikes as a Result of Imperfect Information." *Industrial Relations and Labor Review* 35: 522-38.
- Mazumdar, Dipak. 1989. *Microeconomic Issues of Labor Markets in Developing Countries: Analysis and Policy Implications*. EDI Seminar Paper 40. World Bank, Washington, D.C.
- . 1994. "Wages in Africa." World Bank, Africa Department, Washington, D.C.

- McDonald, I., and R. Solow. 1981. "Wage Bargaining and Employment." *American Economic Review* 71: 896–908.
- Metcalfe, D. 1994. "Transformation of the British Industrial Relations? Institutions, Conduct and Outcomes 1980–1990." In R. Barrell, ed., *The UK Labor Market: Comparative Aspects and Institutional Developments*. Cambridge: Cambridge University Press.
- Minford, P., and J. Riley. 1994. "The UK Labor Market: Micro Rigidities and Macro Obstructions." In R. Barrell, ed., *The UK Labor Market: Comparative Aspects and Institutional Developments*. Cambridge: Cambridge University Press.
- Neumark, D., and W. Wascher. 1992. "Employment Effects of Minimum Wage and Sub-minimum Wages: Panel Data on State Minimum Wage Laws." *Industrial and Labor Relations Review* 46: 55–81.
- Nickell, S. 1982. "The Determination of Equilibrium Unemployment in Britain." *Economic Journal* 11: 187–222.
- OECD (Organization for Economic Cooperation and Development). 1989. *Trade in Services and Developing Countries*. Paris.
- . 1992. *Employment Outlook*. Paris.
- . 1994. *Employment Outlook*. Paris.
- Paldam, M., and L. A. Riveros. 1986. "Minimum Wages and Average Wages: Analyzing the Causality in Argentina, Brazil, and Chile." Washington, D.C. World Bank.
- Papageorgiou, D., M. Michaely, and A. Choksi, 1991. *Liberalising Foreign Trade*. Oxford: Blackwell.
- Paredes, R. 1993. "Job Security Legislation and Labor Market Adjustment in Developing Countries." HRO Working Paper 16. World Bank, Human Capital Operations, Washington, D.C.
- Pascon, P., and M. Ennaji. 1986. *Les paysans sans terre au Maroc*. Casablanca: Editions Toubkal.
- Pencavel, J. H. 1977. "The Distributional and Efficiency Effects of Trade Unions in Britain." *British Journal of Industrial Relations* 15: 136–37.
- . 1991. *Labor Markets under Trade Unionism*. Cambridge, Mass.: Blackwell.
- Podgursky, M., and P. Swain. 1987. "Job Displacement and Earnings Loss: Evidence from the Displaced Worker Survey." *Industrial and Labor Relations Review* 47: 17–29.
- Rees, A. 1963. "The Effects of Unions on Resource Allocation." *Journal of Law and Economics* 6: 69–78.
- . 1993. "The Role of Fairness in Wage Determination." *Journal of Labor Economics* 11: 243–52.
- Reynolds, L. G. 1965. "Wages and Employment in the Labor Surplus Economy." *American Economic Review* 55: 19–39.
- Robinson, D. 1994. "Do Standards for the Workplace Help or Hurt?" Background paper to *World Development Report 1995*. World Bank, Washington, D.C.
- Rothschild, M., and G. Stiglitz. 1976. "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information." *Quarterly Journal of Economics* 90: 629–49.
- Ruhm, C. J. 1992. "Advance Notice and Post-displacement Joblessness." *Journal of Labor Economics* 10: 629–49.
- Russell, S. S., and M. S. Teitelbaum. 1992. *International Migration and International Trade*. World Bank: Discussion Paper 160. Washington, D.C.
- Santiago, C. 1989. "The Dynamics of Minimum Wage Policy in Economic Development: A Multiple Time Series Approach." *Economic Development and Cultural Change* 38: 1–30.
- Sapsford, D., and Z. Tzannatos, eds. 1990. *Current Issues in Labor Economics*. Houndmills: Macmillan.
- . 1993. *The Economics of the Labor Market*. Houndmills: Macmillan.
- Seike, Atsushi, and Hong Tan. 1992. "Job Security and Employment Adjustment: A U.S.-Japan Comparison." In H. Tan and H. Shimada, eds., *Troubled Industries in the United States and Japan*. RAND-JCER Conference Volume.
- Squire, L., and S. Suthiwart-Narueput. 1994. "The Impact of Labor Market Regulations." Policy Research Working Paper 1418. World Bank, Policy Research Department, Washington, D.C.
- Standing, Guy. 1989. *The Growth of External Labor Flexibility in a Nascent NIC: A Malaysian Labour Flexibility Survey*. Working Paper 35. Geneva: ILO World Employment Program.
- . 1992. "Do Unions Impede or Accelerate Structural Adjustment? Industrial Versus Company Unions in an Industrializing Labor Market." *Cambridge Journal of Economics* 16: 327–54.
- Tabellini, G., and M. Rama. 1994. "Endogenous Product and Labor Market Distortions." Paper presented at the labor market workshop, World Bank, Washington, D.C.
- Taira, Koji. 1970. *Economic Development and the Labor Market in Japan*. New York: Columbia University Press.
- Tracy, J. S. 1987. "An Empirical Test of an Asymmetric Information Model of Strikes." *Journal of Labor Economics* 5: 149–73.
- Turnbull, P., and D. Sapsford. 1993. "Conflict and Efficiency in the Docks: A Cross-Country Study of Industrial Relations, Institutional Structures, and their Effects." Paper presented at the conference on institutions and economic growth, Center for Economic Policy Research, Berlin, June.
- Turnbull P., C. Woolfson, and J. Kelly. 1992. *Dock Strike: Conflict and Restructuring in Britain's Ports*. Aldershot: Avebury-Gower.
- Walsh, W. D. 1975. "Economic Conditions and Strike Activity in Canada." *Industrial Relations* 14: 45–54.
- Wellington, A. J. 1991. "Effects of Minimum Wages on the Employment Status of the Youth." *Journal of Human Resources* 26: 27–46.

## REGULATORY POLICIES AND REFORM: A COMPARATIVE PERSPECTIVE

- World Bank. 1993. *The East Asian Miracle: Economic Growth and Public Policy*. New York: Oxford University Press.
- . 1994. *Averting the Old Age Crisis: Policies to Protect the Old and Promote Growth*. Washington, D.C.
- World Bank, 1995. *World Development Report 1995*. New York: Oxford University Press.
- Yoder, D. 1940. "Economic Changes and Industrial Unrest in the United States." *Journal of Political Economy* 48: 222–27.
- Zabalza, A., and Z. Tzannatos. 1985. *Women and Equal Pay*. Cambridge: Cambridge University Press.

# Regulatory policies and reform: the case of land markets

Antônio Salazar P. Brandão and Gershon Feder

Land markets in developing and industrializing countries are subject to regulatory constraints that significantly affect the operation of the market and equilibrium prices and sales, contribute to reduced efficiency, and have negative equity implications. The role of government in land market reform is to remove such regulations, establish a system of predictable market rules, and focus on the provision of information, adjudication of border disputes, enforcement of property rights, and valuation and assessment of land for tax purposes.

Several characteristics distinguish land markets. Land is a factor of production, essential to the provision of urban housing services and the production of agricultural goods. At the same time land is demanded as a financial asset. It is often a good hedge against inflation, especially in countries where financial markets are not well developed. Even in economies with well-developed financial markets and where inflation is not a serious problem, the acquisition of land is frequently part of the portfolio diversification strategies of economic agents. Financial institutions frequently prefer land as collateral for credit operations because, among other reasons, land is immobile, its depreciation is small, and its value is not eroded by inflation (Binswanger and Rosenzweig 1986). Finally, land is a heterogeneous good, a “property,” whose market prices usually reflect not only its value but also its location and attached investments.

Insofar as land is a factor of production and a store of wealth, it is also a source of political power, especially in societies where access to other assets is limited. The evolution of property rights through history shows that landowners have had an upper hand in shaping policies to favor their interest, a situation that is still prevalent in some industrial and many developing countries (see Binswanger, Deininger, and Feder 1995).

The demand for land stock is derived first, from the demand for agricultural goods and housing—essentially a demand for land services. Second, it arises from infra-

structure and environment-related projects, a demand often independent of land prices and determined by government objectives and other concerns. Third, it takes the form of an asset demand, in view of the financial asset characteristics of the land stock. The roles of land as a hedge against inflation, as collateral for credit operations, and as a component of the diversification strategies of economic agents are subsumed in this third type of demand.

The supply of land for the rural and urban sectors is determined by nature—availability, topography, and, in the case of agriculture, soil fertility—and by the volume and quality of prior investments, including structures. The growth of such investments is accompanied by the expansion of services derived from a given stock of land.

Regulatory constraints affect both the demand for and the supply of land. Limitations on land use in urban and peri-urban areas, and ecological zoning, are examples of government-imposed supply restrictions. On the demand side, limitations on the use of land as a collateral for credit operations, or on the exercise of property rights by restricting (or forbidding) sales and rentals, illustrate common constraints by fiat.

Although recent analyses of land markets show a growing concern for policy and regulatory issues, the literature still lacks a robust framework capable of showing how land markets function, the major policy and regulatory constraints to their efficient operations, and the implications for reform.<sup>1</sup> This chapter is a step in that direction.

The first section sets out to characterize land markets—their emergence, closely associated with the evolution of property rights; major imperfections; and key spatial aspects. The focus of the second section is on policies that affect how land markets operate, both directly (such as tenure security, zoning laws, prohibitions of land transactions, speculation, rent controls, and land taxation) and indirectly (such as credit policies and tax and

tariff policies). The third section describes the fundamentals of regulatory policy reform in land markets and suggests a two-phase process. In the first, regulations that are inconsistent with efficient outcomes or that drive inequitable results should be eliminated. In the second phase, a new legal and institutional framework for land administration would be created. The land administration unit should be a technical unit that collects and provides (for a fee) information to the public on relevant aspects of land markets and performs all functions related to titling and conflict resolution. The final section suggests areas for future policy research.

### **Conceptualizing land markets and property rights**

The emergence of land markets is closely related to the evolution of property rights over land. In the rural context the critical factors for the establishment of property rights and the development of enforcement mechanisms were population growth, advances in agricultural technology, and increased trade (Binswanger, Deininger, and Feder 1995; Feder and Feeny 1991). A growing population and greater trading opportunities forced the adoption of fertility-restoring technologies to permit continuous exploitation of the land, ending the reliance on shifting cultivation and long fallow periods to maintain fertility. Insofar as superseding technologies required an investment of effort and capital (tree felling, stone clearing, shrub removal, and terracing, for example), the ability to continuously exploit a tract of land over a reasonable length of time, and reap the related productivity and pecuniary gains, became crucial for agricultural development. In the urban context the appearance of permanent and more secure walled settlements, allowing dwellers to reap economies of agglomeration, created a need to define property rights over tracts of land (and the structures on them). The limited space within the walled city created a scarcity of land—the prerequisite for the constitution of a market.

#### *Evolution of property rights*

In the early stages of agricultural development, individuals were assigned long-term (or even inheritable) use rights to land, with a restricted ability to transfer such rights. This arrangement, while providing sufficient incentives for investment, avoided the social tensions engendered by the emergence of a landless class. In fact the concern for social conflicts was manifested in the earliest agrarian societies. For example, the biblical law of the Israelites (around 1300 B.C.) prescribed that every 50 years land ownership would revert to the original households (or their descendants), regardless

of the circumstances under which transfers had taken place.

The loss of efficiency from restricted transferability was insignificant in such circumstances, since differences among individuals in management capacity mattered less in these times of relatively simple technology. However, as technology advanced, and the differential endowments of management skills, labor, and other nonland productive assets among individuals assumed increasing importance, the lack of transferability of property rights adversely affected productivity, even if individual use rights were secure over the long term. Because larger economic benefits could be realized by making land transferable from low- to high-productivity individuals, transferability became possible, despite the costs associated with the growth of a landless class. Social tensions were attenuated when the urban economy began growing, absorbing the landless in activities with a high marginal productivity of labor.

The emergence of land markets and the consolidation of property rights over land created, in most societies, a powerful class of rural landowners.<sup>2</sup> In some industrial countries the power of this group declined with the relative share of agriculture in the economy. In many developing countries, especially the poorer ones, rural landowners still hold a significant share of political power. Most laws, regulations, and policies promote direct and indirect transfers that benefit these landowners. Among the most conspicuous examples of these are subsidized interest rates, equipment prices, and water tariffs.

#### *Property rights: a categorization*

A bundle of characteristics define property rights over land: exclusivity, inheritability, transferability, and enforcement mechanisms (Alchian and Demsetz 1973). A system of property rights defines the legitimate exclusive uses of land and identifies those entitled to these rights. The complexity of the system allows for situations where, for a specific tract of land, different uses have different holders. For example, in medieval England and contemporary southern India, although the rights to the crop from a given tract of land belong to an individual, the community has a right to graze livestock on the post-harvest residue. Land rights may also include a specific stipulation of the circumstances and conditions for transfers (land cannot be transferred to individuals outside a group or community, for example). Property rights over land also have a temporal dimension: The right to use land can be defined over a short period of time (for example, a year's rental) or a longer period (for example, inheritable and permanent use rights).

The value of property rights (and the functioning of land markets) depends on formal mechanisms for defining and enforcing those rights, including the court system, police, the legal profession, land surveys, record keeping systems, and titling agencies (Feder and Feeny 1991, p. 137), as well as on social norms or religious customs.

For analytical purposes property rights can be categorized into four basic types: open access; communal property; private property; and state property. In an open access regime, property rights are not specifically assigned to any individual or small group, although they may be perceived as belonging to some large group, so that the ability to exclude individuals from using the land is practically nil. In the absence of excludability, there is no incentive for individuals to invest in restoring fertility or in conserving the topsoil, and the resource is usually subject to degradation.

In the case of communal property, rights are assigned to a specific community. Community members are able to exclude outsiders from using the land and to control and regulate its use by members. Although there may still be incentive problems, related to the unwillingness of any individual member to undertake the appropriate fertility-enhancing (or resource-conserving) investments, the group as a whole may overcome these problems by viewing those investments as a public good and using communal tax (or *corvée* labor) authority to finance investment costs. If the community is so large that exercising control is not practicable, the distinction between communal and open access systems disappears.

Under private property rights, land is assigned to specific individuals or corporate entities. Still, certain formal or informal limitations on these rights may be imposed by the state or the community. For example, the state may forbid certain uses of the land or its sale. The fewer restrictions there are, the stronger are the incentives for individuals to invest in the land. In the absence of a proper enforcement apparatus, private property rights may assume the characteristics of an open access regime.

State ownership implies that the state (or extensions of the state, such as local authorities and municipalities) possesses the property rights. The authorities may, however, transfer temporarily some of the rights to private users or to communities (for example, through the rental of state land or by providing permission to graze over state land). When the state does not assert its authority, state property may become *de facto* private property if individuals (squatters) establish their rights by physical possession and acquire informal communal recognition of their *de facto* rights.

Secure individual (or corporate) property rights are critical in establishing a structure of economic incentives for investment in land-based activities. The more these rights are restricted, the weaker will be the investment incentives and the lower the productivity of land. Restrictions on rights can come from formal inhibitions, customary conventions, or inadequate enforcement systems. Certain restrictions pertain to the horizon over which property rights may be held (for example, a lifetime possession provides less investment incentive than an inheritable possession that can be transferred to descendants; a 30-year lease provides greater incentive than a five-year lease). Other restrictions pertain to limitations on use (the absence of any restrictions provides better incentive than a system that limits use to one particular purpose) or to the security of tenure (immunity from uncompensated state confiscation provides more incentive than the right of the state to expropriate with arbitrary compensation procedures; state protection from unsubstantiated challenges by other individuals to property rights provides better incentive than a system without state enforcement of individual property rights).

Restrictions on transferability are often related to inhibitions instituted by the state or the community, typically induced by concerns for social tension. Yet these are commonly circumvented by disguised transactions, because the potential efficiency gain provides incentives for both sides of the transaction to conclude a transfer. For example, in areas where sales to outsiders are forbidden but leases are allowed, a sale will be disguised as a renewable lease. The illegality of the arrangement introduces an element of risk, however, and thus tradability is still negatively affected in the aggregate, with a consequent efficiency loss.

#### *Market imperfections and external effects*

So far it has been argued that the absence of well-defined or adequately enforced property rights in land hampers the functioning of land markets and leads to inefficient outcomes. Several other imperfections, stemming either from particular properties of land or from distortions in other markets that spill over, may also prevent land markets from allocating resources efficiently.

*Asymmetric information and land transferability.* The possessor of land often has more knowledge about the extent to which the rights to the land are (or are likely to become) contested than other individuals (especially those from another community). This limits the tradability of land, because some individuals who might otherwise be interested in acquiring the land (for a higher-value use

than the current one) either may be reluctant to risk purchase or may offer a lower price (reflecting the perceived risk of challenging claims). Both outcomes tend to reduce the extent of land trading, with a consequent loss of efficiency, since land trading generally facilitates the allocation of land to higher-productivity users. It is precisely this loss of efficiency that motivates societies to establish systems of land records and title registration, which enable potential buyers to verify the authenticity of property rights offered for sale.

*Transferability and linkages with the credit market.* Limitations on land trading have a negative spillover effect on the credit market. Credit transactions, and in particular medium- and long-term loans, involve a significant degree of asymmetric information. The potential borrower may know much better than the lender the probability of loan repayment. This asymmetry limits the extent of credit transactions, yielding loss of efficiency, since some high-return investments that would have been financed if information were symmetric are not undertaken. This loss of efficiency induces the introduction of the collateral arrangement, whereby the borrower alleviates the lender's lack of information by offering a reasonably risk-free asset whose conditional sale could be used to repay the loan in the event of default. Land and other real estate are ideal collateral (Binswanger and Rosenzweig 1986), because their physical properties are less amenable to destruction and abuse than other property such as machinery or livestock.

For land to be useful as collateral, however, it must be easily transferable, and the property rights over it must be clearly defined (Feder, Onchan, and Raparla 1988). Thus, the same institutional arrangements that reduce information asymmetry in the land market (for example, land registries and title documents), and thus improve its operation, are also useful for improving the efficiency of credit markets. Similarly, the inefficiencies in land allocation arising from limitations on land transfers are exacerbated by the resultant diminished use of land as collateral. A corollary proposition is that the more developed the credit market, the larger will be the demand for formalizing land rights. Indeed, a study of land policies in Thailand by Feder and others (1988) showed that land registration had a significant effect on the production efficiency of squatters (even when there was reasonable tenure security) and that these efficiency gains were mostly due to credit market linkages.

*Other imperfections in the land market.* The acquisition of land requires a significant outlay of cash. In many developing countries, however, capital markets are imperfect,

and the ability to obtain credit for land purchase requires a significant accumulation of equity before the transaction. This requirement excludes a large proportion of the population from the land market and thus hampers the market's ability to allocate land to the highest-productivity use. The existence of a rental market mitigates somewhat the efficiency loss that this imperfection could generate. In some countries, however, political tensions may engender fears among landowners that the awarding of long-term leases, or prolonged periods of absentee ownership, might weaken the owners' property rights and make them vulnerable to challenging claims by tenants. Such fears may encourage low-intensity utilization under the owners' management (for example, grazing).

Moreover, land's durability and its ability to maintain real value in an inflationary environment make it a desirable asset for storing value in economies where inflation-proof financial instruments are not readily available. As a result, individuals who lack the skills to utilize the land in agriculture or other productive uses may acquire significant amounts of land. Again, the existence of a rental market may make such land available to those who can make a more productive use of it, but the above caveats apply in this case as well.

Finally, insofar as land typically is not traded in international markets, its price reflects various distortions in other goods markets and in the agricultural terms of trade (Jones 1965; Hueckel 1972; Feeny 1982). For example, policies that heavily subsidize agriculture or protect it from foreign competition tend to translate into land prices that are higher than they would otherwise be, and into an excessive allocation of land to agriculture (as in the case of Japan).<sup>3</sup>

#### *Spatial aspects of land markets*

The contribution of land to an individual's income or welfare is dependent on its location. The importance of locational factors for agricultural development was stressed in the seminal work of von Thunen (1966). The work of Schultz (1953) explaining regional income differences in U.S. agriculture and that of Katzman (1974), who analyzes the expansion of the agricultural frontier in Brazil, similarly focused on the spatial aspects of land markets (see Bhadra and Brandão 1993). Katzman, in particular, noted that the price gradient of land should decline with distance from the urban center. At the agricultural frontier, where there is open access to land, the value of land would be zero. With the expansion of agriculture, lower-return activities will move away from the center to the agricultural frontier.

The urban economics literature has similarly developed a conceptual model of a monocentric city. Economies of agglomeration provide cost incentives for the location of business, and increasing transportation costs determine the location of activities along various rings from the center. The price of land declines with distance from the center. At the urban-rural border, the value of land will be the same in the two sectors. Urban activities that use land more intensively will either move away from the center (a phenomenon usually referred to as suburbanization) or substitute capital for land. The land market must be flexible to permit these adjustments to take place.

In developing countries land market problems tend to concentrate at the outer rings, or peri-urban areas, where the market is often driven by prospects of capital gains. Fast urbanization and high population pressure, which characterize many developing countries, exacerbate the disputes over land in the "urban frontier." In the border areas urban and rural activities coexist, and the expected gains by landowners, speculators, and developers give rise to a specific type of land market dynamics. Whereas farmers have an incentive to reduce investment with long gestation periods, speculators have an incentive to precede developers and purchase "cheap" land. The government often steps in, sometimes to "protect" farmers from speculators and to prevent the conversion of land to the urban sector, other times to protect the interests of urban developers. Furthermore, where property rights are not clearly defined, land grabbing becomes pervasive and, in some countries, has the blessing (if not the direct participation) of the government.<sup>4</sup> (Appendix A presents empirical evidence on rural-urban land conversion and the behavior of land prices at fringe areas.)

The discussion above suggests that it is useful to distinguish among the urban, rural, and peri-urban land markets. Because each will have its own dynamics and may respond differently to economic stimuli and policy changes, regulatory intervention in land markets must consider the three separately.

## POLICY ISSUES

A wide range of government interventions influence the operation of land markets. They range from policies aiming to modify the spatial distribution of economic activity (for example, industrial location) to those promoting specific sectors or activities (for example, subsidies to housing). Moreover, they can affect the land market directly (zoning laws) or indirectly (policies that affect

capital markets). These interventions often reduce efficiency and almost as often discriminate against poor people. This section presents an analytical view of selected policy issues in land markets, based on their relative importance and whether the existence of reasonably robust research results allows unambiguous recommendations.<sup>5</sup>

### Policies with a primary focus on land markets

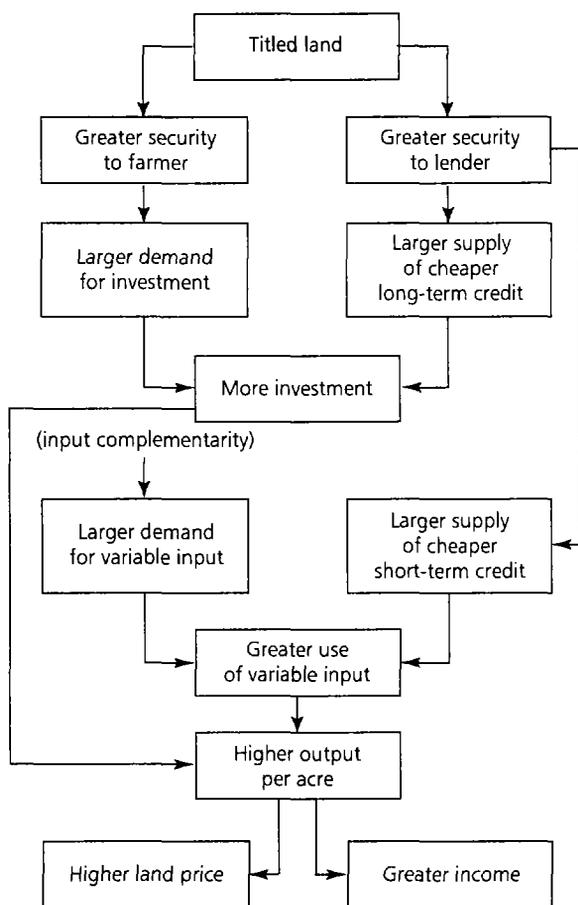
#### *Direct constraints on the exercise of property rights*

*Tenure insecurity.* Tenure insecurity is pervasive in developing countries. It manifests itself through multiple factors: the presence of landowners with no legal titles; inappropriate legislation governing, and legal restrictions on, the issuing of titles (for example, to farms smaller than a certain threshold); institutions unprepared to handle the technical and legal aspects of land registration, leading to multiple titles for the same parcel and improper specification of boundaries; the lack or discretionary enforcement of property rights (for example, in parts of the Amazon region of Brazil, Bolivia, Colombia, and Peru) and the lack of transparency and the high costs of registration and other procedures. Nevertheless, tenure insecurity is fundamentally a consequence of inadequate land administration and of a legal framework incapable of determining boundaries and settling disputes.

Another source of tenure insecurity is the threat of expropriation. Legal provisions in most countries allow land expropriation by the public sector for infrastructure development. Expropriation is also frequently allowed in the context of land reform or for colonization projects. More recently, expropriation for ecological projects has been added to the agenda of policymakers. In several developing countries, however, expropriation rules are either not clearly defined or clouded by procedural difficulties,<sup>6</sup> their implementation is discretionary,<sup>7</sup> and landowners are compensated at prices that understate market values.

The possession of a title can be an important determinant of the degree of tenure security. Feder and others (1988), studying the impact of land policies on farm productivity in Thailand, showed that the higher the degree of tenure security, the higher the demand for investment, especially for goods and services that become attached to land.<sup>8</sup> Access to the formal credit system is easier for titled farmers because they represent a lower risk for the lending institutions. A larger supply of formal long-term credit (usually cheaper than credit obtained in informal markets) helps to further increase the rate of investment.

**Figure 10.1 A conceptual framework for the economics of land titling**



Source: Feder and others (1988).

Tenure security also increases access to short-term credit, which in turn leads to greater use of variable inputs. Consequently, output per hectare, the price of land, and income are higher for titled farmers (figure 10.1).

A relevant consideration is often the probability of eviction. The lower this probability, the greater the incentive for farmers to invest in land-attached improvements and for the formal credit system to extend credit. In regions where tenure insecurity is pervasive, the price of land will not reflect the present value of the income stream associated with the exploitation of land for agricultural production because not all land rights are legitimate or enforceable under the law. The land value will incorporate a speculative element because of the possibility of gains from (eventual) regularization or enforcement of the tenure status. Because this situation brings to the market individuals who may not be primarily interested in agricultural production, total factor productivity is likely to decline.

Tenure insecurity also precludes “landowners” in the urban sector from using this preferred form of collateral in the formal credit market. As in agriculture, it can be a binding constraint for potential borrowers, particularly for the more financially fragile. Similarly, tenure insecurity reduces the incentive to invest, especially in land-attached improvements. The capital-to-land ratio in the housing sector will accordingly be lower than otherwise. In peri-urban areas lack of tenure security is widespread,<sup>9</sup> and so the incentive for investment will be reduced. In addition, as Jimenez (1982) has shown, tenure insecurity reduces the demand for house improvements and government services (see also Jimenez 1984).

Initiatives to increase tenure security can be costly. These costs can be reasonably estimated, whereas estimating the benefits frequently requires extensive household-level data collection and sophisticated econometric analysis. Despite the difficulties, two studies assessed the value of the benefits of increased tenure security through differences in land prices. Feder and others (1988) estimated price differentials for titled and untitled land in four regions in rural Thailand. Using a model in which characteristics of each plot were control variables, they estimated that the value of untitled land as a percentage of the value of titled land varied from 43 to 80 percent and that the net social benefits of providing titles for farmers ranged from 21 to 40 percent of the value of untitled land (p. 145).<sup>10</sup> It is interesting to note that the probability of eviction in the study areas was quite low. The price differential and the high social benefits of titling reflect mostly improved access to capital markets.

Dowall and Leaf (1990), who focused their analysis on urban land markets in Jakarta, found that the mean price differential between land with registered titles and land with weak claims was about 45 percent of the price of the latter in 1989. The price differential decreased as the distance from the center of Jakarta increased: At distances less than 5 kilometers from the center, the price differential was 65 percent, whereas it was 39 percent at distances greater than 15 kilometers. Using a hedonic price model, in which the value of land is regressed against distance from the center and dummy variables for plots with high infrastructure and for those with registered titles, Dowall and Leaf found the title dummy highly significant in the three years they studied (1987, 1988, and 1989). They concluded that the net benefit of providing registered land for both serviced (roads, sewage systems, and the like) and unserved plots is positive, justifying the implementation of a land-titling program on a cost recovery basis.

*Zoning and other restrictions.*<sup>11</sup> Government interventions in the land market are often part of spatial strategies aimed at reducing the growth of large cities, developing small and intermediate-size towns and lagging regions, creating growth poles, and promoting land colonization schemes (see Rondinelli 1990 for a discussion of government interventions in Asia). Several policies have been adopted in support of these strategies, such as subsidized prices of public infrastructure services, concessional loans, and tax incentives, as well as zoning laws and other restrictions to the exercise of property rights. The following are some important examples:<sup>12</sup>

- *Agricultural zoning.* These measures restrict or prohibit the construction of nonfarm buildings in agricultural areas. Without other support programs to increase productivity in agriculture and, more generally, to improve the economic opportunities in the rural sector, the effect of such policies on urban expansion will be limited. Nonetheless, zoning laws are common in developing and industrial countries.<sup>13</sup>

- *Agricultural districting.* Rather than restrict development directly, these policies instead establish districts within which farmers are protected from certain state or local regulations, or from private nuisance suits. Agricultural districting reduces the adverse impact that proximity to urban centers often has on agriculture. Private returns in agriculture are increased at the same time that the rate of conversion of land to its best alternative use is reduced, leading to an inefficient allocation of land. Nevertheless, agricultural districting is common in industrial countries, especially in Canada, the United States, and Western Europe (Barrows and Newman 1990).

- *Public purchase or private transfers of development rights.* To alter the pattern of land use, the government purchases the rights to develop certain tracts of land, with the owner retaining land ownership and other associated rights (Barrows and Newman 1990). The government may, for example, acquire the permanent right to use a certain plot for nonagricultural activities. If the government later deems it appropriate to allow land conversion to the urban sector, the right to use the land for nonagricultural activities is sold in the market. These measures typically help protect agriculture in fringe areas by slowing the expansion of the urban sector.<sup>14</sup>

- *Urban zoning.* Urban zoning, a prevalent feature of land market interventions, is often practiced by city governments, in some cases as part of a "city project." Designated residential areas and industrial districts are common examples of urban zoning. Although it is quite

possible that zoning restrictions reduce efficiency because they prevent land from being allocated to the best alternative use, environmentally motivated and other restrictions, when introduced to correct well-identified and significant market failures, may be justified. Nonetheless, a comprehensive evaluation of the effects of zoning on land use patterns and on conversion does not yet exist.

*Prohibitions of land transactions.* Several developing countries prohibit both the sale and the rental of agricultural land. In the nonsocialist countries, such prohibitions most commonly arise in the aftermath of a land reform process as part of government efforts to impede market mechanisms from changing the structure of land tenure. Prohibitions of sales are sometimes justified as a mechanism to reduce the rate of rural-urban migration and to protect small farmers from the likelihood of foreclosure by commercial banks. The rationale for prohibitions on rentals is often founded in the view that "land is for the tiller."

These prohibitions can have far-reaching implications. A fluid market enables land to move from less to more efficient producers. In addition, where land sales are not allowed, the value of land as a collateral for credit operations disappears, reducing investment and growth.

A well-known example of outright prohibitions on land transactions is the *ejido* system in Mexico. Land was perceived and registered as communal property and consequently could not be sold or rented except within the community (Heath 1992). The Mexican government is now reforming its land legislation. But even before the current reform, there were indications that several communities were willing to join the "private sector," that is, to become fully integrated in the land market and to acquire the right to sell and rent land without any restriction. In the 1980s the government allowed community members to develop partnerships with outsiders, a move that strengthened the informal rental and sales markets on *ejido* land. The recent change in the Mexican agrarian law, and the interest demonstrated by *comunidades de ejidatarios* in joining the "private sector," confirms this trend.

One important aspect of the new Mexican agrarian law that deserves attention in countries undertaking land market reforms is that it allows, but does not require, each *comunidad de ejidatarios* to join the private sector. Local communities, in Mexico and in several other countries, often restrict land sales and rentals to community members; transactions involving outsiders require community permission. This procedure clearly undervalues community land and generates a suboptimal allocation of

resources. However, it also benefits the community by reducing the possibility of social tension, by keeping the bonds that maintain the community together. Small farmers usually have neither access to risk markets to hedge against years of low prices nor access to credit markets to borrow in years of bad crops. Communities commonly provide insurance and supply credit efficiently because information and transaction costs are relatively low within the community.

Outright prohibitions of sales are not as common in urban land markets. Zoning restrictions, as discussed before, and the preservation of buildings for historical reasons seem the most common forms that governments use to restrict land transactions in cities. The latter, however, are significant only in a relatively small number of cities.

*Price-related interventions affecting the exercise of property rights*

*Speculation.* One of the most politically sensitive issues in land markets is speculation. A commonly held view is that speculation distorts resource allocation and is detrimental to the functioning of the land market. Where markets are competitive and information is evenly distributed, speculation provides liquidity to the market and transfers risk to those with a comparative advantage in risk management. However, this situation is not common in land markets in developing countries, where policy distortions and market failures are the general rule, especially at the fringe of large cities and in the agricultural frontier.

Asymmetric information is an important source of speculative gains. Because information costs are usually higher for individuals than for large companies, individuals often are at a disadvantage in transactions involving corporations. Improving the dissemination of public information on government projects, which typically trigger many speculative actions, is one of the most efficient means to reduce the informational advantage of developers and avoid adverse income distribution consequences.

Where property rights are not well defined (or where renting agricultural land is prohibited) speculation at the fringe of large cities may reduce agricultural production. During the period between their acquisition of a plot and its sale to developers or builders, speculators would likely increase their profits if the land were put into production. Speculators usually are not agricultural producers, however, and would have to rent out the land once they had purchased it. But when property rights are not clear, rental may entail risks of challenges by the tenant, in

which case the expected cost of lost land rights might be greater than the forgone rents. This output loss, which can be large depending on the duration of the speculative period, can be avoided or at least minimized if property rights are clarified and enforcement enhanced, or if rental restrictions are removed.

Wrongheaded government policies and ill-conceived legislation are important factors inducing speculation, often with strong adverse efficiency and equity implications. In the city of Karachi, about 70 percent of land available for development is public. The supply of land is determined largely by the development authority, which sells at prices below market values, expecting these sales to help the poor. The initial purchasers, however, are most likely from the middle- and upper-income groups, who are allowed to resell. The large price differential that is observed between the two markets is evidence that supply is short (Dowall 1990a). Contrary to the government's expectations, this mechanism concentrates income, and the short land supply is likely to slow land development in the area. Potential government revenue that could be used to expand housing development for low-income groups is also reduced.<sup>15</sup>

In Bolivia all agricultural land is government-owned and cannot be sold by the government, although it can be transferred to individuals at nominal charges. The pace at which land is transferred is determined to a great extent by bureaucratic procedures rather than economic considerations. Since the recipients of these transfers are allowed to resell the land, the government loses potential revenues directly. This loss is further aggravated by the fact that the government does not tax the appreciation in the value of land resulting from infrastructure development.

A large, densely populated region of an Asian country offers an example of land market policies that lead to regressive income transfers, inefficiency, and reduced government revenue. For projects of "social interest," the government establishes the price for the release of land rights from private owners to developers and builders. Frequently, this price is low relative to market values. Furthermore, the law reduces the bargaining power of local landowners by allowing them to sell their land only to corporations that have obtained development permits for the specific areas in which the land is located. This mechanism implies an income transfer from landowners to developers and builders and, because the determination of a project of "social interest" is subjective, a strong incentive for rent seeking. Exercising their market power, developers set prices for land rights in projects not con-

sidered to be of social interest, on the basis of parameter of the "social projects," especially in areas where small owners with unregistered plots predominate. Strong population pressure and the high growth rate of the capital city ensure substantial gains for developers that purchase land in the periphery for future development. By taxing these gains through a betterment tax, the government could increase revenue. A more important step would be to modify the procedures associated with the release of land for social projects so as to increase competition in the land market.

*Rent controls.* Rent controls, common in both urban and rural sectors of developing countries, have negative consequences for efficiency and equity. The experience with rent control legislation in agriculture is revealing. The common rationale for this legislation is to protect tenants from eviction and to provide them with an income subsidy at the expense of the landlord by limiting the rental value that can be charged (Binswanger, Deininger, and Feder 1995). But as soon as news of impending legislation spreads, landlords often evict tenants and resume cultivation under direct owner management using hired labor. Because hired labor entails supervision costs, producers select activities that require less supervision, even if they must forgo some output. Moreover, landlords, facing more stringent constraints on existing rental contracts, find it less profitable to invest in land improvements, while reduced contract duration fostered by rent control laws diminishes the incentives for tenants to make long-term investments.

One additional difficulty associated with rent control in rural areas is the very high cost of enforcement. In practice tenants, sharecroppers, and landowners find ways to circumvent the legislation. The government uses real resources to enforce the legislation, as does the private sector in attempting to avoid it.

Finally, a common feature of rent legislation in agriculture is the prohibition of shared tenancy or the imposition of an upper limit on the landowner's share. In situations where risk and supervision costs are high and where credit is restricted, prohibitions of share contracts may actually decrease efficiency (Otsuka and Hayami 1988).

In housing markets legislation frequently restricts or prohibits eviction and imposes ceilings on rents. As a consequence incentives to invest diminish, the rate of depreciation of existing residential buildings increases, the rate of construction of new buildings is reduced, and houses are removed from the rental market. Demand pressure causes prices and rental values to increase substantially.

Frequently, due to lack of an adequate judicial system, agreements disregard legal prescriptions, as does the settlement process. The informality of these transactions, the risk they involve, and the segmented markets they create imply an inefficient outcome, although tenants will realize some income gains (at the expense of owners).

*Taxation of rural and urban land.* Economists have long advocated a tax on land.<sup>16</sup> Ricardo (1949), for example, favored such a tax because "a tax on rent would affect rent only; it would fall wholly on landlords, and could not be shifted to any class of consumers" (p. 110). This view is still prevalent among economists for essentially the same reasons: A tax on land causes no distortion in output or input prices, nor does it affect private incentives to produce.<sup>17</sup>

In developing countries agricultural land taxes evolved from payments to landlords or colonizing powers to payments to the central governments of the newly formed states. However, the lack of strong enforcement mechanisms in the wake of the political transformation of these countries reduced revenue, and today taxes on agricultural land are seldom a significant source of revenue in the developing world. Difficulties in implementation often arise not only because of political resistance to land taxes, but also because of the high informational requirements for their administration.

Binswanger, Deininger, and Feder (1995) suggested two necessary conditions for an effective land tax. First, the administration and revenue derived from the land tax must be placed at the local level (municipalities, counties, or the equivalent) so as to lower information costs, facilitate enforcement, and make the benefits of the tax more visible to the community. Earmarking revenue for local investments (as in the United States, for example) creates further incentives for payments. Second, the tax rate must be flat or only slowly progressive so as to decrease political resistance and increase the law's enforceability.<sup>18</sup>

Several countries have attempted to reduce land speculation by imposing higher taxes on unused land. The results have been mixed (Binswanger, Deininger, and Feder 1995). One reason is that the level of taxation is often very low, and efforts to make the land tax progressive meet with political resistance from landowners. More generally, however, there are difficulties in defining precisely whether land speculation is taking place. For example, when the additional tax burden is significant, agricultural landowners are likely to lower their burden by choosing suboptimal activities such as grazing or activities that use low amounts of variable inputs (labor, fertilizers).

At the heart of the problem in the case of both urban and rural land is the fact that a positive economic return is associated with land ownership regardless of whether land is utilized. Although the productive use of land does not yield a profit for some landowners (for example, because there is a high opportunity cost on their time and capital), expected appreciation of the value of their land is sufficient inducement to maintain ownership. Taxation, in principle, could reduce the share of the capital gains accruing to the landowner. But elimination of the incentive to keep land outside the productive process may require a high marginal tax rate,<sup>19</sup> which is likely to be either politically infeasible or economically unenforceable.

Enforcement in urban areas is somewhat easier than in agriculture, but similar difficulties exist. Valuation is a continuous source of disputes (undervaluation being the norm in many developing countries) due to the lack of adequate information systems. Nevertheless, because responsibility for tax collection is more frequently at the local level, urban land taxes are a significant source of revenue.

Finally, it should be noted that differential taxation rates between urban and rural land may be a significant determinant of the rate of land conversion from rural to urban uses. The stylized fact is that land taxation is higher in urban areas than in rural areas. The tax differential is capitalized in agricultural land values, creating an obstacle to prospective buyers, especially in peri-urban areas. In Japan, for example, agricultural land is taxed only lightly (and agriculture is highly protected from external competition). Not surprisingly, the cultivated land in metropolitan Tokyo-Yokohama, Nagoya, and Osaka-Kobe accounts for 16 percent of the land in these urban centers (Australian Bureau of Agricultural and Resource Economics 1988, p. 316).

### **Policies with indirect effects on land markets**

In the fringe areas of large cities and in the agricultural frontier, the present value of expected future land rents is significantly lower than the price of land due to potential capital gains. Government policies in many developing countries contribute to enlarge this wedge by subsidizing housing finance, water prices, and interest rates in agriculture; introducing tax rebate schemes designed to foster specific activities or encourage development of certain regions; and bringing about high and unstable rates of inflation through mismanagement of the macroeconomy.

Consider the Brazilian experience with subsidized interest rates in agriculture. During the 1970s, while infla-

tion averaged 30 percent per year, the average interest rate on agricultural credit was 7 percent per year. The discrepancy between inflation and the nominal interest rate increased even further in the beginning of the 1980s. This subsidy, the most important instrument for agricultural sector support during the 1970s, attracted investment to agriculture and stimulated purchases of agricultural land. As a consequence, land prices rose relative to land rents, attracting new investors to agriculture who had neither the resources to manage an agriculture enterprise properly nor the necessary production knowledge. The mismatch of endowments led to selection of suboptimal activities, such as cattle raising. Through its credit policies, the Brazilian government transferred income to landowners and contributed to a further concentration of property, while reducing the efficiency of the agricultural sector.<sup>20</sup>

Governments in several developing countries have affected the rate of conversion of rural land to urban land by creating wedges between market and social (or shadow) prices of goods and factors. The urban bias, which often materializes in the form of implicit or explicit taxes on agriculture, facilitates the expansion of the urban sector (Krueger, Schiff, and Valdés 1988). The subsidization of manufacturing and construction activities also creates incentives for the expansion of the urban sector. Other policies that have a strong impact on the rate of conversion include subsidies to urban housing and public utilities, and policies that foster the creation of more and better-quality health and educational services in urban centers.

The cost of reconverting land from urban uses to agricultural activities is often prohibitively high. Where urban-based economic activities are promoted by highly distortionary incentive systems, governments should pay close attention to the conversion process and its long-run welfare consequences. Considering such distortions, direct intervention to inhibit land conversion could be welfare-enhancing. Without such intervention, the supply of land for agricultural activities might be suboptimal once discretionary policies are removed (see appendix B).

### **Regulatory reform in land markets**

Imperfections in land markets are common. Although a few are intrinsic to the nature of land itself, others are created by government interventions. In his analysis of the housing market in Bangkok, Dowall (1989) argued that an ample supply of land, strong competition among developers and builders, and an adequate supply of

finance are necessary conditions for the efficient operation of the land market, especially in the fast-growing cities of the developing countries.<sup>21</sup> The three conditions apply directly to urban land markets and, properly paraphrased, would apply as well to rural land markets. However, the reform of land markets cannot be based solely on such an ideal paradigm.

Consider, for example, the issue of finance capital. The purchase of land in meaningful quantities often requires large sums of cash. Since long-term capital markets either are incipient or do not exist in many developing countries, potential purchasers of land must either use their own capital or pay the cost, and bear the risk, of breaking down a long-term borrowing operation into several short-term ones. Because of capital market imperfections, this avenue is feasible only for financially strong individuals; as a consequence, access to land becomes limited. It is not clear that those entering the land market (rural or urban) under these conditions will be the most efficient entrants. But because borrowers know more about their projects than do lenders, it is unlikely that government intervention will be effective. The government tends not to possess any informational advantage over private agents. However, to the extent that financial markets are affected by distortions that are amenable to correction by government action (such as lack of collateral enforcement laws), the policy reforms in the financial sector should be undertaken simultaneously with land market reform in order to make the land reform more effective.

Thus, the reform process should not be confined to the identification and "correction" of market failures. Land market reform in most countries should be undertaken in two phases. First, policies that are currently impeding the market from performing its allocative function should be identified and phased out. The government then must provide the basis for the creation of a legal and institutional framework for land administration whose objectives are compatible with private incentives and that fosters competition. As with all government actions, implementation costs should be borne in mind as some policy changes, though in principle justified, are too costly relative to perceived benefits.

#### *Phase one: dismantling distortionary policies*

The removal of all restrictions on the sale and rental of land, including those related to minimum and maximum size, is essential to improve efficiency in land markets. Where the law does not allow the sale of public lands, or where government sales do not respond to market signals, the removal of such restrictions and revision of proce-

dures for sales will likely increase the effective supply of land and facilitate entry and exit in activities such as agriculture and housing production.

Rent controls (and prohibitions on sharecropping in agriculture) should be completely eliminated since, as discussed earlier, they reduce incentives for investments in the housing sector and may lead to reduced efficiency in agriculture because of risk and supervision requirements. Zoning should also be eliminated, with the possible exception of environmentally motivated restrictions. If society wants to restrict land use in specific areas, other instruments, such as creation of a market for development rights, may be more appropriate.<sup>22</sup>

Land and sectoral policies must be consistent. It is common for governments to implement sectoral and spatial policies that are incompatible with the overall objectives of land policy. Many zoning laws fail because the economic incentives embedded in other policies are not compatible with the restrictions imposed by the zoning legislation. For example, the concern of governments with the excessive conversion of land from rural to urban uses often leads to zoning restrictions. But these frequently coexist with policies, such as taxation (implicit and explicit) of agriculture, subsidies to urban housing and public utilities, and better access to health and educational services in urban centers, that often weaken the zoning laws.

#### *Phase two: institutional and legal reform*

Land administration in developing countries is often performed by institutions that have inadequate technical, administrative, and legal capacity. Bureaucratic procedures are cumbersome and not transparent. The costs of land adjudication are high, titles are often issued with incorrect boundary specifications, and crucial market information is not made available to interested parties at reasonable costs. In addition, the enforcement of property rights is not evenhanded and tends to discriminate against the poor. A high priority must be assigned to the implementation of institutional and legal reforms to eliminate these constraints for the operation of the land market.

*Land law.* One of the most important aspects of land market reform is the creation of a system of stable rules. A land law that establishes basic parameters for the operation of the market is a fundamental component of this system. It facilitates decisionmaking by economic agents (by reducing uncertainty), especially for investments with long gestation periods. In addition, the land law should provide easy and transparent access to the land

administration system and to dispute settlement institutions. Such access guarantees that incentives for rent seeking are minimized and prevents biases against the poor. The law also must take into account that various systems of property rights exist in practice. In Mexico, for instance, the law recognizes communal and private property and allows communities to join the private property system if they so wish. Finally, to ensure the equitable application of established principles and protect politically powerless groups, the law should not grant discretionary powers to members of the land management system.

*Institutional framework.* Another important component of a stable system of land administration is an adequate institutional framework capable of performing the following functions:<sup>23</sup>

- *Facilitate access to land information.* A land information system that is transparent and readily accessible is essential. It will normally be based on a cadastre and will register property, with corresponding data (for example, value and nature of attached investments), as well as boundary information and tenure status.<sup>24</sup> An accessible system provides an incentive for low-income landowners to keep an updated cadastre and to register titles.<sup>25</sup>
- *Adjudicate boundary disputes.* The system should be technically prepared to map the changes in boundaries following land market transactions quickly and at reasonable costs. At the beginning of the reform process, however, it is likely that clarification of existing boundary and title disputes will require most of the resources of the land administration institution.

One important instrument for the reduction of tenure insecurity is the possession of a registered title. As observed earlier, a number of studies show that the economic value of a title is large in both rural and urban areas. Titling is thus a critical component of urban and rural land market reform, and one that many developing countries are unprepared to handle in a timely way. The costs of titling projects can be high, and a careful cost-benefit analysis should be done in each particular case. In addition, it is important to note that rural land titling projects have met with bureaucratic and practical implementation difficulties. Appendix C contains a summary of the problems often found in World Bank projects. One of the most significant, noted by Wachter and English (1992), is the failure of projects to recognize that land titling usually entails a certain amount of land redistribution and political opposition from potential losers. Titling projects with only a technical dimension to formalize an existing situation are rare.

- *Resolve conflicts and enforce property rights.* The land administration system should be able to solve most conflicting claims in the field, which underscores the need for an in-house cadre of technical and legal knowledge. To resolve conflicts that cannot be handled in the field, an efficient appeals process through the judiciary system is of utmost importance. As a consequence of making information easily available, permitting most conflicts to be resolved in the field, and providing an efficient appeals process, the land administration system facilitates the enforcement of property rights in a nondiscretionary way.
- *Value and assess land.* The land administration system should be in charge of land valuation and assessment for purposes of the land tax, key for an efficient land tax administration. The information required for these functions—size, value, ownership status, productive capacity, and market value of outputs and inputs—is usually available in the cadastre.
- *Encourage registration.* One important function of the land administration system is to create procedures and rules that enhance, rather than reduce, the incentives to supply information and comply with registration requirements. Some countries discourage registration, for example, by only allowing registration of agricultural plots larger than a specified minimum size. As properties are subdivided through inheritance and sales, registration eventually ceases, and the cadastre and register become outdated.
- *Provide technical assistance.* The land administration system should provide technical assistance to local governments and communities. In countries where land taxation is managed at the local level, the system will have to provide cadastre information to local authorities for proper valuation and assessment. In countries where ethnic, religious, or other circumstances require special legal status for some communities, the land administration system should be prepared to provide such communities with technical assistance as well as help in conflict resolution.
- *Apply expropriation rules.* Clearly defined criteria for the expropriation of land for public projects must be established in the land legislation. The land administration unit will be responsible for executing the law and determining the compensation in each case according to land valuation studies performed for tax purposes and from cadastre information.

One of the questions in designing the land administration unit is whether it should be public or private. Certainly, several functions of this unit can be performed by the private sector. The land administration system

should be considered a technical unit and only a few of its activities—probably only those related to legislation, taxation, and conflict resolution—should be in the hands of the government. In most countries, however, the land administration unit is likely to begin operations as a unit within the government, as part of the reform process. Because the backlog of unresolved conflicts is usually large and their resolution is expensive, the private sector is unlikely to be attracted to this activity at first. This being the case, it is important to emphasize that all services provided by the government should be charged on a cost recovery basis. This approach is suggested by studies of Feder and others (1988) and Dowall and Leaf (1990), which showed that the private and social benefits created by these services are higher than their respective costs.

### **Suggestions for further research**

The need for additional research is great. One area in which more empirical analysis is needed is the assessment of the benefits of tenure security. Although the results of Feder and others (1988) and of Dowall and Leaf (1990) indicated that the social return of titling projects is positive, the African experience casts some doubt about the general applicability of this finding (see Place and Hazell 1993). Insofar as socioeconomic and cultural factors are fundamental for an understanding of the role of tenure security, empirical analysis in countries of diverse cultural backgrounds would provide firmer grounds for the evaluation of the economic returns of these reforms.

Another theme that deserves further research is land conversion, which has caught the attention of policymakers in several countries and given rise to interventions to halt or slow the process. Nevertheless, even acknowledging that the costs of land conversion are high and quite visible, it is not clear whether government intervention is called for. Observed rates of land conversion may in fact be too low rather than too high. From the point of view of efficient allocation of resources, the questions to be answered are as follows: Should the government intervene in the process, or should land allocation between rural and urban uses be determined entirely by market forces? Are the latter likely to generate an efficient allocation of land between rural and urban uses? In the presence of other distortions in the economy, can direct government intervention be welfare-enhancing? If so, what kinds of policies are likely to be most effective? Why have many of the policies in developing countries aimed at reducing the rate of land conversion been ineffective?

The economic cost of existing regulations must be made explicit so that decisionmaking is well informed. For that purpose a comprehensive modeling effort involving both the agricultural and nonagricultural sectors is called for. Factor markets and the cost elements that underlie the asymmetric nature of land conversion must be carefully considered, and spatial variables must be explicitly introduced in the general equilibrium framework.

The need for land market reform is urgent in the developing world. After the adjustment of macroeconomic and trade policies, factor market reform will be essential to enhance supply response. In particular, land market reform is essential for agricultural development and for the provision of affordable (but not subsidized) housing in urban areas. The task is daunting. The reforms will likely take a long time and require large amounts of resources. They are certain to face political resistance from groups that benefit from the existing system. The research suggested here will hopefully contribute to an improved design of these important reforms.

## **Appendix A**

### **Land price gradients in urban land markets**

The rate of land conversion has been very high in many cities in developing countries, and land market pressure has been greatest at the fringe areas of large cities. Dowall (1990b) noted that in Bangkok “the pace of urban land conversion from the mid-1970s to the mid-1980s was phenomenal, averaging about 21,250 rai per year.<sup>26</sup> But it increased even more during the 1984–88 period—more than doubling to 46,250 rai per year” (p. 5).

Conversion has also been extremely high in Jakarta. According to a World Bank (1990) study,

so intense is land pressure in Java that the need for house-lots alone is estimated to require the conversion of some 10,000 ha [hectares] of agricultural land per year. The Indonesia National Urban Development Strategy Project has calculated that Indonesian cities will also expand by 376,000 ha between 1980 and 1995, of which 222,500 ha would be in Java. Thus, Javanese cities are expected to expand by about 15,000 ha per year. Roads, industries and other uses are expected to increase total land conversion to 40,000 ha per year (p. 45).

Ingram and Carroll (1981), who studied land conversion in Latin American cities, showed that from 1950 to

TABLE 10.A1  
Land price gradients for selected cities

City	Year	Gradient
Bangkok	1988	-.0574
	1989	-.0558
	1990	-.0538
Bogotá	1973	-.1000
	1985	-.0200
Jakarta	1987	-.1813
	1988	-.1735
	1989	-.1690
Karachi	1980	-.0418
	1985	-.0577

Source: Data for Bangkok from Dowall (1990b); for Bogotá from Pachón and Hernandez (1989); for Jakarta from Dowall and Leaf (1990); and for Karachi from Dowall (1990a).

TABLE 10.A2  
Population density gradients for selected cities

City	Year		
	1950	1960	1970
Belo Horizonte	-0.26	-0.28	-0.27
Bogotá	..	-0.25	-0.12
Buenos Aires	-0.21	-0.14	-0.12
Cali	..	-0.41	-0.21
Guadalajara	-0.45	-0.46	-0.41
Mexico City	-0.37	-0.27	-0.17
Monterrey	-0.32	-0.27	-0.19
Recife	-0.25	-0.21	-0.19
Rio de Janeiro	-0.09	-0.08	-0.07
São Paulo	-0.14	-0.13	-0.12

.. Not available

Source: Ingram and Carroll (1981), table 5.

1970 the population density at the periphery of 10 cities<sup>27</sup> rose and, with the exception of Belo Horizonte, Brazil, the growth rate at the periphery was higher than that at the center. Dowall and Treffeisen (1990) reported that density in Bogotá increased faster for rings farther away from the center of the city. Although density in the center of Bogotá declined from 1973 to 1985, it increased outside the center, more so at the outer rings (p. 6). They cited evidence that real land prices in downtown Bogotá peaked in 1970 and have declined since then, whereas prices have increased at the fringe areas of Bogotá and beyond.

Land values and population density gradients can also be used to analyze pressure on the land market in the fringe of large cities. Land price gradients are often estimated based on modified versions of the equation

$$V(x) = V(0) e^{bx},$$

where  $V(x)$  is the land value at distance  $x$  from the central business district,  $V(0)$  is the land value at the central business district, and  $b$  is the estimated land value gradient.

Table 10.A1 displays the estimated land value gradients for selected cities. For three of the four cities, the land price gradient falls (in absolute value) over time.<sup>28</sup> This pattern indicates that the pressure over the land market increases as the distance from the city center increases. The fringe areas are those where the pressure is the highest.

In their study on Bogotá, Dowall and Treffeisen (1990) departed from the monocentric city model and instead used a multicentric model. Their gradient values are coefficients of the distance from the center of an individual neighborhood ("barrio") to the center of the proposed subcenter. Their values differ from those of Pachón and Hernandez (1989) but still show, for most subcenters, a decline over time.

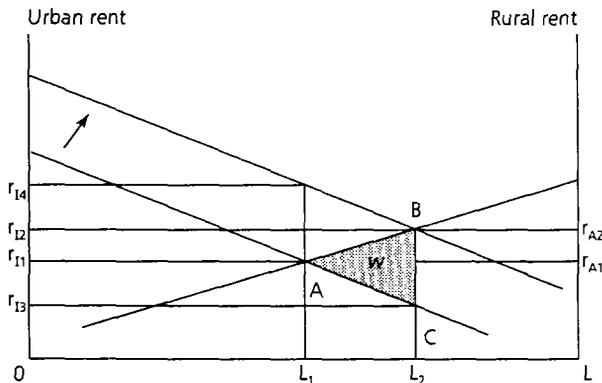
Population densities are positively correlated with land values. The results shown in table 10.A1 can be extended to a larger number of cities by analyzing the density gradients estimated (using an equation similar to that for the value gradients). Population density gradients for several Latin American cities are shown in table 10.A2, which for most cities reveals the same flattening of the gradient over time.

## Appendix B

### Welfare analysis of zoning restrictions

This appendix considers the welfare implications of a tariff to protect the industrial sector both with and without the simultaneous imposition of a restriction on rural-urban land conversion. Figure 10.B1 shows the value of the marginal product of land in the urban and rural sectors. The use of land in the urban sector is measured on the horizontal axis from left to right; the use of land in the rural sector is measured from right to left. The length of the horizontal axis is the total availability of land. Point A represents an equilibrium where there is no restriction on the use of land and no tariff is imposed. In that equilibrium,  $L_1$  units of land will be used by the urban sector,  $L - L_1$  units will be used by the rural sector, and the land rental will be  $r_{A1} = r_{I1}$ . When the tariff is imposed, the value of the marginal product shifts to the right, leading to a new equilibrium, B, where  $L_2 - L_1$  units of formerly rural land are shifted to the urban sector. The (market) rental value of land increases to  $r_{A2} = r_{I2}$ . However, the social value of land for the urban

**Figure 10.B1 Welfare implications of a zoning restriction**



Source: Bhadra and Brandão (1993).

sector is  $r_{13}$ . At each period, the social cost of this policy is represented by the area of triangle  $ABC$ , which we will call  $w$ .

If the policy remains in place for  $t$  periods and the interest rate is  $i$ , the total economic loss is given by the equation

$$TC = \sum_{n=1}^t \frac{w}{(1+i)^n} = \frac{w}{i} \left[ \frac{(1+i)^t - 1}{(1+i)^t} \right]$$

In period  $(t + 1)$ , when the tariff is removed, the equilibrium should return to point  $A$ . But because of the high costs of reconversion of land to the rural sector, it does not return to  $A$ . The social cost of the policy can be as large as  $w/i$ . If, however, the tariff is only a temporary instrument to transfer resources to the urban sector, and if the government simultaneously imposes a restriction on the conversion of land to the urban sector, this cost could be entirely avoided. In this case the urban rent increases to  $r_{14}$  and the agricultural rent remains at  $r_{A1}$ . This difference persists as long as both the tariff and the zoning restriction are in effect.

The above analysis rests on the assumption that the costs of enforcing the zoning restriction are negligible. As long as the costs of enforcement are less than  $w$ , such a policy will be desirable from a welfare point of view.

The analysis here remains fairly partial equilibrium in spirit, for only one market is considered in figure 10.B1. Although this simple representation is useful in illustrating the fundamentals of the problem, there are additional complications. For example, while land is moving from agriculture to industry in response to the tariff, it is likely that both labor and capital will also move in this same direction. However, when the tariff is removed, these two factors of production might return to agriculture more easily than land does. This clearly affects the cost calculations

in a fundamental way. The amount of capital and labor that returns to agriculture will depend on a number of factors related to the characteristics of the technology in the two sectors, other “push” and “pull” factors operating in agriculture and nonagriculture, and other policies that change the relative returns to these factors in each sector. A satisfactory answer thus depends on a careful accounting of all of the relevant elements, some of which may even counteract the initial impact of the tariff (for example, a subsidy for irrigation water). Moreover, tariffs affect demand structures and government budgets; the added revenue may generate further infrastructure construction and hence further shifts in the curves depicted in the figure. An applied general equilibrium model is required to take into account all the elements that influence this process and to assess their quantitative importance.

## Appendix C

### The World Bank’s experience with rural land titling

In their review of the World Bank’s experience with rural land titling, Wachter and English (1992) included 12 projects approved during 1971–81 and evaluated during 1982–91. Only one of these projects was a dedicated land titling project. Wachter and English concluded that almost all the projects suffered from one or more of the following problems:

- *Lack of political support.* Because land titling usually involves considerable redistribution of land rights (and of land itself), it is likely to suffer strong opposition from those who stand to lose in the process. Only under special circumstances will a land titling project consist of purely technical operations designed to formalize (or legalize) an existing situation.
- *Conflicting bureaucratic priorities or infighting.* In most of the projects Wachter and English reviewed, land titling was not the main component. Consequently, the agency responsible for the land titling component was not always properly involved in the project and often was not persuaded to change its own priorities to focus on the land titling component.
- *Lack of institutional capacity or unwillingness to commit adequate resources.* Wachter and English stressed that although land titling requires state participation, public administrations in several cases were not prepared to perform specific tasks. The major difficulties include deficient land records, lack of reliable maps, and low capacity of agencies to distribute titles.
- *Underestimation of the complexity or costs of the tasks to be carried out.* Substantial cost overruns occurred for

those projects for which cost data were available. This, coupled with qualitative statements in evaluation reports, led Wachter and English to conclude that the complexity of land titling projects was substantially underestimated.

### Notes

The views expressed here are those of the authors and should not be attributed to their respective organizations. The authors thank Claudio Frischtak, David Feeny, Eliseu Alves, José L. Carvalho, Robert Schneider, and Gervásio C. Rezende for their useful comments on drafts of this chapter.

1. The land market literature has been developed by urban and agricultural economists more or less independently. The agricultural land market literature was recently reviewed and extended by Binswanger, Deininger, and Feder (1995). The literature on urban land markets began with the work of Isard (1956), Alonso (1964), Muth (1961), Mills (1967), and others. It was expanded to take into account the economies of developing countries by Henderson (1982, 1988), Kelley and Williamson (1984), and Becker, Williamson, and Mills (1992). A third branch of the literature, which addresses issues associated with the process of rural-urban land conversion, is not as well developed as the other two, as Bhadra and Brandão (1993) argued in their recent survey.

2. "History has few examples of the uninterrupted transformation of general cultivation rights to land into owner-operated family farms. . . . Nearly always, there has been an intervening period under a class of rulers who extracted tribute, taxes or rent from cultivator families. . . . The landholdings of these overlords . . . were allocated temporarily or as permanent patrimony or ownership holding, along with the right to tribute, taxes, or rent (in cash, kind or *corvée* labor) from the peasants residing on the estate" (Binswanger, Deininger, and Feder 1995, p. 10).

3. It should of course be stressed that, as with other resources, the determination of land use by market forces does not take into account environmental and other externalities.

4. The spatial dimension of the interactions between the rural and urban sectors deserves more attention than it has received in the recent literature (see Bhadra and Brandão 1993).

5. The term "policy" is used here in a broad sense. The persistence and recurrence of certain situations reflect a policy, even if it is adopted by inaction of the public sector.

6. An interesting example is the concept of "reversion" in the Bolivian agrarian law of 1954, whereby land granted to farmers (except for small farms) can be reclaimed by the state if it remains idle for more than two years. In practice most reversion processes start with a request by a third party, but the law allows the state to take the initiative too. A special judicial body functioning under the Ministry of Agriculture is responsible for examining reversion requests, and the minister has the final decision. This judicial body responds directly to political interests and is clearly an additional

source of tenure insecurity for commercial farmers who have not yet fully developed their lands but intend to do so.

7. Expropriation rules frequently discriminate against the poor, who are not in a position to sustain legal battles with the public sector.

8. A formal model is presented by Feder and others (1988). Feder and Feeny (1991) offer a more general model.

9. In their careful analysis, Dowall and Leaf (1990) provide indirect evidence for Jakarta. They noted that "in the case of plots with low infrastructure availability, the impact of more secure tenure first declines, then rises with distance. This may reflect the fact that, in the unserved periphery of the city, the greatest conflict is occurring between formal developers and small land-owners who rarely have registered claims to their land" (p. 20).

10. Evidence for rural Africa, however, indicates that land rights did not significantly affect the use or modern inputs and yields of land-improving investments (see Place and Hazell 1993). The authors, who used the same conceptual framework as Feder and others (1988), explained the results as arising because capital and land markets are still undeveloped in these countries.

11. This section draws heavily on Bhadra and Brandão 1993.

12. These examples were for the most part adapted from Fisher 1982 and Murchison 1980.

13. See, for example, World Bank 1990 and United Nations reports (1986, 1987a, 1987b, 1988, and 1989). In Japan, for example, several restrictions to conversion exist. In most cases farmers who wish to transfer land to other uses must request permission from the governor of the prefecture or the Ministry of Agriculture, Forestry, and Fisheries. See Australian Bureau of Agricultural and Resource Economics 1988, p. 75.

14. Despite government intervention, large metropolitan areas and small and medium metropolises of developing countries continue to grow. The World Resources Institute (1988) estimated that urban population will grow from about 2.2 billion in 1990 to 5.0 billion in 2025 and account for about 90 percent of world population growth during this period. Crosson and Anderson (1992) estimated that to accommodate this population, about 125 million hectares of land—or approximately 10 percent of the potential crop land of developing countries—will have to be converted to urban uses.

15. Dowall (1990a, p. 22) showed that the forgone revenue was equal to about 317 percent of actual revenue in 1980 and 52 percent of that in 1985.

16. Hoff (1993) recalled that Henry George (1879) argued for a single revenue source for the government, which he identified as a tax on the rent of unimproved land.

17. Binswanger, Deininger, and Feder (1995) noted that a land tax based on the potential monetary yield of a certain plot under normal conditions has minimal disincentive effects, facilitates the taxation of the domestic agricultural sector while being much less regressive than poll taxes, and as long as the tax base is changed infrequently, it does not discourage investment in land improvement (p. 69).

18. The Brazilian land tax is a frequently mentioned example. As written in the law, it is a progressive tax, but a large number of provisions allow deductions that make enforcement difficult. The amount of tax collected has always been insignificant and, although systematic empirical evidence is not available, many economists in Brazil and elsewhere believe that in practice this land tax is regressive.

19. An extreme example may clarify the point. If land and currency were the only stores of wealth in an economy, the marginal tax on idle land would have to tax away the full nominal appreciation of the land price.

20. Empirical evidence confirms the impact of the credit subsidy on the price of land in Brazil (Brandão and Rezende 1992) and elsewhere (Shalit and Schmitz 1982). Econometric studies also show that inflation affects real agricultural land prices in the United States (Just and Miranowski 1988) and in Brazil (Brandão and Rezende 1992). Subsidized water in California has certainly affected agricultural land prices in that state, as it has in Colombia and northeastern Brazil. Casual evidence also indicates that urban land prices respond to inflation and to subsidies to housing.

21. "First, land markets require ample supply of land for residential development, and they must be free of bottlenecks and constraints which slow the delivery of residential lots to homebuilders or households. To meet this requirement, infrastructure—including roads, electricity, water, and sewage disposal—must be continuously made available. A second requirement for efficiency is competition. No specific land developer or housing builder should have sufficient market power to charge prices above what would prevail in an open and competitive marketplace. This implies that entry into the land and housing development industry must be fluid. Also no substantial barriers to entry which would hinder new firms from entering the marketplace should stand. A final requirement for efficient land markets is an ample supply of finance capital to support residential construction and to fund long-term mortgages for buyers. If these three conditions are met, there will be minimal land speculation, and housing prices will be held down to actual costs plus a reasonable profit for the developer" (pp. 1–2).

22. The creation of such a market is not a trivial question. Its discussion is beyond the scope of this paper.

23. This discussion benefits from the work of Barnes (1992).

24. In some countries the property registration system must remain (for constitutional or other reasons) in the judiciary system. In situations like this, a direct, preferably electronic, connection between the register and the cadastre system must be established to keep records up to date on both ends.

25. In several developing countries the total cost of title registration (including time, transportation, and sometimes subsistence outside the home town) is extremely high.

26. Since 6.25 rai is approximately equal to 1 hectare, this amounts to 3,400 hectares.

27. Belo Horizonte, Bogotá, Buenos Aires, Cali, Guadalajara, Mexico City, Monterrey, Recife, Rio de Janeiro, and São Paulo.

28. These regressions are not strictly comparable. In the case of Bangkok, for example, a dummy variable was introduced to control for the presence of services in the areas.

## References

- Alchian, A., and H. Demsetz. 1973. "The Property Rights Paradigm." *Journal of Economic History* 33: 16–27.
- Alonso, W. 1964. *Location and Land Use*. Cambridge, Mass.: Harvard University Press.
- Australian Bureau of Agricultural and Resource Economics. 1988. *Japanese Agricultural Policies: A Time of Change*. Canberra: Australian Government Publishing Service.
- Barnes, Grenville. 1992. "Technical and Institutional Issues Related to Land Tenure, Titling, and the Cadastre in Bolivia." Consultant's report. World Bank, Latin America and the Caribbean—Country Development III, Washington, D.C.
- Barrows, Richard, and Martha Newman. 1990. "A Review of Experience with Land Use Zoning." University of Wisconsin—Madison, Department of Agricultural Economics.
- Becker, Charles M., Jeffrey G. Williamson, and Edwin S. Mills. 1992. *Indian Urbanization and Economic Growth Since 1960*. Baltimore: Johns Hopkins University Press.
- Bhadra, Dipasis, and Antônio Salazar P. Brandão. 1993. *Urbanization, Agricultural Development, and Land Allocation*. World Bank Discussion Paper 201. Washington, D.C.
- Binswanger, H. P., and M. R. Rosenzweig. 1986. "Behavioral and Material Determinants of Production Relations in Agriculture." *Journal of Development Studies* 22: 503–39.
- Binswanger, Hans P., Klaus Deininger, and Gershon Feder. 1995. "Power, Distortions, Revolt, and Reform in Agricultural Land Relations." In Jere Behrman and T. N. Srinivasan, eds., *Handbook of Development Economics*, Chapter 42, vol. III. Amsterdam: Sevier Science B.V.
- Brandão, Antônio Salazar P., and G. C. de Rezende. 1992. "Credit Subsidies, Inflation, and the Land Market in Brazil: A Theoretical and Empirical Analysis." World Bank, Agricultural Policies Division, Washington, D.C.
- Crosson, Pierre, and Jock R. Anderson. 1992. *Resource and Global Food Prospects: Supply and Demand for Cereals to 2030*. World Bank Technical Paper 134. Washington, D.C.
- Dowall, D. E. 1989. "Bangkok: A Profile of an Efficiently Performing Housing Market." IURD Working Paper 493. University of California at Berkeley, Institute of Urban and Regional Development.
- . 1990a. "The Karachi Development Authority: Failing to Get the Price Right." IURD Working Paper 513. University of California at Berkeley, Institute of Urban and Regional Development.

- . 1990b. "A Second Look at the Bangkok Land and Housing Market." IURD Working Paper 527. University of California at Berkeley, Institute of Urban and Regional Development.
- Dowall, D. E., and M. Leaf. 1990. "The Price of Land for Housing in Jakarta: An Analysis of the Effects of Location, Urban Infrastructure, and Tenure on Residential Plot Prices." IURD Working Paper 519. University of California at Berkeley, Institute of Urban and Regional Development.
- Dowall, D. E., and P. Alan Treffeisen. 1990. "Spatial Transformation in Cities of the Developing World: Multinucleation and Land-Capital Substitution in Bogotá, Colombia." IURD Working Paper 525. University of California at Berkeley, Institute of Urban and Regional Development.
- Feder, Gershon, and David Feeny. 1991. "Land Tenure and Property Rights: Theory and Implications for Development Policy." *World Bank Economic Review* 5: 135-53.
- Feder, Gershon, T. Onchan, and T. Raparla. 1988. "Collateral, Guarantees, and Rural Credit in Developing Countries: Evidence from Asia." *Agricultural Economics* 2: 231-45.
- Feder, Gershon, T. Onchan, Y. Chalamwong, and C. Hongladarom. 1988. *Land Policies and Farm Productivity in Thailand*. Baltimore: Johns Hopkins University Press.
- Feeny, David. 1982. *The Political Economy of Productivity: Thai Agricultural Productivity, 1880-1975*. Vancouver: University of British Columbia Press.
- Fisher, P. S. 1982. "Introduction: Public Policy and the Urbanization of Farmland." *International Regional Science Review* 7: 249-56.
- George, Henry. 1879. *Progress and Poverty: An Enquiry into the Cause of Industrial Depression, and of Increase of Want with Increase of Wealth—The Remedy*. New York: H. George & Co.
- Heath, John R. 1992. "Evaluating the Impact of Mexico's Land Reform on Agricultural Productivity." *World Development* 20: 695-711.
- Henderson, J. V. 1982. "The Impact of Government Policies on Urban Concentration." *Journal of Urban Economics* 12: 280-303.
- . 1988. *Urban Development: Theory, Fact and Illusion*. New York: Oxford University Press.
- Hoff, Karla. 1993. "Land Taxes, Output Taxes, and Sharecropping: Was Henry George Right?" In Karla Hoff, Avishay Braverman, and Joseph E. Stiglitz, eds., *The Economics of Rural Organization: Theory, Practice, and Policy*. New York: Oxford University Press.
- Hueckel, Glenn Russell. 1972. "The Napoleonic Wars and Their Impact on Factor Returns and Output Growth in England, 1793-1815." Ph.D. diss., University of Wisconsin-Madison, Department of Economics.
- Ingram G., and A. Carroll. 1981. "The Spatial Structure of Latin American Cities." *Journal of Urban Economics* 9: 257-73.
- Isard, Walter. 1956. *Location and Space Economy*. New York: Technology Press, MIT, and John Wiley and Sons.
- Jimenez, Emmanuel. 1982. "The Value of Squatter Dwellings in Developing Countries." *Economic Development and Cultural Change* 30: 739-52.
- . 1984. "Tenure Security and Urban Squatting." *The Review of Economics and Statistics* 66: 556-67.
- Jones, Ronald W. 1965. "The Structure of Simple General Equilibrium Models." *Journal of Political Economy* 73: 557-72.
- Just, R. E., and J. A. Miranowski. 1988. "U.S. Land Prices: Trends and Determinants." In A. Maudner and Alberto Valdés, eds., *Agriculture and Governments in an Interdependent World*. Proceedings of the Twentieth International Conference of Agricultural Economists. Aldershot, England: Dartmouth Publishing Company.
- Katzman, Martin T. 1974. "The von Thunen Paradigm, the Industrial-Urban Hypothesis, and the Spatial Structure of Agriculture." *American Journal of Agricultural Economics* 56: 683-96.
- Kelley, Allen C., and Jeffrey G. Williamson. 1984. *What Drives Third World City Growth?* Princeton: Princeton University Press.
- Krueger, A. O., Maurice Schiff, and Alberto Valdés. 1988. "Agricultural Incentives in Developing Countries: Measuring the Effect of Sectoral and Economywide Policies." *World Bank Economic Review* 2: 255-72.
- Mills, E. S. 1967. "An Aggregative Model of Resource Allocation in a Metropolitan Area." *American Economic Review* 57: 197-210.
- Murchison, G. W. 1980. "Ways of Redistributing Benefits Created by Land-Use Policies." *Habitat International* 4 (4/5/6): 533-42.
- Muth, R. F. 1961. "Economic Change and Rural-Urban Land Conversion." *Econometrica* 29: 1-23.
- Otsuka, K., and Y. Hayami. 1988. "Theories of Share Tenancy: A Critical Survey." *Economic Development and Cultural Change* 37: 31-68.
- Pachón, Alvaro, and Sonia de Hernández. 1989. "La Vivienda en Colombia, 1973-1985: La Distribución Espacial de La Población en Las Áreas Metropolitanas." *Boletín de Estadística* (December): 245-58.
- Place, F., and P. Hazell. 1993. "Productivity Effects of Indigenous Tenure Systems in Sub-Saharan Africa." *American Journal of Agricultural Economics* 75: 10-19.
- Ricardo, David. 1949. *The Principles of Political Economy and Taxation*. London: J. M. Dent & Sons.
- Rondinelli, D. 1990. "Policies for Balanced Urban Development in Asia: Concepts and Reality." *Regional Development Dialogue* 11: 25-51.

# Incentive regulation: market-based pollution control for the real world?

Raymond S. Hartman and David Wheeler

Much of the environmental policy literature analyzes and contrasts the theoretical and institutional appropriateness of a variety of policy instruments. The most basic (and most common) policy instruments are effluent fees and effluent standards. Extensions of these basic instruments include tradable emissions permits, emissions banks, and technology standards.

Depending on the assumptions and predispositions that one brings to the analysis, one set of instruments usually is preferred. For example, most economists are proponents of price-based instruments, for the usual reasons. The preponderance of lawyers, however, favor quantity-based instruments, in the form of effluent or technology standards. Some economists have analyzed the conditions under which effluent prices or effluent standards are preferred,<sup>1</sup> but in all cases there has been scant attention directed to the incentive structures best suited for stimulating compliance with environmental regulations, *whatever form those regulations take*.

The importance of incentive structures, independent of the policies they are designed to effectuate, has received growing attention in the regulatory literature. This new approach takes account of the principal-agent problem: A regulatory body acts as the principal (for society) in delegating to or mandating certain actions of the regulated agents (for example, producers or consumers). The agents, however, have their own agendas, and probably will not fully respond to the principal's regulatory mandates in the fashion desired. In such situations the principal must use both normative economic theory to identify the behavior desired of the regulated individuals, and positive behavioral theory to identify the incentive structures necessary to stimulate compliance by the agent. The appropriate policy instruments and incentive structures must account for the motives of the agents and the nature of the information held by each party. The principal attempts to design a mechanism that will induce the agent to come as close as possible to maximizing the principal's performance criteria.<sup>2</sup>

Designers of environmental regulations face these same principal-agent problems. For example, effluent fees designed by the principal (an environmental protection agency) may be Pareto superior in theory. However, if economic agents subvert the fees by submitting falsified information or avoiding monitoring, the results will not be Pareto superior. Likewise, effluent standards designed by the principal may have certain desirable properties. However, if economic agents avoid or subvert the standards, compliance will be incomplete. We believe, therefore, that compliance incentives for environmental policy must be analyzed explicitly and independently. Fortunately, we can draw on a rich literature that analyzes incentive regulation and principal-agent issues.

The literature on incentive regulation focuses on improving the structure of incentive mechanisms in regulated industries by explicitly accounting for principal-agent issues. Incentive regulation mechanisms have been used to rationalize prices, improve the economic efficiency of long-term strategic and short-term operational plans, and encourage the attainment of specific policy goals. The incentive regulation literature provides an important point of departure because many incentive mechanisms have already been reviewed or implemented in the United States (at the U.S. Federal Energy Regulatory Commission, the Federal Communications Commission, and state public utility commissions), in Britain, and in other member countries of the Organization for Economic Cooperation and Development (OECD).

The major insights of the incentive literature for environmental policy modeling are threefold. First, simple price-based and quantity-based regulatory directives are often naive. Examples include directives to "price at social marginal cost" or "offer energy conservation programs until their marginal cost equals their marginal social value." In most cases the regulator has insufficient information to evaluate whether the regulated producers are implement-

- Schultz, T. W. 1953. *The Economic Organization of Agriculture*. New York: McGraw Hill.
- Shalit, H., and A. Schmitz. 1982. "Farmland Accumulation and Prices." *American Journal of Agricultural Economics* 64: 710–19.
- United Nations. 1986. "Global Report on Human Settlements." UN Centre for Human Settlements, New York.
- . 1987a. "The Prospects of World Urbanization." *Population Studies* 101. Department of International Economic and Social Affairs, New York.
- . 1987b. "Population Growth and Policies in Mega-Cities: Dhaka." Population Policy Paper 8. Department of International Economic and Social Affairs, New York.
- . 1988. *World Demographic Estimates and Projections, 1950–2025*. New York.
- . 1989. "Population Growth and Policies in Mega-Cities: Jakarta" Population Policy Paper 18. Department of International Economic and Social Affairs, New York.
- von Thunen, J. H. 1966. *The Isolated State*. 1826. Reprint. London: Pergamon Press.
- Wachter, Daniel, and John English. 1992. "The World Bank's Experience with Rural Land Titling." Divisional Working Paper 1992–35. World Bank, Environment Department, Washington, D.C.
- World Bank. 1990. *Indonesia: Sustainable Development of Forests, Land, and Water*. A World Bank Country Study. Washington, D.C.
- World Resources Institute. 1988. *World Resources 1988–89*. New York: Basic Books.

ing such directives. Because the regulated producers (the agents) are usually governed by different objectives than those of the regulators, the agents will maximize their objectives subject to regulatory constraints.

The second conclusion is that the principal must develop incentive mechanisms that make policy compliance economically beneficial to the regulated parties. In order to do so, the incentive mechanisms should:

- Recognize that the regulating principal has less information than the agents being regulated, who in turn are interested in profit, not social welfare or some other social goal.
- Make it beneficial, other things being equal, for the regulated parties to reveal information regarding product demand, supply costs, and the cost of compliance.
- Reward the regulated parties for both effort and outcomes.
- Involve some method of sharing the risks and the rewards of compliance.

The third conclusion is that incentive mechanisms do indeed work. As will be discussed below, incentive regulation mechanisms have proved effective in stimulating, or “incentivizing,” regulated behavior that was not stimulated by more traditional regulatory instruments that disregarded incentives.

All the arguments in favor of incentive regulation in industrial countries are even more persuasive for developing countries. In developing countries there are frequently greater asymmetries of power and information and stronger tendencies toward status quo biases.<sup>3</sup> In such cases it is particularly naive to think that the environmental principal can merely announce a set of fees or standards and expect that the regulated agents will comply. The agents will find noncompliance and avoidance of monitoring economically beneficial and quite easy.

In developing countries it would therefore be preferable to “incentivize” the regulated agents to comply with Pareto superior behavioral rules.<sup>4</sup> The principal should make it economically beneficial for the regulated parties to reveal information, cooperate with monitoring, and comply with regulatory directives. It should be stressed here (as we will note more completely below) that we are not advocating the use of subsidies for compliance.<sup>5</sup> Rather, we will demonstrate that the incentives proposed for compliance are analogous to the fees paid for tradable emissions permits within an emissions bank. Indeed, we contend that our proposed incentive regulatory system can provide the transition to a tradable permit or emissions bank system.

The discussion in this chapter proceeds as follows. The first section discusses traditional methods of regula-

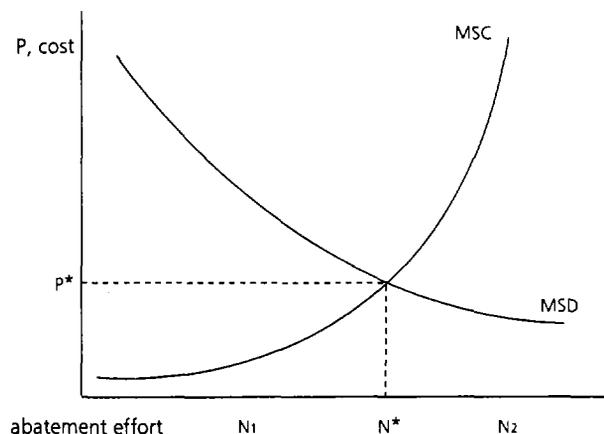
tion, both in principle and in terms of their application internationally, and reviews price-based and quantity-based methods. The second section discusses incentive regulation and how it has been used to overcome principal-agent problems. Several versions of incentive regulation mechanisms are described, and a variety of situations in which incentive mechanisms have been employed successfully are identified. The third section draws from the preceding sections and suggests a paradigm for setting better incentives for environmental policy compliance.

**Traditional methods of regulation**

Figure 11.1 depicts, in an idealized fashion, the major issues faced by an environmental protection agency when implementing environmental policy: the value of, or demand for, pollution abatement by society and the cost of that abatement to society. The demand for abatement is summarized by a marginal social demand (*MSD*) schedule, which summarizes society’s valuation of or willingness to pay for pollution abatement. *MSD* declines with abatement effort, which is measured (as *N*) on the horizontal axis. Hence, society is willing to pay considerably more for the first marginal unit of abatement (when pollution is most serious) than it is willing to pay for a marginal unit of abatement when a significant abatement effort has already been undertaken (when *N* is large).

Pollution can be abated by scaling back polluting activities or by diverting resources to cleanup. In either case there will be a cost to society. Diminishing returns will apply; more resources will have to be devoted to cleaning up each additional unit of pollutant. This escalation is traced by the marginal social cleanup cost (*MSC*) schedule in figure 11.1.

**Figure 11.1 Social determinants of optimal pollution abatement efforts**



If the government knows the location of both  $MSD$  and  $MSC$ , it can readily identify the socially optimal level of pollution in figure 11.1. Suppose the control region before regulation is characterized by pollution level  $N_1$ , where  $MSD$  is significantly above  $MSC$ . In this case society is willing to pay considerably more to reduce pollution by one unit than the social cost of that abatement. Logically, cleanup is worthwhile. At  $N_2$ , by contrast, cleaning up a unit of pollutant will cost far more than it is worth to society. The optimum is at  $N^*$ , where  $MSD = MSC = P^*$ ; the gain to society from abating one unit of pollutant exactly matches the social cost.

*Price-based or quantity-based regulation?*

In the stylized world of figure 11.1, two basic ways are available to an environmental protection agency to move industry to the optimum pollution point. It can choose a price-based approach, insisting that all firms pay the shadow price  $P^*$  per unit of pollution, thereby ensuring optimal environmental use at  $N^*$ . Or it can opt for a quantity-based solution, ordering firms to cut back pollution in such a way that aggregate environmental use is restricted to  $N^*$  (at opportunity cost  $P^*$ ).

The most common price-based instrument is the effluent charge, which is levied per unit of pollutant discharged into the relevant environmental medium (air, water, or land). Where monitoring discharges is difficult, governments sometimes opt for deposit-refund systems, which require firms to pay in advance for estimated pollution and provide rebates on proof of lower emissions. (In the discussion that follows, effluent charge refers to both charge systems.) Tradable permit systems specify an aggregate limit on pollution, but allow it to be allocated among firms in a secondary market for pollution rights.<sup>6</sup>

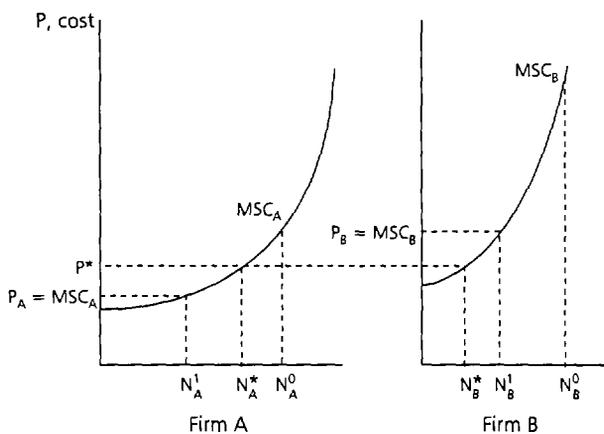
Quantity-based regulation may operate through baseline emissions standards or mandated installation of pollution control equipment when discharge monitoring is difficult.

It is easy to show that effluent standards can be less economically efficient than an effluent charge or tradable permit system. Suppose, as in figure 11.2, that the government dictates pollution reduction for two representative firms from different industries (A and B), ordering both to cut back the quantity of their pollution by 50 percent from  $N_A^0$  and  $N_B^0$ , respectively. The regulator may implement these requirements through either technology or effluent standards. The marginal social cleanup cost curves for A and B have different slopes and positions, reflecting the underlying differences in their production economics. Notice that firm A is the more efficient abater, since its abatement cost curve is uniformly lower than that of firm B. As a result, at the new constraints imposed by the government ( $N_A^1$  and  $N_B^1$ ), firm A and firm B confront very different costs of abatement.

Firm B now faces a much greater marginal abatement cost than firm A ( $P_B = MSC_B > P_A = MSC_A$ ). Since the analysis presumes perfect competition in all other markets, this implies that society as a whole would be better off if the marginal unit of abatement could be transferred from B to A, since firm A is the more efficient abater. Indeed, society can abate more efficiently by transferring abatement responsibility from B to A as long as  $P_B = MSC_B > P_A = MSC_A$ . By implication, the government cannot do better than specifying a common emissions charge  $P^*$  or, equivalently, setting  $N^*$  and allowing firms to trade freely in rights to pollute until equilibrium is attained at  $P^*$ . It would have to be uncommonly lucky to specify an emissions standard rule that matched the allocation of pollution cutbacks effected automatically by the effluent charge or tradable permit approaches.

Numerous simulation studies in the United States have supported this theoretical result, suggesting very high costs for actual emissions standard-based pollution regulation when compared with equivalent pollution reduction in a hypothetical effluent charge system. Air pollution studies have considered sulfur dioxide (Roach and others 1981; Spofford 1984), nitrogen dioxide (Krupnick 1983; Seskin, Anderson, and Reid 1983), and suspended particulate (Atkinson and Lewis 1974). The estimated emissions standards-effluent charge cost ratios for meeting the same pollution target have a median value of 4.2. For water pollution, simulation studies of biological oxygen demand (Eheart, Brill, and Lyon 1983; Johnson 1967) have a median cost ratio of 1.6 (under the

**Figure 11.2 Interfirm and interindustry differences in abatement**



assumption that the quantity-based approach is proportional reduction from existing effluent levels).

#### *Sources of variation in instrument choice*

Despite these impressive efficiency results, price-based approaches have rarely been implemented. Even where price-based systems are in effect, quantitative limits such as baseline effluent standards are always employed. Policymakers also have frequently resorted to another quantity-based approach—mandated installation of specific pollution abatement control equipment.

The reason for this revealed preference for effluent standards and pollution abatement control equipment is verification and monitoring. Pollution abatement control equipment is preferred in many cases because of the expectation that verification of installation can substitute for continual inspection. Although there are obvious reasons why this expectation might not hold—equipment can be badly maintained or simply disconnected—such requirements have featured prominently in industrial regulation almost everywhere.<sup>7</sup>

What should determine the selection of policy instruments? The following are important in deciding whether a price-based or a quantity-based system has a significant advantage in a specific control region.

*Heterogeneity of industrial activity.* Other things being equal, quantity-based regimes are suboptimal to the extent that industrial diversity gives firms very different cleanup cost curves (as in figure 11.2, where  $P_B = MSC_B > P_A = MSC_A$ ). If these curves were nearly identical, there would be few potential “gains from trade,” and government simulation of a market through a price-based regime would result in very little cost savings.

In practice the potential gains appear to be large, because most pollutants are emitted by plants in diverse industrial sectors with very different cleanup costs. For one example, Crandall (1983) contrasted an airborne particulate cleanup cost range of \$36 to \$680 per ton removed in the utility sector with \$1,010 to \$3,030 in a secondary aluminum plant and \$30,880 in a coke oven. For a second example, Hartman, Wheeler, and Singh (1994) found that air pollution abatement costs vary across U.S. manufacturing sectors by a factor of 100.

*Nature of uncertainty about environmental damage and abatement cost curves.* Another significant problem is the frequent absence of any good information about the shape or position of the marginal social demand (MSD) and marginal social cleanup cost (MSC) curves. Theoretical work by Weitzman (1974) considered the relative merits of quantity-based and price-based instru-

ments under uncertainty when there are alternative prior beliefs about the slopes of these schedules. Adar and Griffin (1976) extended this analysis to the case of environmental regulation.

At issue is the cost of being wrong, on the assumption that policies are hard to change quickly once implemented. A price-based system will be superior if the MSD curve is relatively flat (society is relatively indifferent to small changes in pollution and pollution abatement) and the MSC curve is relatively steep (abatement is very expensive for firms). In such a case small reductions in pollution may radically increase industry's cost without much benefit for society as a whole. Price-based policies are better in this case because they ensure that the abatement cost for industry will not be inordinately high.

The converse is true when the pattern is reversed: steep MSD curve, relatively flat MSC curve. In such a case a small deviation of the actual impact of an effluent charge from its expected impact can cause a large change in pollution and society's valuation of social damage and abatement. Conversely, a quantitative restriction (or a tradable permit system) can offer considerable certainty about marginal social damage and the value of abatement but will not change industry's cost much even if it is somewhat overdone.<sup>8</sup>

*Pollutant toxicity.* When a plant generates toxic emissions, the regulatory agency has to impose quantity controls. While some highly poisonous substances are easy to identify (for example, plutonium), the identification of most other pollutants as sufficiently toxic to warrant quantity-based controls involves somewhat arbitrary judgments. The number of such substances is growing rapidly, in part because of widespread worry that stock pollutants may haunt future generations. It also owes something to uncoordinated environmental regulation in the OECD countries.<sup>9</sup>

*Employment impacts.* Models of optimal pollution in a control region generally assume that social and private marginal costs differ only in the use of environmental services. The conventional MSC curve does not include any provision for the social costs of adjustment in other input markets. If a heavily polluting firm is also a large employer, however, these costs may include mass layoffs and serious community disruption. Findley (1988) cites an illustrative, albeit extreme, case for Brazil. A large cement plant in Contagem, in the metropolitan region of Belo Horizonte, was ordered closed in 1975 by the municipal authority because of failure to install stack filters to control hazardous emissions. Plant employees resisted bitterly, waging a pitched battle with government

troops called in to enforce the order. After closure, the plant owner appealed to Brasilia for relief and got a presidential restraining order that overrode the local decree. The ensuing negotiations resulted in a postponement of compliance in exchange for a company agreement to install the required filters.

*Significant number of public enterprises with soft budget constraints.* If public enterprises can draw on public funds to cover deficits, they will have little incentive to reduce pollution when price-based regulation is applied. For exactly the same reason, they will have little reason to resist quantity- or technology-based regulation. When the regulations call for closure (which they sometimes do), access to public funds will not matter.

*Information held by the regulated firms.* If the regulated firms are fully informed about the costs and technologies of abatement in their industry and in other industries, price-based regulation will be more appropriate. Conversely, if the regulated firms are not well informed, quantity-based or technology-based regulation will be more appropriate.

To summarize, the advantage of price-based regulation over quantity-based approaches will depend on the degree to which:

- All markets in the system are competitive.
- The control region is populated by many privately owned, cost-minimizing, fully informed firms engaged in heterogeneous activities with low transaction costs.
- The government is fully informed about industry's environmental demand curve and the control region's environmental supply curve (MSD and MSC in figure 11.1).
- The pollutants subject to regulation are not highly toxic (or are well dispersed) at current levels of emission.

*Comparing price-based approaches: the merits of effluent charges and tradable permits*

In the United States environmental economists have advocated tradable permit systems because they combine some of the advantages of quantity-based and price-based instruments. Since tradable permit systems mandate aggregate reduction of emissions, some believe they have more political appeal than emissions charges.

There also is a presumption in the literature that a tradable permit system may be better than effluent charges because the first has lower regulatory costs. The most commonly cited operational critique of effluent charges focuses on the problem of finding the appropriate charge when the long-run social demand curve is not known. In the idealized effluent charges system, some fee is assessed

at the outset, pollutant reduction is tracked, and the fee is adjusted to compensate for overshooting or undershooting a pollution target that is presumed to be known. But implementation could clearly be a problem. The idealized experiment is conducted holding constant other variables (that obviously do not remain constant), and the tracking period is of indeterminate length, making it difficult to measure the independent effect of the effluent charge. Furthermore, because the charge is assessed per unit of effluent, good assessment requires accurate, continuous monitoring of effluents from all relevant sources. The cost of a comprehensive system might well outweigh the benefits of undertaking the program in the first place.<sup>10</sup> Finally, effluent charges systems clearly involve considerable "fine tuning," requiring variation in the levy over long periods of time. Implementation would call for frequent, large, and seemingly arbitrary shifts in charge schedules whose impact on profitability would have to be significant if it were to be effective.

Is a tradable permit system superior? It is the dual solution to the optimization problem, and duality implies rough symmetry. In fact, the two major weaknesses of effluent charges—high monitoring costs and iterative adjustment—afflict tradable permit systems as well. Monitoring requirements are basically the same: Effluent volumes must be measured for taxation under effluent charges systems and for compliance with permit limits under tradable permit systems. The two systems have dual iterative solutions: Under an effluent charges system, the charge must be adjusted as the pollution consequences are revealed; under a tradable permit system, the overall pollution limit may be adjusted as its true social opportunity cost (rising unemployment, falling output and profits, and so on) is revealed.<sup>11</sup> However, a tradable permit system probably requires fewer adjustments (in aggregate quantities) than would be required for an effluent charges system.<sup>12</sup> Under both systems, there is convergence to some politically tolerable pairing of pollution limit and shadow price.

In short, on operational grounds, a tradable permit system requires fewer iterations to achieve total emissions targets. However, the conventional argument for the superiority of tradable permits over effluent charges is not persuasive on monitoring grounds. Both systems will be partially monitored at best; tradable permit systems have a particularly difficult monitoring problem because the regulatory system may not be capable of adjusting quickly when permits are traded. Firms may then be able to pollute illegally for long periods, purchasing traded permits only when they are put on notice of inspection.<sup>13</sup> A trad-

able permit system remains clearly better than an effluent charges system in cases where there is considerable uncertainty about the pollution abatement demand and cost curves but a presumption that the first is steeper.<sup>14</sup>

#### *Practice in the OECD economies*

Recent studies comparing pollution control policies and results in the member countries of the OECD are close to unanimity in their assessments of instrument choice.<sup>15</sup> Quantitative controls have been dominant in the formal regulatory systems of Canada, Japan, the United States, and the member countries of the European Union (EU). While price-based incentives have been used in some situations, there seems to be no situation in which they have been used to the exclusion of quantity controls.

High baseline emission standards have always been maintained, and regulatory strictness generally seems to reflect relative population density. By one measure—reported cost of pollution abatement and control as a proportion of total investment—Japan has been the strictest, followed in order by Western Europe, the United States, and Canada (USCBO 1985). Both quantitative emission controls and mandated “best practice” technology requirements are commonplace everywhere but the United Kingdom, which has developed a unique system of informal adjustment through negotiations between the pollution inspectorate and its industrial partners (Vogel 1986).

*EU systems.* Unlike the United States, which has relied almost exclusively on quantitative controls, Japan and Western Europe have made some use of effluent charges as a supplementary tool for environmental management. These are all “polluter pays twice” systems: Although charges are levied, effluent standards are maintained as well. France has levied charges for both air and water pollution since 1985. Since 1981 Germany has subjected all open water discharges to a fee based on the estimated amounts of different pollutants in the waste stream. In Sweden plants that discharge into municipal facilities (including solid waste plants) must pay the full cost of collection and treatment. The Swedish government has attempted to levy an additional “environmental charge” to neutralize profits accruing to noncompliance with standards, but problems in establishing legally acceptable verification of noncompliance levels seem to have limited the effectiveness of the system.

In almost all EU cases, such charges are set too low to have much incentive impact. Only in the Netherlands have effluent charges apparently exceeded the marginal cost of abatement, and significant effluent reductions

have resulted.<sup>16</sup> That country’s uniquely high charge seems to have been politically feasible only because the historical necessity of hydraulic control has made strict systems acceptable to Dutch industry. Most European effluent charges systems are designed basically to raise revenues for financing the regulatory system, R&D, and subsidies for industry purchases of abatement equipment.

Japan has long maintained a pollution charge system for sulfur dioxide emissions. A 1973 law designates areas where sulfur dioxide concentrations are thought to be high enough to be significantly damaging to health. Nationally, charges are levied on stationary sulfur dioxide emissions sources in the ratio of 9:1 between designated and nondesignated areas. The funds are used for compensating recognized groups of pollution victims (O’Connor 1993).

*Tradable permit systems.* Emissions trading has been a U.S. innovation for the most part (OECD 1989; USCBO 1985). True tradable permit systems have been rare even in the United States, where the system evolved out of quantity-based policies that have focused on controlling point emission sources within plants. This evolution began in the 1970s, when very restrictive and inflexible quantity-based policies threatened to strangle economic growth in highly polluted areas. The regulatory statutes were gradually revised to allow for some compensating emissions adjustments. Typically, these have allowed for changing the pattern of point emissions within a control region (known as a “bubble”), provided the overall results stay within the regulatory limits. In principle, interfirm offsets (“offset trading”) were made allowable. In practice, however, almost all offsets take place across emissions points within individual plant sites because the transactions costs of interfirm exchanges have been very high (Hahn 1989).

Emissions offset trading has apparently resulted in huge cost savings in some cases. It has allowed economic growth to continue in highly polluted areas where inflexible point source policies would undoubtedly have stifled it. Most analysts agree with Hahn (1989), however, that emissions offsetting has not significantly reduced total pollution.

If interfirm offset arrangements have been rare, formal markets in tradable permits to pollute have been even rarer. Hahn (1989) and Tietenberg (1988) cite only a handful of cases that predate the U.S. experiment with sulfur dioxide permits incorporated into the 1990 Clean Air Act Amendments. One past notable success has been the U.S. lead permit trading program for gasoline refiner-

ies, while one notable failure has been an attempt to establish a tradable permit market for one region of the Fox River in Wisconsin.

The U.S. Environmental Protection Agency's lead trading program was initiated in 1982 and scheduled from the outset for termination at the end of 1987. Its purpose was to promote flexibility and cost savings in the refinery industry's response to mandated reductions in the lead content of gasoline. Superior performers were allowed lead credits, which could be traded to inferior performers in an open market. Since the aggregate targets were fixed by law and easily monitored, the purpose of this program was explicitly *cost-effective* compliance with the regulations, not extra reduction in pollution. Hahn (1989) estimates that total yearly savings to refiners under the program must have been somewhat higher than \$250 million (in 1994 dollars), when compared with an emissions standards regime forcing concurrent reduction at every refinery.

Hahn (1989) cites both geographic limitation and imperfect competition as reasons for the failure of the Fox River program. The program was originally designed to allow for the trading of permits to discharge a variety of organic pollutants into the Fox River at two specific sites. Initial simulations had estimated the yearly program savings over emissions standards to be \$7 million. However, the program generated only one permit trade in its first six years of existence. Neither of the two major discharging industries—pulp and paper and municipal waste treatment—was competitive in structure. Furthermore, the two points on the river where pollution tended to peak were treated as separate control regions, with no trading between regions allowed. This restriction limited the potential market to six or seven firms per region, a number that seems to have been too small to support permit trading.

Interest in tradable permit systems is clearly growing. In the United States, the 1990 amendments to the Clean Air Act established a market in permits for sulfur dioxide emissions, which are targeted for an approximate 50 percent reduction by 2000 (Coy and Sieminski 1990; Weisskopf 1990). These amendments may reflect a trend toward increased acceptance of the concept among policymakers, but they may also reflect the relative certainty that comes with experience. The same amendments imposed direct controls on 189 other substances, most of which had not previously been regulated.

In the European Union, Germany provides one example of transferable pollution rights. Under the Plant Renewal Clause licenses for new plants are refused in

areas that have not attained mandated ambient standards, unless they replace plants of the same kind with substantially lower emissions. This example is similar to the offset arrangement in the United States and is far from a true tradable permit system.

*Effluent charges versus tradable permits: the balance to date.* Although economists have long regarded tradable permits as a realistic alternative to effluent charges, policymakers seem to disagree. Charges are much more common than permits, although they generally bear little resemblance to idealized instruments. All recent surveys of policies in the OECD countries (Hahn 1989; Kopp, Portney, and DeWitt 1990; OECD 1989) report that governments have almost never raised effluent fees to a point anywhere near the marginal cost of abatement. They simply use modest fees as a source of revenue for public abatement investments and as subsidies to firms for abatement projects. A striking exception seems to be the Netherlands case cited earlier, but Bressers (1983) found that the Dutch government also intended water charges only to be a source of funding for public abatement facilities. Because the Netherlands is densely settled, the required facilities were very expensive. The resulting fees were high enough to induce serious private abatement, but this reduction in pollution was an unintended benefit.

Environmental economists generally prefer effluent charges because they provide appropriate incentives for pollution reduction, not because they raise money. The revenue is considered a windfall gain for the treasury and should ideally be allocated among public expenditures in proportion to social welfare weights. But governments seem to use effluent charges only to raise money, and they use the revenues only to promote pollution abatement.

Several practical reasons for this behavior can be cited. Governments that go beyond modest effluent fees may induce small-firm demands for costly monitoring systems. In addition governments do require a steady source of funds for investments in public pollution treatment facilities. Anderson (1990) argues that targeted effluent charges serve this need very well. The winds of political support for any program shift over time, and drawing pollution expenditures entirely from the general revenue fund makes them vulnerable to the short-run exigencies of the political process. Continuity is ensured if pollution abatement investments are linked to revenues from effluent charges. The industries that are taxed also are more likely to cooperate if they perceive a direct linkage between their expenses and environmental cleanup. Finally, the effective planning and implementation of regulation itself depends on a group of experts whose con-

tinued financial support can be ensured by effluent charge revenues.

In short, modest effluent charges systems mixed with emissions standards have prospered because they finance regulatory operations and provide at least some incentive for cleanup in the long run. Operationally, effluent charges have also been better than tradable permits for control regions with too few firms to support a market in permits. If firms in the control region have significant market power, additional complications arise.<sup>17</sup> Tradable permits can be viewed as property rights; polluting firms cannot operate without them. Dominant firms therefore have a natural incentive to take preemptive positions, particularly when there is substantial uncertainty about the future market value of the permits. They may also be able to bargain strategically for lower permit prices.

*Summary of OECD experience.* Most OECD countries have achieved notable reductions in some forms of pollution, using broadly similar approaches to regulation. There has been strong reliance on quantity-based systems with end-of-pipe controls. Apparently, the combined impacts of political preferences, uncertainty, and concern about toxic substances have outweighed the abstract efficiency argument for price-based instruments in most cases. There is scattered use of charges, which are generally below abatement costs and intended primarily for fundraising. Control region heterogeneity also seems to create many local problems that cannot be handled by central regulators. As a result, negotiations between local agents and firms seem to play an extremely important role throughout the OECD.

#### *Practice in the non-OECD economies*

*Poland: a former socialist economy.* Polish regulatory policy goes quite far toward institutionalizing appropriate shadow pricing principles for environmental services. Effluent charge structures exist for both water and air pollution; fees are explicitly regarded as charges for use of environmental services, thereby institutionalizing a concept that is much advocated by Western economists but almost never implemented. The fees have a detailed structure that is not replicated in any market economy. Polish policy for air pollution, for example, has specific charge rates for 54 separate pollutants (Wilczynski 1990). By world standards, Poland is also well endowed with technical and scientific manpower that could be mobilized for staffing an effective environmental protection agency. Nevertheless, Poland is an environmental disaster area.

Much of the country's pollution problem can be traced, directly or indirectly, to more general problems of distortionary incentive regimes. However, as Hughes

(1990) notes, several institutional and economic factors contributed heavily to the failure of direct pollution control. First, as in the OECD countries, much of the actual regulatory implementation was left in the hands of local governments. Many Polish communities have been dependent on a few large enterprises that fund local communal services, which has drastically reduced regulators' leverage in negotiations with plant managers. Second, the "polluter pays" principle could not operate effectively in a system where most major enterprises operated with soft budget constraints and many goods and services were not allocated in markets.

Thus the elaborate structure of Poland's environmental regulation bore no resemblance to the actual pollution costs faced by enterprises. The overwhelming dominance of state industry and the primacy of other objectives ensured very poor support for environmental regulatory agencies, lax monitoring, and general failure to enforce existing regulations. Many plants were essentially unmonitored; both fines and effluent charges were far below marginal abatement costs in all but a few isolated cases; and collection was sporadic (Wilczynski 1990).

*Brazil and Korea: Two newly industrialized countries.* Pollution control in both Brazil and the Republic of Korea has had a strong quantity-based orientation.<sup>18</sup> As in the OECD countries, information problems stemming from control region heterogeneity have led to the delegation of considerable discretionary power to local regulatory agents. Brazilian regulators, like their Korean counterparts, have at their disposal a graduated system of notifications, fines, suspensions, and shutdowns for standards violations. However, the Brazilian system of implementation seems to be "softer" than the Korean system, in the sense that Brazilian policy has seldom depended on explicit emissions standards or mandated installation of abatement equipment. Rather, negotiations have generally been handled in the British style, on a case-by-case basis. This approach has had the advantage of permitting greater flexibility in adaptation to particular circumstances, although it is easily abused in practice.

In Korea there has been substantially more reliance on mandated controls, with the threat of highly punitive sanctions for violations. In fact, the Korean system incorporates some elements of a basic emissions charge system. Since 1986 Korea's Environmental Administration has been required to levy charges on plants that are in violation of their emissions standards. However, the charge rate is not linked to the level of excess emissions and has sometimes been set lower than the operating costs of pollution abatement facilities.

Both support for enforcement and the incidence of meaningful sanctions have increased markedly in both Brazil and Korea, as democratically elected governments have come under great popular pressure to improve the environment, even at substantial cost. Although extremely spotty, the available evidence suggests that both countries have attained some improvement in environmental quality. Major investments in pollution abatement equipment by some of the largest polluters in the São Paulo region of Brazil have reduced the incidence of extremely hazardous conditions there. Estache (1991) reported substantial improvements in São Paulo's air quality. Comparable time series data for Korea are unavailable, but the implementation of measures such as forced switching toward lower-sulfur fuels in the Seoul region will in all probability improve air quality.

*Indonesia, Nigeria, and Thailand: Three developing countries.* Indonesia, Nigeria, and Thailand do not have well-established regulatory systems.<sup>19</sup> Various ministries in all three countries have fragmented responsibility for pollution control. Where pollution regulation has been implemented in response to local crises, it has almost always been quantity-based. Indonesia has witnessed the forced installation of abatement equipment by a few firms in Surabaya in the aftermath of a water quality emergency. The Indonesian Ministry of Population and Environment recently began negotiating cleanup agreements with major water polluters in Jakarta and other urban and industrial areas.

Similar conditions have prevailed in Thailand, although Kritiporn, Panayotou, and Charnprateep (1990) noted two interesting precedents for pollution charges there. In the early 1970s the Mae Klong River experienced a rapid increase in biochemical oxygen demand as a result of effluent discharges from new sugar mills. In response to a perceived crisis, the Thai government established the country's first central treatment facility for industrial waste water. Funds were taken initially from the sugar price stabilization fund, but in succeeding years the fund was compensated with charges levied on the industrial users. In instituting this policy, the Thai government was in effect moving toward the mixed system that has long characterized water pollution control in Western Europe.

In 1988 the Thai government established the Bang Khuntien Treatment Center to remove heavy metals from the waste water of 200 small- and medium-size electroplating factories in Bangkok. The government assesses factories for the effluent according to the type and quan-

tity of waste and the transport distance to the treatment center. Long-term contracts ensure an adequate treatment volume for the center. At present, the system is considered an operational success.

In Nigeria there has been almost no use of pollution control instruments, given the absence of an appropriate regulatory apparatus. A traditional corpus of law has in principle given local authorities the right to take gross polluters to court and private parties the right to sue for damages under tort law, although most attempts are reported to have been ineffective. Because problems of official corruption have been rife, it is not clear how an effective program of local inspection and enforcement would be carried out.

*Summary of non-OECD experience.* Experience outside the OECD seems closely related to experience in OECD countries. Most regulation is quantity-based, with existing pollution charge arrangements more closely akin to fines than to true price-based instruments. One exception to the general rule seems to be the recent introduction of tradable permits by Singapore as a means of flexibly carrying out its self-imposed limitation on use of ozone-depleting chlorofluorocarbons (CFCs) under the Montreal Protocol. O'Connor (1991) reports the successful implementation of an auction-based system. Local CFC prices escalated rapidly, and many substitution possibilities were quickly revealed. In this sphere, at least, the Singapore system seems to have defined the world "state of the art" for tradable permit systems. Otherwise, quantity-based systems have clearly been the norm, despite much recent discussion of alternatives.

### **Incentive regulation mechanisms**

Having briefly reviewed the theory and experience of environmental policy instruments, we turn now to the more general issue of creating the incentives needed to engender the appropriate response to policy instruments, whatever form they take.

#### *Overview of the mechanisms*

Regulated firms are required to perform a variety of tasks. They are asked to produce and price their products efficiently, where efficient production involves the appropriate choice of technique and operation of plant and equipment, and efficient pricing involves appropriate attention to production costs. In certain situations, firms are required to undertake particular activities deemed to be welfare-improving. One example is the regulatory requirement that energy utilities offer energy conservation programs to their customers.

While it is often easy to identify the general normative prescriptions needed to effectuate these goals, it is frequently difficult to identify and incentivize the specific behaviors required to implement these prescriptions. For example, one regulatory goal would be for the firm to price at marginal cost. However, in many cases, neither the firm nor the regulator knows the marginal cost. And even if the firm knew marginal cost, it might not be in its self-interest to report that information to the regulator.

In these situations the regulator must use normative economic theory to identify the behavior desired of the regulated individuals and positive behavioral theory to identify the incentive structures necessary to stimulate their compliance. Incentive regulation mechanisms are designed precisely for these situations. The mechanisms are designed to incentivize both the revelation of information to the regulator and the desired welfare-maximizing behavior. The focus of the mechanisms is incentives, independent of what is being incentivized.

As mentioned earlier, there exist a variety of incentive regulation mechanisms, each of which incentivizes different behavior. In order to portray the breadth of their applications, we briefly describe several here.

*Price caps.* Historically, energy and telecommunications utilities have set the prices of their services to cover their costs (total revenue requirements), no matter what the level of the costs. The result often has been inefficient choice of technique, inefficient operation of plant and equipment, and inefficient service pricing, because the utilities have no incentive to minimize costs.

Price caps have been proposed as a mechanism to improve the relevant incentives. Under a price cap regime, price increases for the products of a regulated producer are constrained by an index of producer prices corrected for productivity increases. Specifically, denoting the vectors of regulated prices and quantities in period  $i$  as  $p^i$  and  $q^i$ , respectively, under price caps the regulator will allow the regulated producer to increase prices from  $p^0$  to  $p^1$ , if  $p^1$  is constrained as follows:

$$(11.1) \quad p^1 q^0 / p^0 q^0 = FPI^1 / FPI^0 - X,$$

where  $p^i q^j$  is the inner product of the relevant price and quantity vectors; FPI is the factor price index in period  $i = 0$  and  $i = 1$ ; and  $X$  is the percentage increase in factor productivity.<sup>20</sup>

This incentive mechanism has a number of desirable characteristics.<sup>21</sup> First, it provides the correct incentives because the prices charged by the regulated firm are not directly related to its costs. Hence, if the regulated firm

can increase its efficiency, it will be able to retain the profits thereby generated. Furthermore, the regulated firm will have no reason to misrepresent its costs. The regulatory system implementing the price caps requires little in the way of administrative burden. Finally, price caps continue to provide downward pressure on the prices of regulated products.

*Automatic rate adjustment mechanisms (ARAMs).* ARAMs are formulas that allow a regulated producer to incorporate unpredictable factor price changes (increases and decreases) into regulated product prices. For example, an automatic rate adjustment mechanism may allow a regulated producer of electricity to adjust prices to reflect an unanticipated increase in oil prices.<sup>22</sup>

ARAMs also are frequently used to incentivize the improved use of specific factors of production. For example, in an attempt to promote optimal capital investment and utilization, the U.S. Federal Energy Regulatory Commission has proposed rate-setting procedures based on target capacity utilization rates. Accordingly, utilities would lose money (through reduced product prices) when capacity utilization falls below target and would make money (through increased product prices) when capacity utilization is above target.

This ARAM takes the following form:

$$(11.2) \quad p^1 q^0 / p^0 q^0 = 1 + a^*(CU - CU^*),$$

where  $p^1$  is the price vector allowed by the regulator;  $CU$  and  $CU^*$  are the actual and targeted capacity utilization; and  $a$  is the proportion of the ARAM passed through. For simplicity, assume that there are no increases in factor prices.<sup>23</sup> If  $a = 0$ , no adjustment is allowed and regulated product prices are held constant, no matter what the level of actual capacity utilization. If  $a = 1$  and actual capacity utilization is greater than the target, the producer benefits completely in the form of higher prices. If  $0 < a < 1$ , the producer and consumers share the benefits of efficient capacity use. And if  $CU < CU^*$ , the producer will be penalized with reduced rates and reduced profits.<sup>24</sup>

*Sliding-scale profit-sharing plans.* Profit-sharing plans are analogous to the ARAM in equation 11.2. In equation 11.2,  $a$  is a sliding scale that distributes the rewards and penalties from improved capacity utilization between the regulated producer and the consumers.<sup>25</sup> With profit-sharing mechanisms, the sliding scale is designed to distribute profits in excess of a target rate of return.

Under traditional rate-of-return regulation, a regulated producer is allowed a target rate of return of  $s^*$ . If that producer were to earn  $s > s^*$  because of efficiency

improvements, the producer would be unable to retain any of the gains from productivity ( $a = 1$  in equation 11.3). The amount  $(s - s^*)$  would be returned to the rate payers. Under a sliding-scale profit-sharing plan, the regulated producer would be allowed to retain some of the profits earned from cost reductions. Specifically, for a single-product regulated producer, we have:

$$(11.3) \quad p^1 = p^0 + a^*(s^* - s^0)K_0/q_0,$$

where  $p^i$  is the regulated price in period  $i$ ;  $s^*$  is the target rate of return;  $s^0$  is the actual rate of return in period 0;  $K_0$  is the rate base; and  $q_0$  is the quantity produced in period 0. Again, assume that factor prices have not increased from period to period. Notice that in this case, if  $s^0 > s^*$ , the regulated producer will lower regulated prices to reflect efficiency gains. However, if  $a < 1$ , the new prices will not reflect the entire productivity savings. The regulated producer retains a portion of the gains.

*Specific performance incentive mechanisms.* Some regulators have used sliding-scale sharing plans to stimulate regulated producers to perform specific activities deemed to be in the public interest. For example, suppose that a given activity, for example energy conservation, is projected to increase consumer surplus by the amount  $CS$ . The regulated producer will have greater incentive to implement the programs that generate  $CS$  if some portion of  $CS$  is shared with that producer. That arrangement can be accomplished as follows:

$$(11.4) \quad p^1q^0 - p^0q^0 = a^*CS,$$

where the producer retains some proportion,  $a < 1$ , of  $CS$  in the form of higher prices.

#### *Selected experience with incentive mechanisms*

There has been considerable experience with all of the incentive mechanisms introduced above. To be brief, we merely indicate when and where most of the experiments have been implemented.<sup>26</sup> However, we discuss specific performance incentives in some detail.

Experimentation with price caps has occurred most frequently in the telecommunications industry. The best known implementation involved the privatization of British Telecom in 1984.<sup>27</sup> The retail price index was used for  $FPI$  in equation 11.1, and a 3 percent limit was set for  $X$ .<sup>28</sup> The program has been considered successful. The earliest example of telecommunications price cap regulation in the United States was adopted by the Michigan Public Service Commission for the Michigan Bell

Telephone Company in 1980 (see Face 1988). A variety of other telecommunications applications have followed.<sup>29</sup> The success of the British Telecom and the U.S. programs prompted the Federal Communications Commission (1987) to propose a similar regulatory scheme in the United States.

Because they come in so many forms, ARAMs are the predominant form of incentive regulation in the United States. Most of the mechanisms allow for a partial pass-through of cost savings or overruns, where savings and overruns are calculated by comparing actual performance with targeted performance, as in equation 11.2. For one example, the Arizona Public Utilities Commission initiated the Operative Incentive Plan in 1984 to stimulate operating efficiency at two power plants owned by the Arizona Public Service Company—the Palo Verde #1 nuclear plant and the coal-fired Four Corners plant. Under the plan performance is measured by the “capacity factor,” a measure of capacity utilization. “Yardstick” performance targets for capacity utilization were derived from an industry-wide analysis of similar units.

The target for the Palo Verde plant has been a deadband of capacity factors ranging from 60 to 75 percent.<sup>30</sup> The deadband allows for unforeseeable events and differences between the regulated producer and the control group built into the yardstick. If actual performance falls within the deadband, no penalty or reward occurs. If the actual capacity factor falls within the range of 75 to 85 percent (50 to 60 percent), the Arizona Public Service Company is rewarded (penalized) by 50 percent ( $a = 0.50$  in equation 11.2) of the fuel savings (additional costs). For capacity factors above 85 percent (between 35 and 50 percent), the company is rewarded (penalized) with 100 percent ( $a = 1.0$  in equation 11.2) of the fuel savings (additional fuel costs). The Arizona ARAM has been considered a success overall.<sup>31</sup> Other ARAMs are analogous and numerous.<sup>32</sup>

The earliest regulated sliding-scale profit-sharing scheme seems to have been a program adopted by the District of Columbia Public Utility Commission in 1925 for Potomac Electric Power Company (Pepco). In 1925 the District’s electricity rates were among the highest in the country. Under the program, a target rate of return of 7.5 percent was set for Pepco. If the actual rate of return exceeded the target as a result of productivity increases, Pepco retained the full amount of the excess in the first year. In each subsequent year in which there was a productivity gain, Pepco could retain 50 percent of the previous year’s reward. For example, if Pepco were to institute production changes that generated \$100,000

savings in year  $t$ , it could retain all of those savings in year  $t$ . If the production savings continued in subsequent years, Pepco could retain 50 percent of \$100,000, or \$50,000, in year  $t+1$ ; 50 percent of \$50,000, or \$25,000, in year  $t+2$ ; and so on. Both Pepco and its customers benefited. Residential rates fell by 50 percent from 1926 to 1934, while national rates fell by only 29 percent. Pepco earned rates of return in excess of 8.8 percent during this period. Target rates of return were later lowered to 7 percent, 6.5 percent, and 5.5 percent in 1948.<sup>33</sup>

Finally, specific-performance incentive mechanisms have been designed to stimulate a regulated producer to undertake a variety of activities deemed to be conducive to the public welfare. For example, since passage of the Public Utility Regulatory Policy Act of 1978, a variety of utility-sponsored energy conservation and load management programs have been offered as a result of pressure from state public utility commissions. In the years immediately following enactment of the 1978 law, these programs were mandated by the commissions. By the mid-1980s, the programs were less aggressively pursued. However, renewed interest has recently emerged, and incentive regulation has played a role in the current efforts.

The experience of the California utilities is instructive. From August 1989 through January 1990, representatives of California's major energy policy stakeholders collaborated to identify regulatory approaches that would be most productive in continuing the energy conservation momentum of the 1980s.<sup>34</sup> In light of increased discussion of regulated incentive mechanisms, the group agreed to allow the major California utilities to develop incentive mechanisms to promote programs that stimulated durable, persistent, and reliable energy efficiency savings. A wide variety of "demand-side management" programs were candidates for the incentive mechanisms, including the following:

- Programs that reduce energy use while maintaining service levels.
- Programs that encourage builders to add efficient designs and features to new buildings in excess of state standards.
- Programs encouraging the purchase of energy-using equipment that is more efficient than state and federal equipment efficiency standards.
- Low-income programs.
- Programs with "resource value."
- Marketing and information programs operated in conjunction with energy efficiency programs.
- Innovative energy efficiency programs.
- Certain load management and load-shifting programs.

The California utility commission set minimum performance targets for these candidate programs and left day-to-day design and operation of the programs to the utilities. Regulatory involvement therefore focused on compensating utilities for achieving program goals while measuring energy savings (and other relevant information) for the purpose of setting future incentive levels.

The specific incentive programs proposed by the major utilities as part of the collaborative exercise included the following:

*Pacific Gas and Electric (PG&E).* PG&E proposed a mechanism whereby the benefits of investment in demand-side management programs are shared by the ratepayers and PG&E shareholders. The incentives apply to either program expenditures or program-induced savings. Program expenditures are simply those resources spent on program implementation. Program-induced savings are defined to be the difference between the costs of the specific conservation and load-management programs and the costs of the supply power avoided (benefits equal "avoided costs") as a result of the programs.<sup>35</sup>

Specifically, PG&E shareholders receive 5 percent ( $a = 0.05$ ) of the expenditures for such programs as direct weatherization, low-income assistance, information, and several other programs that do not produce energy savings directly but that fulfill customer service, equity, or social policy goals. PG&E shareholders receive a percentage ( $a = 0.15$ ) of the estimated savings produced by conservation and load management programs that are cost-effective, contribute to energy efficiency, and create resource value. PG&E recovers administrative costs under all programs.

*San Diego Gas and Electric (SDG&E).* For purposes of incentive mechanisms, SDG&E assigned its programs to two categories. The first category includes programs that can achieve measurable energy savings. For these programs, SDG&E proposed to retain 30 percent ( $a = 0.30$ ) of the present value life-cycle savings above a certain minimum threshold, where savings are calculated as the difference between "avoided costs" and program costs. The second category includes those programs for which energy savings are uncertain or difficult to quantify. SDG&E proposed a 10 percent incentive for all program expenditures in this category.

Notice that these two categories are analogous to those of PG&E but that the sharing factors,  $a$ , differ from those of PG&E.

*Southern California Edison (SCE).* SCE proposed to amortize a portion of its demand-side management pro-

gram expenses as an incentive mechanism. Amortization is limited to programs for which energy savings are measurable. Amortization defers the recovery in rates of current program expenditures by creating a “demand-side management asset.” The rate-making treatment of these deferred expenditures parallels the rate-base treatment of all utility-owned generation assets.

*Southern California Gas Company (SoCalGas).* For purposes of incentive mechanisms, SoCalGas identified two categories of programs: mature, full-scale programs that are known to deliver savings to the service territory; and experimental, “test” programs whose effectiveness remains questionable. For the mature programs, shareholders are allowed to receive a maximum of 16.6 percent ( $a = 0.166$ ) of the net social benefits (“avoided costs” minus program expenditures) above a minimum threshold. An increasing sliding scale operates up to the 16.6 percent level. For experimental programs, SoCalGas shareholders receive 16.6 percent of program expenditures.

#### *Lessons learned*

A wide variety of hybrid incentive mechanisms have been crafted to improve the overall operational performance of regulated producers and stimulate their specific performance. The accumulated experience indicates that it is invariably not enough to propose “normative” behavioral rules to the regulated agents.<sup>36</sup> Additional attention is required to incentivize the behavior sought by regulators.

Experience indicates that incentive mechanisms work: They affect and alter behavior. They do so by offering explicit rewards for desired behavior and explicit penalties for undesired behavior. These rewards are *not* subsidies. They are payments for actual social services.

The best mechanisms are those that elicit information from the regulated agents and that involve the sharing of the risks, costs, and benefits of regulatory compliance between the regulated producers and the consuming public.

Most instructive, perhaps, is the experience of the California regulators in stimulating regulated utilities’ participation in the conservation program. Before the implementation of incentive regulation, only some utilities complied with conservation targets. Hence, it was simply not sufficient for the regulators to state that “conservation is important, and utilities should encourage it as long as the program gains (avoided cost of supply power) are greater than the program cost.” For some utilities, it is necessary to develop incentives making conservation explicitly in the utility’s self-interest.

### **A proposal for environmental regulation incentive mechanisms**

This section suggests a regulatory procedure for incentivizing environmental policy compliance. The methods proposed make use of effluent standards. Specifically, target effluent standards and deadbands around the targets are developed by the regulator for each polluter (for our purposes, a firm). The targets are based on a yardstick method and communicated to the regulated firms at the beginning of each production period. Firms then undertake their production activities for a complete accounting period and report their final effluent levels to the regulator. Rewards are paid to firms that meet or surpass the targets, whereas penalties are imposed on firms that do not meet their targets.<sup>37</sup>

The reporting procedure is designed to be incentive-compatible and to elicit information regarding producers’ compliance. We expect that the firms that report their effluent levels will be those in compliance—hence, those that expect to receive a reward for exceeding their targets. The remaining firms, some of which may not report, will be those that have not met their targets. This self-selection of firms into reporting and nonreporting groups will provide useful information for monitoring overall compliance.<sup>38</sup>

We would argue that the operation of the proposed incentive system will mimic an emissions bank, or a market for tradable permits. Ultimately, the proposed incentive mechanism could be privatized as an emissions bank.

#### *Proposed incentive system*

The following discussion is intended to be heuristic and to provide an overall understanding of the proposed incentive system.<sup>39</sup>

*Developing effluent targets and deadbands.* Regulation of public utilities requires scrutiny of a relatively small number of regulated parties. As a result, it is not overwhelmingly difficult to develop targets for each regulated producer. With environmental policy compliance, however, targets must be developed for all polluters in a relevant environmental area. This task may seem daunting.

To begin, the regulator first must identify the pollutants that will be subject to regulation. The most basic list would include the “criteria” air and water pollutants prescribed by the U.S. Environmental Protection Agency pursuant to the Clean Air Act Amendments and Clean Water Act Amendments. Some toxic pollutants identified by the Clean Air Act Amendments, the Clean Water Act Amendments, or the Comprehensive Environmental Response, Compensation, and Liability Act (Superfund)

should also be included.<sup>40</sup> The comprehensiveness of the list is left to the discretion of the regulator. At the start of the implementation of this regulatory scheme, the list can be fairly short. The list can be extended once the regulatory structure is in place.

Second, the regulator must identify the control region that will be subject to regulation. This region's boundaries will depend on the pollutants being controlled. For example, some air and water pollutants dissipate within a given airshed or watershed, which would be the appropriate jurisdictional area for those pollutants. Other pollutants (air, water, and land) are global.

Third, the regulator must calculate total allowable effluent levels for the specific pollutants in specific environmental areas. For example, in a local airshed, total levels of carbon monoxide may be prescribed to ensure an ambient air quality that is locally appropriate. Or, total levels of nitric oxide and sulfur dioxide may be prescribed for a broader national or international airshed. Likewise, total effluent levels for waterborne residuals and solid waste must be prescribed.

Fourth, the regulator must set effluent targets for each economic agent that produces a specific pollutant within a specific environmental area. The naive method for allocating effluent targets among the  $N$  firms in a specific environmental area would be to divide the total allowable effluent level by  $N$ . The targets incorporated into the recent U.S. program for tradable permits for sulfur dioxide emissions are analogous to such a system.<sup>41</sup>

As long as free and fully informed trading occurs, it makes little difference from an efficiency point of view how the initial targets are set.<sup>42</sup> If, however, markets are not fully developed initially, and if players have differential information regarding abatement technologies and possibilities, then there will be advantages for the regulator to set differentiated targets, moving the targets in the direction that would occur under trading. Such differentiation should take account of the following six factors:

- Scale and scope of production and the nature of the products produced by each firm.
- Technology in place.
- Vintage of the plant and equipment.
- Abatement efforts already undertaken.
- Type of effluents emitted.
- Abatement technologies available, including those "reasonably available" and "best available," and their costs.

For example, firms characterized by high levels of emissions, large production scale, and low marginal abatement costs (induced, say, by type and scale of production

process) are more efficient abaters, and thus would sell their emissions permits. From an efficiency standpoint, these firms should be given higher abatement targets (equivalently, lower emissions targets) at the outset.<sup>43</sup>

The best method for incorporating these considerations is a yardstick approach that combines regression and engineering analysis. This approach would begin by identifying a control group of firms with similar products, production technologies, factor demands, and environmental impacts. A starting criterion for such a control group would be inclusion in some set of sectors identified by their International Standard Industrial Classification (ISIC).<sup>44</sup> For this group, regression models could be estimated relating effluent levels to the six factors above.

Once the effects of the six factors are assessed, the impact and costs of reasonably available control technologies (RACT) and best-available control technologies (BACT) must be included. These effects can be included through engineering analysis. Alternatively, regression analysis could be used, if there exist data on plants and firms that have put BACT and RACT in place.

Once the yardstick analyses are completed, we can differentiate targets for pollutant  $i$  for a specific firm  $j$  as follows. Letting  $\text{target}_{ij} = \text{tgt}_{ij} = \text{pollutant}/\text{firm}_j$  (or alternatively,  $\text{pollutant}/\text{output}_j = \text{target}_{ij} = \text{tgt}_{ij}$ ), we have

$$(11.5) \quad \text{tgt}_{ij} = F$$

where  $\text{pollutant}_i$  is the level of effluent  $i$  emitted by firm  $j$ ;  $\text{output}_j$  is some measure of production for firm  $j$ ; and  $\text{firm}_j$  is characterized by all of the attributes: scale and scope of production and nature of products produced by firm; technology in place; capital vintage; abatement efforts already undertaken; abatement technology available (RACT/BACT); and cost of abatement technology available. Equation  $F$  could be estimated by regression analysis and refined by engineering analysis for all plants and firms in the control group.<sup>45</sup>

Once the differentiated targets are ascertained for each pollutant  $i$  and firm  $j$ , we can develop deadbands<sup>46</sup> of the following sort:

$$(11.6) \quad (\text{tgt}_{ij} - s_{ij}, \text{tgt}_{ij} + s_{ij}),$$

where  $s_{ij}$  can be determined by a variety of methods. It can vary by firm and pollutant; it can be constant across firm and pollutant; or it can be set at an arbitrary level, say 5 percent ( $X$  percent) of the target. Alternatively, if regression techniques are used, the standard error of the regression can be set to make equation 11.6 a 95 percent (or 90 percent) confidence interval around the target.<sup>47</sup>

*Structuring the rewards and penalties.* Once the targets and deadbands are set, rewards are calculated as:

$$(11.7) \quad R_{ij} = SV_i^* a_{ir}^* Q_i^* [(tgt_{ij} - s_{ij}) - EI_{ij}]$$

and penalties as:

$$(11.8) \quad P_{ij} = SV_i^* a_{ip}^* Q_i^* [EI_{ij} - (tgt_{ij} + s_{ij})],$$

where  $R_{ij}$  is the reward paid to firm  $j$  for its exceeding the lower deadband limit around its effluent target  $\{[(tgt_{ij} - s_{ij}) - EI_{ij}] > 0\}$ , where  $EI_{ij}$  is the effluent intensity of firm  $j$  and pollutant  $i$ ;  $P_{ij}$  is the penalty paid by firm  $j$  for its noncompliance with the upper deadband limit around its effluent target  $\{[EI_{ij} - (tgt_{ij} + s_{ij})] > 0\}$ ;  $SV_i$  is a measure of the social value of the abatement of a unit of effluent  $i$ ;  $Q_i$  translates effluent intensities and targets into units of effluent  $i$ ; and  $a$  is the sharing factor, which can be differentiated by pollutant and by reward or penalty (that is,  $a_{ir}$  or  $a_{ip}$ ).<sup>48</sup>

Performance within the deadband is neither rewarded nor penalized. The deadband allows for the possibility that an actual firm may differ from the target, due to the following: special circumstances; inability to measure the variables entering into  $F$  in equation 11.5; or exclusion of important variables from equation 11.5.

*Information flows.* One primary responsibility of the regulator is to gather and update economic, epidemiological, and engineering data for estimating and improving the components of equations 11.5–11.8. A second responsibility is to announce the targets and deadbands at the beginning of a production period. The length of the production period should be set to accommodate the incentive structure. For example, a one-year period is probably too short, whereas a five-year period is probably too long.<sup>49</sup> For purposes of the current exposition, let us say that the accounting period is three years.

At the beginning of each accounting period, the regulator announces the targets and deadbands for all firms and all pollutants subject to control, either on an annual basis or for the entire accounting period. The firms then proceed to undertake their normal production decisions, subject to the environmental system of rewards and penalties and all other constraints.

At the end of the accounting period, the reporting will proceed as follows:

First, all firms that are due rewards (by their own calculation) are invited to report their environmental record (that is, their actual levels of  $EI_{ij}$ ). These firms will come forward eagerly. The reporting procedure could be made

part of the normal annual corporate report. Firms that exceed their effluent targets will, no doubt, tout the accomplishment in their annual reports. After scrutinizing the environmental record of each firm, the regulator will pay rewards to those that substantiate above-target performance. If a firm does not substantiate its environmental performance, the regulator will impose a serious fine. It is assumed that this system will stimulate a reporting self-selection such that firms that report will be those in compliance.

Second, all firms that fall within the compliance deadband (by their own calculation) are invited to report their environmental record. As above, this reporting procedure could be made part of the normal annual corporate report. These firms also could publicize the fact that they met their effluent targets. These firms will come forward naturally, though not as eagerly as those receiving rewards. Again, the regulator will scrutinize the environmental record of each of these firms and, if the environmental performance is not substantiated, will impose serious fines.

Finally, all remaining firms that have not come forward can be assumed to be in noncompliance and subject to fines. They will be invited to report their environmental record. The regulator will impose fines on these firms according to equation 11.8. Furthermore, if the environmental performance of this group is worse than claimed, the regulator will impose additional serious fines.

This reporting procedure will economize on monitoring requirements and resources, relative to a system in which all firms have the same incentive to report compliance performance.

*Implementation issues.* As with any new regulatory procedure or institution, it would pay to start modestly, perhaps with a few pollutants. The regulatory scope could then be expanded gradually as the institutions and procedures become set.

*Monitoring and enforcement.* The international experience described earlier indicates that if regulators lack the political will to enforce the regulations, they might as well not waste social resources by pretending to implement environmental regulations. The relevance to our proposed system is that the proposed rewards must be paid. Likewise, penalties must be imposed, and serious additional penalties (fines or imprisonment) must be levied for evasion and misrepresentation.

*Asymptotic shadow institution.* The incentive mechanisms proposed above are founded on the well-established principles of rewarding approved behavior and punishing disapproved behavior. However, the incentives

are quite similar to those underlying a tradable emissions permit system. This is not serendipitous. We have designed our incentive regulation system to be undertaken with the explicit intent of ultimately transforming it into a *tradable permit system* under an emissions bank. The transition to the emissions bank with tradable permits would be accomplished as follows.

Once the incentive regulation mechanisms have been put in place, they can be privatized by transferral to an emissions bank. In this case the targets would still be imposed on every polluter, as calculated naively or using equation 11.5. The deadbands may still be used; however, it is possible that they may be eliminated. Once the targets are set, each firm must comply with them or pay another firm to comply for it. In this case, firms that are out of compliance (that is, produce effluents in excess of their assigned targets) and inefficient abaters will find it cost-effective to pay firms that are in compliance and efficient abaters, to reduce their effluent levels below their assigned targets.

As is well known, such a system has a variety of desirable properties. Total pollution can be reduced to target levels, but abatement activity is not across the board.<sup>50</sup> Rather, more abatement activity is provided by firms that are more efficient in abating. The total social cost of abatement is thereby minimized (see figure 11.2). The initial payment by noncompliant polluters to compliant polluters that are efficient abaters could be the penalty payments in equation 11.8. If the penalties are set equal to the rewards, then the noncomplying polluters would pay a fee to the complying polluters that were efficient abaters.<sup>51</sup> If an abatement market is thereby stimulated, the penalties plus rewards will be equal to the social cost of abatement.<sup>52</sup>

As discussed at the beginning of this chapter, such abatement procedures have precedent in the “offset” policies introduced by the 1977 Clean Air Amendments. Under those amendments, new pollution sources (plants) for a given firm in a nonattainment area can be built only if the firm offsets the “new” pollution by halting an equivalent amount of “old” pollution at an existing pollution source (plants). Ambient air quality must be maintained under the hypothetical bubble drawn over a group of productive activities. Ambient quality will be maintained if new pollution sources are offset by stopping old pollution sources. In the original amendments, the bubbles were drawn over the plants of a single firm and the offsets were within-firm exclusively.<sup>53</sup> However, recent applications have been regional, allowing offsets between firms.<sup>54</sup> Many of these between-firm offsets have been negotiated as part of state implementation plans.

The potential for trading offsets has been explored by the Bay Area Air Quality Management District in San Francisco. The procedures specify the extent of the offset that must be obtained to compensate for any new emission sources. This is determined by the type of pollutant and the distance between the new emission source and the offsetting emission obtained elsewhere. For example, for a nitrogen oxide offset within 15 miles of a new nitrogen oxide emission source, the offset must be within a 1.2 to 1 ratio, that is, 1.2 tons of nitrogen oxide in offset for 1 ton of new nitrogen oxide emissions.

Given this specificity, informal markets are developing for trading the offsets among firms. For example, a new terminal for petroleum processing being constructed by Wickland Oil Company will result in higher emissions of sulfur dioxide and hydrocarbons. To offset the emissions, Wickland will pay for the abatement equipment of a nearby dry cleaning operation, will buy and close down a chemical company, and will buy low-sulfur fuel for some shipping in San Francisco Bay.

These offsets are true Coasian market responses, which could be formalized through an emissions bank. Polluters in compliance (that is, meeting their targets) that are efficient abaters (that is, produce abatement so that their effluents are below the deadbands in equation 11.6) could deposit in the emissions bank abatement credits (offsets) in excess of their targets (or deadband limits). These credits could be auctioned to polluters out of compliance (that is, with effluents above their targets or deadband limits) or to new sources of pollution. If the market proves to be an auction market, the auction price will be the social cost of abatement.

### **Agenda for further research**

Clearly, further research is required to more completely articulate the details of the incentive regulation scheme introduced heuristically above. Once the details have been articulated, research is needed to clarify how much more preferable our proposed system will be than regulatory systems currently in operation in the real world. We focus here on the former research agenda.

#### *Determination of aggregate abatement targets*

In figure 11.1 we made use of standard textbook paradigms: the social demand for abatement and the social supply of abatement. These curves are the starting point of almost all analyses and discussions of environmental policy. Their intersection identifies the socially desirable level of aggregate pollution by identifying the required level of abatement ( $N^*$ ) and the social price of abatement ( $P^*$ ).

Estimation of the social demand curve (MSD) for abatement requires either market prices or proxies for market prices as the measures of social value. For existing goods and services provided into already-functioning markets, market prices should be available for both measuring value and cost-benefit calculations. In the absence of a market for abatement services or environmental quality, contingent valuation survey methods are required. These methods have been used extensively to value nonmarketed environmental goods and services.<sup>55</sup> They have also been used to value changes in the quality and reliability of existing goods and services, such as improved electrical service.<sup>56</sup>

The social abatement cost function in figure 11.1 represents the horizontal summation of marginal cost curves of abatement for all potential and actual abaters in the economy. At least three approaches are possible for estimating these cost curves: (a) econometric cost functions fitted to data from economies where extensive regulation is already in place; (b) standard, equipment-specific engineering cost functions adjusted for local input costs; and (c) cost functions fitted to data for the country itself. If the economy in question is not formally regulated in the first period, the third option might seem unrealistic. However, recent research has revealed tremendous variation in emissions-reducing behavior, even in “unregulated” developing economies. This variation seems to reflect several factors, including “informal regulation” by local communities, which is more or less effective under differing circumstances.

Research in this area should consist of, but not be limited to, the following tasks:

- Identification of the determinants of the appropriate size of the watershed or airshed (that is, control region) for each of the pollutants being regulated.
- Identification of the appropriate units in which to measure abatement ( $N$ ) for relevant pollutants.
- Adaptation, design, and implementation of contingent valuation survey methods for the estimation of the social demand for abatement of specific pollutants within the relevant air- and watersheds.
- Estimation of the abatement cost functions for all possible abaters, using all three approaches identified above;<sup>57</sup> comparison of the alternative cost estimates; and development of the social cost curve by aggregating the cost curves across all abaters.
- Development of aggregate regional incentive regulation targets,  $N^*$  (and implicit shadow charges  $P^*$ ), for selected countries and economies. In each control region, the intersection of the marginal social demand (MSD)

and the marginal social cleanup cost (MSC) curves will be strongly affected by local industry structure, climate, topography, demography, and ecology.

#### *Determination of firm-specific targets*

Perhaps the most controversial issue for our incentive regulation proposal will be the setting of targets, which we have heuristically described using the “black box” equation 11.5. Two alternative extremes are possible:

- *Alternative A:* A “zero information” system that (a) imposes equal percentage reductions or equal abatement requirements or identical effluent concentrations on firms; and (b) uses relatively large, symmetric payments and charges for over- and underachievement, respectively.
- *Alternative B:* A “maximum information” system in which the targets would equalize marginal abatement costs across firms at the globally appropriate shadow charge on emissions. In this system neither rewards nor penalties would be necessary, since the incentive regulation targets would already be optimal.

Alternative A would generate less political controversy regarding equity, since all polluters are treated “equally.” However, these targets are analogous to the quantity regulations found in figure 11.2 and are therefore not the most efficient. As a result, the rewards and penalties paid to effectuate these targets would be useful for (a) overcoming status quo biases; (b) stimulating the desirable information flows; and (c) stimulating the firms to think about interfirm trading of abatement responsibility. However, if all firms were to meet these targets, abatement would not be Pareto optimal.

Alternative B is not possible in its ideal form; the information requirements are too onerous. However, some version is possible that would incorporate more information into the targeting equation  $F$  (11.5). For example, the econometric estimation of abatement costs by firm discussed in the preceding subsection can be introduced so that the least-cost abaters are given greater abatement responsibility.

Additional research is required to address the following questions:

- Given estimates of optimal abatement  $N^*$  (developed in the preceding subsection), what are the appropriate units in which to articulate the targets to firms?
- What equity arguments will arise with setting abatement targets according to alternative A? How can they be defused or countered?
- How can additional abatement cost information be introduced into equation 11.5? Examples of econometric cost analysis are provided by Hartman, Wheeler, and

Singh (1994), who estimated abatement costs conditional on the nature of products produced and the abatement efforts already undertaken. Current extensions of this analysis are designed to include the scale and scope of production.

- What equity arguments will arise for setting abatement targets according to alternative B (or some version of alternative B)? How can they be defused or countered?
- How will the desired transition to an emissions bank for tradable permits determine the nature of the incentive regulation targets? Under alternative A, there will be a great demand for trading permits because abatement requirements will be a long way from least cost. Under the ideal version of alternative B, there will be no demand for trading permits because abatement requirements will be set at least cost. If the regulators plan for a quick transition to tradable permits, alternative A may be the best method for assigning targets. If plans for transition to a tradable permit system are still vague and unformed, some version of alternative B will be preferable. Can these qualitative issues be made more quantitative?
- Can econometric methods be used to quantify equation 11.5? If so, can we identify the control groups over which equation 11.5 is to be estimated?<sup>58</sup>
- What is the best method for designing and estimating the deadbands?

#### *Determining time periods for compliance and revising targets*

As information is revealed by firms under incentive regulation, it may be desirable to revise the targets. This is particularly true if the targets are set according to alternative A, above. However, the desirability of revising targets toward socially efficient levels must be balanced against the expectation problems that arise for abaters when targeted responsibilities change in unanticipated ways.

For an incentive regulation system to work best, annual firm-level targets should reflect anticipated progress toward standards set on a five- to ten-year horizon. During the “plan period,” firms’ target paths should not be revised. This approach will, of course, have a “cost of being wrong,” but it will ensure generally appropriate incentives for improvement.

At the end of the plan period, however, there seems to be a good case for revision of targets. Long-run overperformance almost certainly implies overestimation of abatement costs, whereas long-run underperformance implies the converse. Clearly, of course, the planning period should be long enough and the penalties for underperformance high enough that recalcitrant firms

would not find it worthwhile to stall on emissions reduction until the revision round.

The research problem is therefore relatively clear: intertemporal stability and consistency must be balanced against the need for error correction, with optimal timing undoubtedly affected by differences in the characteristics of pollutants and control regions.

#### *Introduction of trading*

While we have focused primarily on the nature of the incentive regulation system, it is clear that many of its elements will be determined by the speed and nature of the transition to tradable emission permits. The administrative and operational aspects of this transition must be clarified.

#### **Conclusions**

Experience in regulating industrial pollution suggests three main lessons. First, true Pigouvian charges, while theoretically optimal in a world without transactions costs, have almost never been successfully implemented. Regulation almost invariably relies on emissions or technology standards. Where charges have been used, they have generally been intended to cover costs for regulatory agencies and have remained substantially below the marginal cost of abatement. Nevertheless, the case for market-based instruments remains compelling.

The second lesson is that monitoring and enforcement of regulation are always limited in practice by principal-agent problems. Regulated firms always know considerably more than the regulators, and their degree of compliance depends on the incentive to comply.

Third, status quo bias is a very significant problem for price-based (indeed any) regulatory policy: Research on firms’ responses to large energy price changes has repeatedly demonstrated great stickiness over long periods of time.<sup>59</sup>

From these lessons, we draw the following conclusions:

- Explicit targeting of firm performance standards in conjunction with the explicit use of rewards and penalties for performance can be very useful as a means of motivating or “incentivizing” behavior, and overcoming status quo bias and the principal-agent problem.
- Incentive regulation significantly reduces the information problems that arise in principal-agent situations. It does so by isolating inferior performers, which can then be made the focus of selective monitoring and enforcement efforts. Both superior and target performers will find it in their interest to reveal their status voluntarily in an incentive regulation system.<sup>60</sup>

- While incentive regulation systems have faced difficult operational and equity issues, they have been successfully implemented in a wide array of industries.
- The targeting embedded in our incentive regulation system can be made approximately consistent with market-based approaches. Payments and charges are respectively assigned to over- and underachievement of performance goals. If correctly instituted, incentive regulation can operate as a “shadow emissions bank,” which can be transformed into a tradable permit system.
- Indeed, we propose incentive regulation as a means of effectuating transition from a system of effluent standards to a privatized emissions bank that will formally oversee the trading of permits.

To summarize, our reading of actual experience suggests that incentive regulation will be superior to quantity-based and price-based approaches in many cases.<sup>61</sup> Incentive regulation is closest to, and allows for easy transition to, a formal tradable permit system. Both are founded on a target level of overall emissions; both achieve efficiency gains by encouraging more reduction by low-cost abaters. However, incentive regulation is unique in imposing performance standards on individual firms while introducing the financial rewards and penalties of a market-based system. In practice incentive regulation has successfully attacked the principal-agent and status quo bias problems, which traditional regulatory regimes do not address.

### Notes

The authors gratefully acknowledge the comments of Carter Brandon, Claudio Frischtak, and Robert Hahn.

1. This comparison is usually performed in a highly stylized setting. See Weitzman 1974, 1977.
2. Examples abound. Vogelsang and Finsinger (1979) [also Finsinger and Vogelsang (1981, 1985) and Vogelsang (1989)] have proposed institutional methods for inducing regulated producers to price efficiently, even when the regulators have considerably less information than the firms being regulated, and the regulated firms find it in their self-interest to exploit that informational asymmetry. Hartman and others (1986, 1991) and Hausman (1979) have identified and analyzed situations where consumers resist Pareto superior allocations, because of “status quo” biases or other forms of “irrational” behavior. In these situations the consumers must be incrementally and independently stimulated to choose the Pareto superior allocations they resist.
3. Status quo bias refers to the observed tendency of firms and households to remain at the status quo even after very significant changes in relative prices.

4. Notice that we are looking to improve compliance only for Pareto superior performance. In the real world, it is highly unlikely that we can attain Pareto optimal results.

5. It is well known that subsidizing compliance may introduce certain inappropriate incentives. See Seneca and Taussig 1984, pp. 214–16.

6. Notice that we use an unconventional taxonomy here. Tradable permits are generally classified as a quantity-based instrument, since they impose an aggregate quantity limit on emissions. However, because we ultimately want to focus on the trading of these permits and the market prices at which they are traded, we classify them as price-based. The discussion in the text is invariant to their regulation classification as either price-based or quantity-based.

7. In the United States, for example, the Clean Water Act of 1972 required industrial dischargers to meet effluent limitations by 1977 based on the “best practicable control technology currently available.” By 1983 this language was to become the “best available technology economically achievable” (Tietenberg 1988, p. 412). Obviously, there can be a close correlation between technology and effluent standards. For example, in many cases, an effluent performance standard is a de facto technology standard.

8. Weitzman’s (1974) judgments about instrument choice derive from a stylized normative model that does not try to reflect political constraints. In the real world, judgments about instrument choice should be case-specific. In this regard, there are few published studies. Tietenberg 1988, p. 329 reports a study by Kolstad (1982), who considered the relative merits of effluent charges and tradable permits for pollution from electricity production near the Four Corners wilderness area in the western United States. Kolstad concluded that tradable permits were generally preferable.

9. For example, the installation of stack scrubbers and settling basins has generated large volumes of solid wastes that contain previously dilute effluents in highly concentrated form (Wynne 1987). These stocks become “toxic wastes” from the regulatory perspective because they are hazardous on short exposure.

10. See Mishan 1990. OECD 1989 suggests that Mishan’s assumption is often valid. In France, for example, water pollution charges are kept quite low and assessed on bulk flows of water rather than discharges of individual pollutants. The regulators monitor only the largest polluters; all others are charged at rates taken from standard tables. But this method provides small firms with an implicit bargaining chip. If the standard discharge fee is set very high, small firms can be expected to lobby actively for individual monitoring on the grounds that their own discharges are overestimated. The regulator stays with modest fees, knowing that total monitoring will not be cost-effective.

11. The collapse of the Pittsburgh steel complex provides a good illustration of how the social opportunity cost of much stricter pollution standards reveals itself over time. The U.S. iron and steel indus-

try was among those hardest hit by rapidly tightening environmental standards in the 1970s. At the same time, major new steel sources entered the U.S. market from newly industrialized countries such as the Republic of Korea and Mexico, which tolerated pollution levels much higher than the U.S. standards. See Merrifield 1988.

In the United States, much of the necessary iteration seems to occur before the fact in committee hearings and media reports. According to Rosewicz (1990), petrochemical producers began filing cost-based exemption cases from the moment the U.S. Congress passed the 1989 revision of the Clean Air Act.

12. There is little evidence that standards for acceptable ambient concentrations of any given pollutant fluctuate very much or very often. When changes do occur, they are usually in the direction of overall reduction.

13. The government can prevent such pollution by remaining the sole auctioneer of permits at regular intervals, but the attendant uncertainty will reduce the economic appeal of permits as implicit environmental property rights.

14. These uncertainties diminish, however, once we allow for iterations in setting the policy parameters.

15. See USCBO 1985; OECD 1989; Kopp, Portney, and DeWitt 1990; Hahn 1989; and Vogel 1986.

16. By the late 1960s waste discharge into Dutch surface waters had drastically exceeded their assimilative capacity. Under the 1970 Pollution of Surface Waters Act, the government instituted an emission charges system based on a unit charge per "inhabitant-equivalent" (IE)—the approximate amount of organic pollution of waste water normally produced by one person. In 1969, fourteen pollution-intensive sectors produced about 29.2 million IE—90 percent of total industrial organic water pollution in the Netherlands.

Starting in 1970, a fee of \$4.00 per IE was levied on polluting plants. During the following decade the fee rose about 83 percent (in 1970 dollars), reaching \$7.30 per IE in 1980, and total organic pollution from the fourteen sectors dropped by 69 percent, to 9.1 million IE. Since real Dutch industrial output grew about 27 percent during the decade, the implication is an approximately unitary long-run response elasticity.

It is important to note that this is a partial equilibrium result that undoubtedly overstates the response elasticity of total industrial pollution to water charges. High inter-medium substitution elasticities exist for many processes. For example, if water pollution charges are high but cheap landfill sites remain available, toxic pollutants can be separated from the waterborne waste stream and transported offsite as "solid waste." Wynne (1987) and others have noted the rapid growth of toxic waste shipments from the Netherlands and other West European countries during the period when water pollution was falling rapidly. See also Bressers 1983 and OECD 1989.

17. See Misiolek and Elder 1989 and Krishna 1988, 1990 for a discussion of similar problems in export quota markets.

18. The treatment of Brazil in this section relies heavily on Findley 1988. The material on Korea is drawn from Clifford 1990; Gresser 1979; Song 1989; Gadacz 1986; Engineering Science, Inc. 1984; Chung and Sang 1992; and O'Connor 1993.

19. The Nigerian treatment draws on Osaic-Addo 1990 and Olokesusi 1987, 1988. Indonesian material was taken from USAID 1987 and unpublished World Bank reports.

20. Notice that in this formulation, the price cap constraint is merely a Laspeyres index limitation on the product prices of a multiproduct producer.

21. The problems with traditional cost-of-service regulation and the improvements induced by price caps are developed in USFERC 1989.

Bradley and Price (1988), Vogelsang (1989), and Vogelsang and Finsinger (1979) demonstrate that a regulatory regime that correctly implements the price caps in equation 11.1 will eventually constrain a profit-maximizing producer to adopt Ramsey prices.

22. A variety of ARAMs are possible. See USFERC 1989.

23. If factor price changes were included, the ARAM would take the following form:  $p^1 q^0 / p^0 q^0 = FPI^1 / FPI^0 + a^*(CU - CU^*)$ . Obviously, this is the form that is most common, since ARAMs are used to pass along particular factor price increases.

24. Other ARAMs have been articulated to account for the use of all factors (that is, total factor productivity), rather than capital alone.

25. Some method of sharing rewards and penalties is usually preferable to the situation in which either the principal or the agent assumes all of the risk. See Baron and Myerson 1982, Cox and Isaac 1987, Cross 1970, Holmström 1979, Laffont and Tirole 1986, Shavell 1979, and Sibley 1989.

26. Other summaries are found in Edison Electric Institute 1987, Joskow and Schmalensee 1986, USFERC 1989, and USNRC 1987.

27. For discussion of this experiment, see Littlechild 1983, Bradley and Price 1988, USFERC 1989, OFTEL all years, and Bhattacharyya and Laughunn 1987.

28. See OFTEL 1985.

29. Twenty-nine states have replaced rate-of-return regulation with a price cap mechanism for prices of intrastate services provided by AT&T. Mathios and Rogers (1989) identify these states and document that AT&T's daytime, evening, nighttime, and weekend rates are significantly lower in those states with the price caps.

The British Gas Company and the British Airports Authority have been regulated through price caps; see Bradley and Price 1988.

30. "Deadbands" define the range of variation around targets within which neither rewards nor penalties are implemented.

31. See Arizona Corporation Commission 1987.

32. ARAMs have been used in regulating the performance of energy utilities in the following states: Arkansas, California, Colorado,

Connecticut, Delaware, Florida, Maryland, Massachusetts, Michigan, New Hampshire, New Jersey, New York, Ohio, Oregon, Pennsylvania, Utah, and Virginia. See Pacific Gas and Electric 1985, New York Public Service Commission 1984, 1986, Delaware Public Service Commission 1984, and USFERC 1989.

33. Sliding-scale profit-sharing incentive mechanisms have been used in Mississippi (Mississippi Public Service Commission 1986) and New York (New York Public Service Commission 1986, 1987).

34. See California Public Utilities Commission 1991, California Public Utilities Commission and California Energy Commission 1987, 1990, San Diego Gas and Electric Company 1991, Southern California Edison Company 1991, and Southern California Gas Company 1991.

35. Note the analogy to the comparison of social benefits and costs of pollution abatement in figure 11.1.

36. Naive versions of the regulated rules would have been the following: "Price efficiently, or Ramsey price" (equation 11.1); "Produce efficiently" (equations 11.2 and 11.3); and "Provide those conservation programs that are socially beneficial" (equation 11.4).

37. Of course, some notion of a budget constraint must condition the aggregate rewards and penalties.

38. This system is incentive-compatible. Even if firms understand the selection process, firms in noncompliance will have no incentive to report.

39. Many specific details (regarding, for example, estimation and assignment of regulatory targets, definition of compliance, and the day-to-day operation of the system) remain to be clarified more fully. We address some of these details in the discussion of further research later in this chapter.

40. More than 300 toxic chemicals are identified in the Toxic Chemical Release Inventory.

41. See the 1989 and 1990 Clean Air Act Amendments for sulfur dioxide trading.

Notice that the targets can be specified as either allowable load per firm (total allowable emissions divided by the total number of relevant firms) or allowable intensity per firm (total allowable emissions divided by the aggregate output of the relevant firms). Notice further that assigning the same pollution intensity target to all firms differentiates the targeted load for each firm. Our proposed incentive system can use either form of target. However, the use of pollution load targets has more intuitive appeal. The information requirements of the two targets are similar; they can be translated into one another using the firm's total production.

42. This assumes that there are no meaningful transactions costs. If transactions costs do exist, the original distribution of targets will affect efficiency. In that case, the original distribution should aim to avoid the transactions costs, for example, by placing higher targets on the least-cost avoiders. See Hartman 1982.

Even if there exist no real transactions costs, the initial setting of targets still has obvious distributional effects for later trading. If,

for example, firms that are believed to be more efficient abaters are given higher abatement targets, they will make less money abating, other things being equal. Other distributional contention will arise if and when targets are differentiated across older and newer technologies and larger and smaller producers.

43. This is merely a form of placing the liability with the least-cost avoider in the presence of transactions costs that limit Coasian trading of liabilities. See Hartman 1982.

Obviously, from a distributional point of view, these firms would argue that they should have the same emissions targets as all other firms and should be rewarded for exceeding them. These (and other) distributional issues will of course be the subject of staggering political debate and rent-seeking behavior. Analogous debate and rent-seeking behavior have been evidenced at the implementation of other incentive regulation schemes (see the section on incentive regulation mechanisms). Methods have been developed to defuse this debate and behavior. Specifically, many incentive regulation systems take the status quo as the point of departure for distributional issues. If the status quo were used as the point of departure for the pollution target system, the first period targets would reflect current pollution less some across-the-board percentage reduction. See USFERC 1989, chapter 3 for more discussion of equity and status quo fairness.

44. Issues involved in identifying control group members for ARAM targets are discussed in the section on traditional methods of regulation.

45. As discussed above (note 41), the targets can be articulated in terms of allowable emissions per firm or allowable emissions per unit of output. However, it is easier to understand and manipulate the system when the targets are expressed as loads. Once a firm knows its allowable load and compares that with its actual load, it can make efficient abatement decisions (figure 11.2). Furthermore, expressing targets as loads is not much more information-intensive than expressing the targets as pollution intensities. If a firm is given its target in terms of intensity, it will most likely translate that intensity target into a targeted load (multiply by a measure of output) in order to understand its abatement requirements.

For the current discussion,  $F$  remains a black box. The estimation of  $F$  will require a research effort that must address efficiency, equity, and information issues (see the agenda for further research, later in this chapter). For example, if all firms are fairly well informed regarding the abatement technologies and abatement capacities of all other firms, and if each firm has an equal inalienable right (the property right rests with the firms) to the use of environmental resources (regardless of its production levels), then equation 11.5 will be simple. It will assign the same constant allowable pollution load for each and every firm.

On the other hand, if firms are poorly informed regarding the abatement technologies and abatement capacities of all other firms, and if each firm has the right to use only those environmental

resources assigned by the regulator (or purchased from other firms), then equation 11.5 will be more complicated. Under these assumptions the target should be denominated in intensity units, and the intensity targets should be differentiated by the abatement costs of each firm. Each firm would translate its intensity target to a load target by multiplying by total production.

Issues faced when estimating a targeting equation such as equation 11.5 are introduced in USFERC 1989, which discusses both regression and engineering approaches.

46. See the discussion of deadbands in the discussion of traditional methods of regulation.

47. One of our reviewers expressed concern about the enormous number of legal battles that could arise with the setting of these targets. We share this concern. However, regulatory and legal fights will arise with any regulatory system. If our incentive regulation system is well designed, the targets will be fairly simple to understand; firm behavior will not be rigidly constrained in the face of expansion and production diversification; and the implementation of the incentive regulation system will be such that participants will participate willingly. All of these conditions have characterized the examples of incentive regulation discussed in the second section of this chapter. In each case, incentive regulation was more flexible and preferred by the regulated firms over the regulatory system that it replaced.

Obviously, the first and foremost task of our future research is to more fully articulate the details of our incentive regulation system, in order to alleviate these concerns. The most important details will include the precise formulation of equation 11.5; the precise delineation of the timing of the assignment of targets and the evaluation of performance; the definition of performance (for example, should it include abatement by the targeted firm alone or abatement purchased from other firms); and the timing of the transition to an emissions bank.

48. All theoretical analyses of the principal-agent problem suggest that rewards and penalties should be shared (see the section on incentive regulation mechanisms). The extent of the sharing depends on the risk-bearing preferences of the principal and agent. The value of the sharing parameter must be examined further, particularly within the context of the optimal incentive clause literature (Laffont and Tirole 1986; Joskow and Schmalensee 1986).

49. The empirical literature on incentive mechanisms provides discussion regarding appropriate lengths of time.

50. See Seneca and Taussig 1984, chapter 10 for the distinction between across-the-board and point-to-point standards.

51. Of course, a budget deficit or surplus may occur, which may require additional political intervention.

52. Specifically, since the most efficient abaters are stimulated to abate, the aggregate of rewards paid to these abaters equals the integral of payments per unit of abatement times the amount abated. See Hartman 1982.

53. Maloney and Yandle (1980) examined the bubble approach

for 52 DuPont plants. They found that the bubble approach meets ambient air regulation quality standards at a 60 percent cost savings relative to traditional source-by-source emission standards.

54. For example, Pennsylvania changed the composition of its highway paving materials in order to reduce hydrocarbon emissions sufficiently to offset those of a new Volkswagen auto assembly plant in New Stanton, Pennsylvania.

55. See Cummings, Brookshire, and Schulze 1986; Freeman 1979, 1982; Mitchell and Carson 1981; Rowe, d'Arge, and Brookshire 1980; and Schulze, d'Arge, and Brookshire 1981. For a summary, see Hartman 1992.

56. See Chao and Wilson 1987; Doane, Hartman, and Woo 1988a, 1988b; Hartman, Doane, and Woo 1991a, 1991b; Munasinghe 1979, 1980; and Munasinghe and Gellerson 1979.

57. Examples of econometric cost estimates are found in Hartman, Wheeler, and Singh 1994.

58. Hartman, Wheeler, and Singh 1994 used four-digit ISIC control groups in their analysis of abatement costs.

59. See note 2.

60. Of course, the extent of the reduction in monitoring costs is an empirical issue.

61. Examples of incentive regulation are becoming extensive in the electric power industry, in the telecommunications industry, and in postal services. See the section on incentive regulation mechanisms and its references.

## References

- Adar, Z., and Griffin, J. 1976. "Uncertainty and Choice in Pollution Control Instruments." *Journal of Environmental Economics and Management* 3(3): 178-88.
- Anderson, D. 1990. "Environmental Policy and the Public Revenue in Developing Countries." Environment Department Working Paper 36. World Bank, Washington, D.C.
- Arizona Corporation Commission. 1987. Docket nos. U-1345-86-062 and U-1345-85-367. Phoenix, Arizona.
- Atkinson, S., and D. Lewis. 1974. "A Cost-Effective Analysis of Alternative Air Quality Control Strategies." *Journal of Environmental Economics and Management*. 1(3):237-50.
- Baron, D. P., and R. DeBondt. 1979. "Fuel Adjustment Mechanisms and Economic Efficiency." *Journal of Industrial Economics* 27(3):243-61.
- Baron, D. P., and R. B. Myerson. 1982. "Regulating a Monopolist with Unknown Costs." *Econometrica* 50(4):911-30.
- Bhattacharyya, S. K., and D. J. Laughhunn. 1987. "Price Cap Regulation: Can We Learn from the British Telecom Experience?" *Public Utilities Fortnightly* 120(8):22-29.
- Bradley, I., and C. Price. 1988. "The Economic Regulation of Private Industries by Price Constraints." *Journal of Industrial Economics* 37(1):99-106.

## REGULATORY POLICIES AND REFORM: A COMPARATIVE PERSPECTIVE

- Bressers, I. 1983. "The Role of Effluent Charges in Dutch Water Quality Policy." In P. Downing and K. Hanf, eds., *International Comparisons in Implementing Pollution Laws*. Boston: Kluwer Nijhoff.
- California Public Utilities Commission. 1991. *Request for Proposal: Evaluation of DSM Shareholder Incentive Mechanisms*. San Francisco, December 12.
- California Public Utilities Commission and California Energy Commission. 1987. *Standard Practice Manual: Economic Analysis of Demand-Side Management Programs*. Sacramento.
- . 1990. *An Energy Efficiency Blueprint for California*. Report of the Statewide Collaborative Process. Sacramento.
- Cass, G., R. Hahn, and R. Noll, eds., 1982. *Implementing Tradable Emissions Permits for Sulfur Oxides Emissions in the South Coast Air Basin: Final Report to the California Air Resources Board*. Pasadena: Environmental Quality Laboratory.
- Chao, H. P., and R. Wilson. 1987. "Priority Service: Pricing, Investment and Market Organization." *American Economic Review* 77(5): 899–916.
- Chung, C.-S., and D. L. Sang. 1992. "Environmental Management in Korea." Paper prepared for the OECD Development Center, Paris, October.
- Clifford, M. 1990. "Kicking up a Stink: South Korean Government Reels from Anti-Pollution Backlash." *Far Eastern Economic Review*, October 18, pp. 72–73.
- Coursey, D. L., J. L. Hovis, and W. D. Schulze. 1987. "The Disparity Between Willingness to Accept and Willingness to Pay Measures of Value." *Quarterly Journal of Economics* 102(3): 679–90.
- Cox, J. C., and R. M. Isaac. 1987. "Mechanisms for Incentive Regulation: Theory and Experiment." *Rand Journal of Economics* 18(3): 348–59.
- Coy, D., and A. Sieminski. 1990. "The Clean Air Act of 1990." Washington Analysis Corporation, New York City, November.
- Crandall, R. 1983. *Controlling Industrial Pollution: The Economics and Politics of Clean Air*. Washington, D.C.: Brookings Institution.
- Crew, M. A., and P. R. Kleindorfer. 1985. "Governance Structures for Natural Monopoly: A Comparative Institutional Assessment." *Journal of Behavioral Economics* 14(Winter): 117–40.
- . 1987. "Productivity Incentives and Rate-of-Return Regulation." In M. A. Crew, ed. *Regulating Utilities in an Era of Deregulation*. New York: St. Martin's Press.
- , eds. 1991. *Competition and Innovation in Postal Services*. Boston: Kluwer Academic Publishers.
- Cross, J. G. 1970. "Incentive Pricing and Utility Regulation." *Quarterly Journal of Economics* 84(2): 236–53.
- Cummings, R. G., D. S. Brookshire, and W. D. Schulze, eds. 1986. *Valuing Environmental Goods: An Assessment of the Contingent Valuation Method*. Totowa, N.J.: Rowman and Allanheld.
- Dales, J. H. 1968. "Land, Water and Ownership." *Canadian Journal of Economics* 1(November): 791–804.
- Delaware Public Service Commission. 1984. Findings and Order no. 2770. Docket no. 84-21. Dover.
- Doane, M. J., R. S. Hartman, and C. K. Woo. 1988a. "Household Preference for Interruptible Rate Options and the Revealed Value of Service Reliability." *Energy Journal—Special Issue 9*: 121–34.
- . 1988b. "An Econometric Analysis of Perceived Value of Service Reliability." *Energy Journal—Special Issue 9*: 135–50.
- Edison Electric Institute. 1984. *Incentive Regulation in the Electric Utility Industry*. EEI Publication SR-84-03. Washington, D.C.
- . 1987. *Incentive Regulation in the Electric Utility Industry*. EEI Publication 04-87-19. Washington, D.C.
- Eheart, J., E. Brill, and R. Lyon. 1983. "Transferable Discharge Permits for Control of BOD: An Overview." In E. Joeres and M. David, eds., *Buying a Better Environment: Cost-Effective Regulation Through Permit Trading*. Madison: University of Wisconsin Press.
- Engineering Science, Inc. (in association with Hyundai Engineering Co., Ltd.). 1984. "Summary Report of the Han River Basin Environmental Master Plan Project." Cape Coral, Florida.
- Estache, A. 1991. "Municipal Environmental Policy Issues in Brazil." World Bank, Washington D.C.
- Face, H. K. 1988. "The First Case Study in Telecommunication Social Contracts." *Public Utilities Fortnightly* 121(9): 27–31.
- Federal Communications Commission. 1987. *Notice of Proposed Rule Making*. CC Docket no. 87-313. Washington, D.C.
- Findley, R. 1988. "Pollution Control in Brazil." *Ecology Law Quarterly* 15(1): 1–68.
- Finsinger, J., and I. Vogelsang. 1981. "Alternative Institutional Frameworks for Price Incentive Mechanisms." *Kyklos* 34 (3): 388–404.
- . 1985. "Strategic Management Behavior Under Reward Structures in a Planned Economy." *Quarterly Journal of Economics* 100(1):263–69.
- Freeman, A. M. III. 1979. *The Benefits of Environmental Improvement Theory and Practice*. Baltimore: Johns Hopkins University Press.
- . 1982. *Air and Water Pollution Control: A Benefit-Cost Assessment*. New York: John Wiley & Sons.
- Gadacz, O. 1986. "House Cleaning, Sort of." *Business Korea* 4(1) : 36–39.
- Gerard, R. 1982. "The Effects of Public Utility Regulation on the Efficiency of a Market for Emissions Permits." In G. Cass, R. Hahn, and R. Noll, eds., *Implementing Tradable Emissions Permits for Sulfur Oxides Emissions in the South Coast Air Basin: Final Report to the California Air Resources Board*. Pasadena: Environmental Quality Laboratory.

- Gonzalez, C. 1981. "Markets in Air: Problems and Prospects of Controlled Trading." *Harvard Environmental Law Review* 5: 377-54.
- Gravelle, H. S. E. 1985. "Reward Structures in a Planned Economy: Some Difficulties." *Quarterly Journal of Economics* 100(1): 272-78.
- Gresser, J. 1979. "Managing Industrial Development with Environmental Management in the Republic of Korea." Report 79-3. World Bank, Urban and Regional Economics Division, Washington, D.C.
- Hagerman, J. 1990. "Regulation by Price Adjustment." *Rand Journal of Economics* 21(1): 72-82.
- Hahn, R. 1984. "Market Power and Transferable Property Rights." *Quarterly Journal of Economics* 99(4): 753-65.
- . 1989. "Economic Prescriptions for Environmental Problems: How the Patient Followed the Doctor's Orders." *Journal of Economic Perspectives* 3(2): 95-114.
- Hahn, R. W., and R. G. Noll. 1982a. "Implementing Tradable Emissions Permits." In L. Gramer and F. Thompson, eds., *Reforming Social Regulation*. Beverly Hills: Sage Publications.
- . 1982b. "Designing a Market for Tradable Emissions Permits." In W. Magat, ed., *Reform of Environmental Regulation*. Cambridge, Mass.: Ballinger Publishing.
- . 1983. "Barriers to Implementing Tradable Air Pollution Permits: Problems of Regulatory Interactions." *Yale Journal on Regulation* 1(1): 63-91.
- Harrison, G. W., and M. McKee. 1985. "Monopoly Behavior, Decentralized Regulation and Contestable Markets: An Experimental Evaluation." *Rand Journal of Economics* 16(1): 51-69.
- Hartman, Raymond S. 1982. "A Note on Externalities and the Placement of Property Rights: An Alternative Formulation to the Standard Pigouvian Results." *International Review of Law and Economics* 2(1): 111-18.
- . 1988. "Self-Selection Bias in the Evaluation of Voluntary Energy Conservation Programs." *Review of Economics and Statistics* 70(3): 448-58.
- . 1992. "Issues in the Valuation and Aggregation of Goods and Services: A Concept Paper." World Bank, Socio-Economic Data Division, International Economics Department, Washington, D.C.
- Hartman, Raymond S., and M. J. Doane. 1986. "Household Discount Rates Revisited." *Energy Journal* 7(1): 139-48.
- Hartman, Raymond S., K. Bozdogan, and R. Nadkarni. 1979. "The Economic Impacts of Environmental Regulations on the U.S. Copper Industry." *Bell Journal of Economics* 10(2): 589-618.
- Hartman, Raymond S., M. J. Doane, and C. K. Woo. 1991a. "Consumer Rationality and the Status Quo." *Quarterly Journal of Economics* 106(1): 141-62.
- . 1991b. "Status Quo Bias in the Measurement of Value of Service." *Resources and Energy* 12 (2): 197-214.
- Hartman, Raymond S., D. R. Wheeler, and M. Singh. 1994. "The Cost of Air Pollution Abatement." World Bank, Washington, D.C.
- Hausman, J. A. 1979. "Individual Discount Rates and the Purchase and Utilization of Energy-Using Durables." *Bell Journal of Economics* 10(1): 33-54.
- Holmström, B. 1979. "Moral Hazard and Observability." *Bell Journal of Economics* 10(1): 74-91.
- Hughes, G. 1990. "Are the Costs of Cleaning Up Eastern Europe Exaggerated? Economic Reform and the Environment." University of Edinburgh, Department of Economics Working Paper, November.
- ICF, Inc. 1981. *Emissions Reduction Banking Manual* (also U.S. EPA Emission Reduction Banking and Trading Publication BG200). Washington, D.C.
- Isaac, R. 1982. "Fuel Cost Adjustment Mechanism and the Regulated Utility Facing Uncertain Fuel Prices." *Bell Journal of Economics* 13(1): 158-69.
- Johnson, E. 1967. "A Study in the Economics of Water Quality Management." *Water Resources Research* 3(1): 291-305.
- Joskow, P. L., and R. Schmalensee. 1986. "Incentive Regulation for Electric Utilities." *Yale Journal of Regulation* 4(1): 1-49.
- Kolstad, C. 1982. *Economic and Regulatory Efficiency*. Report LA-9458-T. Los Alamos: Los Alamos National Laboratory.
- Kopp, Raymond, Paul Portney, and Diane DeWitt. 1990. "International Comparisons of Environmental Regulation." Discussion Paper QE90-22-REV. Resources for the Future, Washington, D.C.
- Krier, P. 1982. "Some Legal Aspects of Tradable Emissions Permits for Air Pollution in Southern California." In Cass, R. Hahn, and R. Noll, eds., *Implementing Tradable Emissions Permits for Sulfur Oxides Emissions in the South Coast Air Basin: Final Report to the California Air Resources Board*. Pasadena: Environmental Quality Laboratory.
- Krishna, K. 1988. "The Case of the Vanishing Revenues: Auction Quotas with Oligopoly." NBER Working Paper 2723. National Bureau of Economic Research. Cambridge, Mass.
- . 1990. "The Case of the Vanishing Revenues: Auction Quotas with Monopoly." *American Economic Review* 80(4): 828-36.
- Kritiporn, P., T. Panayotou, and K. Charnprateep. 1990. *The Greening of Thai Industry: Producing More and Polluting Less*. Prepared for the Thailand Development Research Institute Conference on Industrializing Thailand and Its Impact on the Environment, December 8-9, Bangkok.
- Krupnick, A. 1983. "Costs of Alternative Policies for the Control of NO<sub>2</sub> in the Baltimore Region." Working Paper. Resources for the Future, Washington, D.C.
- Laffont, J. J., and J. Tirole. 1986. "Using Cost Observation to Regulate Firms." *Journal of Political Economy* 94(3): 614-41.

- Landau, W. 1979. "Who Owns the Air? The Emissions Offset Concept and its Implementation." *Environmental Law* 9(3): 575-600.
- Littlechild, S. C. 1983. "Regulation of British Telecommunications' Profitability." Report to the Secretary of State, Department of Industry, London, February.
- Loeb, M., and W. A. Magat. 1979. "A Decentralized Method for Utility Regulation." *Journal of Law and Economics* 22 (2): 399-404.
- Maloney, M. T., and B. Yandle. 1980. "Bubbles and Efficiency." *Regulation* 4(May/June): 49-52.
- Massachusetts Governor's Task Force on the Fuel Adjustment Clause. 1981. *The Fuel Adjustment Clause Question: A Report and Recommendations on Reform of the Fuel Adjustment Clause in Massachusetts*. Interim Report. Boston.
- Mathios, A. D., and R. P. Rogers. 1989. "The Impact of Alternative Forms of State Regulation of AT&T on Direct-Dial, Long-Distance Telephone Rates." *Rand Journal of Economics* 20(3): 437-53.
- Merrifield, J. 1988. "The Impact of Selected Abatement Strategies on Transnational Pollution, the Terms of Trade, and Factor Rewards: A General Equilibrium Approach." *Journal of Environmental Economics and Management* 15(3): 259-84.
- Mishan, E. J. 1990. "Economic and Political Obstacles to Environmental Sanity." *National Westminster Bank Quarterly Review* (May): 25-42.
- Misiolek, W., and Elder, H. 1989. "Exclusionary Manipulation of Markets for Pollution Rights." *Journal of Environmental Economics and Management* 16(2): 156-66.
- Mississippi Public Service Commission. 1986. "Performance Evaluation Plan: Rate Schedule 'PEP.'" PSC Schedule no. 28. Jackson, March 31.
- Mitchell, R. C., and R. T. Carson. 1981. "An Experiment in Determining Willingness to Pay for National Water Quality Improvements." Draft report prepared for the U.S. Environmental Protection Agency. Resources for the Future, Washington, D.C.
- Montgomery, R. 1972. "Markets in Licenses and Efficient Pollution Control Programs." *Journal of Economic Theory* 5(3): 395-418.
- Munasinghe, M. 1979. *The Economics of Power System Reliability*. Baltimore: Johns Hopkins University Press.
- . 1980. "Cost Incurred by Residential Electricity Consumers due to Power Failures." *Journal of Consumer Research* 6: 361-69.
- Munasinghe, M., and M. Gellerson. 1979. "Economic Criteria for Optimizing Power System Reliability Levels." *Bell Journal of Economics* 10(1): 353-65.
- New York Public Service Commission. 1984. Opinion no. 83-17, Case 27741, September 19, 1983; Case 28896, December. New York City.
- . 1986. Opinion no. 85-17(A), Case 28961, May 2. New York City.
- . 1987. Opinion no. 85-17(D), Case 28961, May 11. New York City.
- Noll, R. G. 1983. "The Feasibility of Tradable Emissions Permits in the United States." In J. Finsinger, ed., *Public Sector Economics*. New York: St. Martin's Press.
- O'Connor, D. 1991. "Policy and Entrepreneurial Responses to the Montreal Protocol: Some Evidence from the Dynamic Asian Economies." Technical Paper 51. OECD Development Center, Paris, December.
- . 1993. "Managing the Environment with Rapid Industrialization: Lessons from the East Asian Experience." Draft paper. OECD Development Center, Paris.
- OECD (Organization for Economic Cooperation and Development). 1989. *Economic Instruments for Environmental Protection*. Paris.
- OFTEL (Office of Telecommunications). 1985. "British Telecom's Price Changes, November 1985." Statement issued by the director general of telecommunications, London, December 16.
- . 1986. "Review of British Telecom's Tariff Changes, November 1986." London.
- . 1988. "The Regulation of British Telecom's Prices: A Consultative Document." London.
- Olokesusi, F. 1987. "Characteristics of Environmental Problems in Nigeria and Management Prospects." *The Environmentalist* 7(1): 55.
- . 1988. "An Overview of Pollution in Nigeria and the Impact of Legislated Standards on Its Abatement." *The Environmentalist* 8(1): 31.
- Osaë-Addo, A. 1990. "Nigeria—Environmental Assessment Project." World Bank, Washington, D.C., July 16.
- Pacific Gas and Electric Company. 1985. "Application to Establish a Rate Adjustment Procedure for Its Diablo Canyon Nuclear Power Plant." Application 85-08-025 before the Public Utilities Commission of the State of California. San Francisco.
- Palmisano, P. 1982. "Have Markets for Trading Emission Reduction Credits Failed or Succeeded?" Working Paper 2. U.S. Environmental Protection Agency, Office of Policy and Resource Management, Washington, D.C.
- Pedersen, R. 1981. "Why the Clean Air Act Works Badly." *University of Pennsylvania Law Review* 129: 1059-1109.
- Riordan, M. H. 1984. "On Delegating Price Authority to a Regulated Firm." *Rand Journal of Economics* 15(1): 108-15.
- Riordan, M. H., and D. E. M. Sappington. 1987. "Awarding Monopoly Franchises." *American Economic Review* 77(3): 375-87.
- Roach, F., C. Kolstad, A. Kneese, R. Tobin, and M. Williams. 1981. "Alternative Air Quality Policy Options in the Four Corners Region." *Southwestern Review* 1(2): 29-58.

- Rosewicz, B. 1990. "Price Tag Is Producing Groans Already." *Wall Street Journal*, October 29, p. A7.
- Ross, S. A. 1973. "The Economic Theory of Agency: The Principal's Problem." *Papers and Proceedings of the American Economic Review* 63(2): 134-39.
- Rowe, R. D., R. C. d'Arge, and D. S. Brookshire. 1980. "An Experiment on the Economic Value of Visibility." *Journal of Environmental Economics and Management* 7(1): 1-19.
- San Diego Gas and Electric Company. 1991. "Annual Summary of Demand Side Management Activities." San Diego, March.
- Sappington, D. E. M. 1980. "Strategic Firm Behavior under a Dynamic Regulatory Adjustment Process." *Bell Journal of Economics* 11(2): 360-72.
- . 1982. "Optimal Regulation of Research and Development under Imperfect Information." *Bell Journal of Economics* 13(2): 354-68.
- Schulze, W. D., R. C. d'Arge, and D. Brookshire. 1981. "Valuing Environmental Commodities: Some Recent Experiments." *Land Economics* 57(2): 151-72.
- Seneca, J. J., and M. K. Taussig. 1984. *Environmental Economics*. New York: Prentice Hall.
- Seskin, E., R. Anderson, and F. Reid. 1983. "An Empirical Analysis of Economic Strategies for Controlling Air Pollution." *Journal of Environmental Economics and Management* 10(2): 112-24.
- Shavell, S. 1979. "Risk Sharing and Incentives in the Principal and Agent Relationship." *Bell Journal of Economics* 10(1): 55-73.
- Shleifer, A. 1985. "A Theory of Yardstick Competition." *Rand Journal of Economics* 16(3): 319-27.
- Sibley, D. S. 1989. "Asymmetric Information, Incentives and Price-Cap Regulation." *Rand Journal of Economics* 20(3): 392-404.
- Smith, M. J., and W. Dickter. 1984. "Living with Standards of Performance Programs." *Public Utilities Fortnightly* 114(4): 26-30.
- Song, J. 1989. "Too Much Rubbish." *Korea Business World* 5(10): 24-38.
- South Coast Air Quality Management District. 1981. "Sulfur Dioxide/Sulfate Control Study." Staff Report. El Monte, Calif.
- Southern California Edison Company. 1991. "Filing of 1990/1991 Demand-Side Management (DSM) Annual Report." Application 86-12-047, I.87-01-017 before the Public Utilities Commission of the State of California, San Francisco.
- Southern California Gas Company. 1991. "Demand Side Management Report." Los Angeles, March.
- Spofford, W. 1984. "Efficiency Properties of Alternative Source Control Policies for Meeting Ambient Air Quality Standards: An Empirical Application to the Lower Delaware Valley." Working Paper D-118. Resources for the Future, Washington, D.C., February.
- Tam, M.-Y. S. 1981. "Reward Structures in a Planned Economy: The Problem of Incentives and Efficient Allocation of Resources." *Quarterly Journal of Economics* 96(1): 111-28.
- Tietenberg, T. 1988. *Environmental and Natural Resource Economics*. Chicago: Scott Foresman.
- USAID (U.S. Agency for International Development). 1987. *Natural Resources and Environmental Management in Indonesia: An Overview*. Jakarta. October.
- USCBO (U.S. Congressional Budget Office). 1985. *Environmental Regulation and Economic Efficiency*. Washington, D.C.: U.S. Government Printing Office.
- USEPA (U.S. Environmental Protection Agency). Office of Planning and Management. 1981. *Smarter Regulation*. Washington, D.C.
- USFERC (U.S. Federal Energy Regulatory Commission). Office of Economic Policy. 1989. *Incentive Regulation: A Research Report*. Washington, D.C., November.
- USNRC (U.S. Nuclear Regulatory Commission). 1987. *Incentive Regulation of Nuclear Power Plants by State Public Utility Commissions*. NUREG-1256, 1. Washington, D.C., December.
- Vogel, D. 1986. *National Styles of Regulation: Environmental Policy in Great Britain and the United States*. Ithaca: Cornell University Press.
- Vogelsang, I. 1989. "Price Cap Regulation of Telecommunications Services: A Long-Run Approach." In Michael Crew, ed., *Deregulation and Diversification of Utilities*. Boston: Kluwer Academic Publishers.
- Vogelsang, I., and J. Finsinger. 1979. "A Regulatory Adjustment Process for Optimal Pricing by Multiproduct Monopoly Firms." *Bell Journal of Economics* 10(1): 157-71.
- Weisskopf, Michael. 1990. "Clean Air 'Milestone' Is Sent to President." *Washington Post*, October 28, pp. A1, A16.
- Weitzman, M. L. 1974. "Prices vs. Quantities." *Review of Economic Studies* 41(4): 477-91.
- . 1977. "Is the Price System or Rationing More Effective in Getting a Commodity to Those Who Need It Most?" *Bell Journal of Economics* 8(2): 517-24.
- Wheeler, David R. 1991. "The Economics of Industrial Pollution Control: An International Perspective." Draft report. World Bank, Policy, Planning and External Affairs, Industry Development Division and Environmental Assessments and Programs Division, Washington, D.C., July 26.
- Wilczynski, P. 1990. *Environmental Management in Centrally Planned Non-Market Economies of Eastern Europe*. Working Paper 35. World Bank, Environment Department, July.
- Wynne, B. 1987. *Risk Management and Hazardous Waste: Implementation and the Dialectics of Credibility*. Berlin: Springer-Verlag.

# Regulatory policies and reform in telecommunications

Ioannis N. Kessides

In recent years worldwide demand for better and more varied telecommunications services has increased substantially. A large number of commercial activities—such as banking and international finance, tourism and travel, publishing, commodity exchange, and to a large extent all export-oriented manufacturing—are becoming critically dependent on global information and efficient electronic exchange. In a global information economy characterized by intense competition for new markets, telecommunications is rapidly becoming a vital component of national economic policy. Consequently, the quality of a nation's information infrastructure is increasingly viewed by many as an important determinant of its success in improving its balance of trade and overall economic performance.

The explosion of demand in telecommunications has been stimulated by the precipitous decline in the cost of information transmission and processing that followed the rapid technological change of the last three decades. Breakthroughs in microelectronics, computers, digital microwave, optical memory, and satellite relay have dramatically accelerated the pace of productivity growth in information technology. Merging communications and computer technologies have sparked innovations that are radically transforming the very character of most economic activities. In merchandising, retail transactions at the cash register are integrated with purchasing and inventory management. In manufacturing, worldwide sourcing and production are integrated with inventory and orders. In oil exploration, seismic data from rigs are transmitted and analyzed at a central location, facilitating the efficient management of rigs' activities. In publishing, global networks permit a book or a newspaper article written in one country to be mocked up in a second, typeset by a computer in a third, and then transmitted by satellite for printing anywhere in the world.

Information has become a means for firms to perceive and seize new opportunities and new markets, and to satisfy new needs. Information is vital to corporate survival;

it is critical to an economy's viability. In fact in the last two decades, telecommunications policy in many industrial countries has been formulated in the context of far-reaching global strategies. In Great Britain, for example, a liberalized telecommunications regime was intended to support and augment London's role as an international financial center. In the Netherlands national telecommunications policy was formulated so as to stimulate the development of electronic publishing and to promote Amsterdam as a point of access to Europe for international networks, in direct competition with London. A key objective of Australia's telecommunications policy has been to attract commercial traffic destined for Southeast Asia and to encourage financial services business to locate in Australia. In Japan national pricing structures were adopted to stimulate the growth of sectors with vital links to the information infrastructure.

For most of its history and in most countries, telecommunications has been provided as a user-pay service administered by the central government. In a few exceptional instances, a compromise has been struck between market competition and government ownership through the institution of regulated private utilities; franchised monopolies were created, and then the firms within these monopolies were subjected to detailed regulatory scrutiny.

Both government ownership and the public utility paradigm of governmental regulation have been expressly based on the premise that the telecommunications industry constitutes a natural monopoly. Competition, it was believed, would duplicate investment, raise costs, inflate rates, and compromise the affordability of service to the subscriber public. Competitive rivalry was seen as the equivalent of economic waste. Technology and market structure remain neither static nor fixed, however, and institutions do not endure forever. The telecommunications industry today is undergoing massive worldwide alterations in manufacturing, investment, and market orientation. The industry is experiencing the throes of entry,

rivalry, and diversity. Long regarded as a natural monopoly, telephony today is beset by competition on a multi-industry dimension.

The primary force underlying this competitive drift is technology. Advances in telecommunications technology are facilitating the transfer of larger volumes of information than ever, at faster speeds, with superior accuracy, and at rapidly declining costs. In some countries these changes have led to the introduction of competition in the provision of telecommunications outputs and services that formerly were the exclusive domain of a single source of supply. The emerging evidence suggests that the resulting benefits are substantial: a proliferation of exciting new products and services; more efficient and sophisticated means for transferring information; more efficient use of scarce telecommunications resources; a greater variety of price options; and greater responsiveness to the needs and desires of the consuming public.

The ongoing technological revolution has generated enormous pressures on many countries to modify their public policies toward the telecommunications sector. The traditional telephone systems are increasingly seen as incapable of responding sufficiently to the information challenge. Many industrial nations around the world have already made profound changes in the policy framework, structure, and regulation of their telecommunications sector.

Much can be learned about actual and potential industry structure and performance, and about policies designed to improve performance, from information about market demands for the industry's products and about the productive techniques available to the industry's firms. Indeed, policy decisions regarding the appropriateness of public intervention or the efficiency of multiproduct offerings by a single firm (for example, the desirability of carriers offering both local and long-distance service) should be explicitly based on the underlying characteristics and technological conditions of production. For these reasons, the next section offers a description of the fundamental economic characteristics of the telecommunications industry. The second section provides a general perspective of the policy issues raised by these underlying characteristics. Finally, issues for future research are presented in the concluding section.

### **Economic characteristics of the telecommunications industry**

The telecommunications industry encompasses a large number of highly complex and multilayered activities, including transmission of voice messages, images, and

data; provision of access to networks; research and development; and manufacturing of supplies. The discussion in this section focuses on the technical and economic features of modern telecommunications supply as well as the demand for telecommunications outputs. The analysis is restricted to the service segment of the industry.

#### *Components of a telecommunications system*

In analyzing the economic characteristics of the industry, it is convenient to view telecommunications systems from two related perspectives. First, telecommunications distribution systems can be examined in terms of their physical components; from this vantage the key element of the analysis is the "facilities network" and its underlying characteristics. Alternatively, telecommunications distribution systems can be examined in terms of the services they facilitate; from this perspective a system can be viewed as a traffic-directed network.

*The network.* A telecommunications network is a system of interconnected, possibly disparate facilities designed to carry voice, data, image, and other traffic units between a multiplicity of users and locations. The network is comprised of three physical components: terminals or subscribers' equipment, switching systems, and outside plant.

The location of a customer's telephone is called a telephone station. In view of the large number of telephones, it would be extraordinarily expensive for every telephone to have a direct path to every other one. Consequently, each telephone station is connected via a subscriber line to a local switching center known as the telephone exchange or central office, which connects an appropriate pair of subscribers' lines, as required, for each telephone call. The network of lines connecting the telephone stations to the switching center is called the local network. Each temporary path set up through the network to facilitate a telephone call is called a connection.

It is generally more efficient to divide a large town into separate areas, each served by its own switching center, than to have only one centrally located large switch. In the second case the lines would be very long and the cost of the network excessive. The cost of providing additional switching centers is more than offset by the cost savings resulting from the shorter lines. Lines called junctions or trunks interconnect the switching centers, permitting connections between customers attached to different centers. The network comprising all these lines is called the junction network.

In a very large metropolitan area with a large number of switching centers, it would be uneconomic to provide

trunks between all these switching centers. Instead an additional switching center—which has trunks to all other switching centers and serves solely to make connections between them—is installed. This switching center is called a tandem exchange or tandem office. The switching centers that are connected to subscribers' lines, to distinguish them from the tandem exchanges, are called local exchanges or class-5 offices. Different cities and towns are joined by long-distance circuits known as trunk circuits or toll circuits. These circuits comprise the toll network, and the switching centers that they link together are called trunk exchanges or toll offices.

The national public switched telephone network (PSTN) consists of a hierarchy of networks, each with its own switching center. Telecommunications networks have a very natural vertical structure. Subscribers are linked to local exchanges, which in turn are linked by trunk to local tandem exchanges. Local tandem exchanges are linked by toll circuits to regional and then to national tandem exchanges. The network permits a connection to be made between any pair of telephones in the country. Whenever a route does not exist at a particular level in the hierarchy, the connection is routed through switching centers at a higher level. Normally, more than one path through the PSTN exists between any two stations. If one path is busy, another can be utilized through alternative routing. Finally, the national network is connected by a gateway to the international network, which links the countries of the world.

*Subscribers' apparatus.* The apparatus attached to the network by users include telephones, fax and telex machines, television sets, and computing equipment. Large customers may also have their own switching systems to enable calls to be made both between their extension telephones and between their extensions and the PSTN. These private systems may be either manual switchboards, known as private manual branch exchanges (PMBXs), or automatic systems, known as private automatic branch exchanges (PABXs). The private switching systems illustrate why the distinction between network and apparatus is not always clear. A PABX is both terminal equipment and a part of the overall network.

*Switching systems.* Originally, exchanges were manually operated. Operators connected lines as subscribers requested. To make the desired connection, the operator searched over the switchboard for the appropriate jack into which to plug a cord. Beginning in the 1910s manual operation was increasingly replaced by automatic electromechanical switching technology (Strowger switches). In the Strowger system the metal fingers of an electro-

mechanical selector moved over a bank of contacts to reach the required outlet. The subscriber's action of requesting a number from the operator was replaced by means of a dial on the telephone. Digital electronic switching systems began replacing the electromechanical systems in the mid-1970s. In addition to their substantially higher speeds, these electronic systems enjoy another important advantage: They require much less engineering maintenance than the older electromechanical systems, thus enabling the network to expand without creating excessive manpower requirements.

*Outside plant.* The outside plant comprises all physical components of the network that are located between terminal stations and switching centers and between switching centers. Originally, open-wire lines on poles provided connections among the components of the network. These were replaced by cables (overhead or underground), varying in size from two to a few thousand pairs. For long-distance transmission, coaxial cables replaced the simple copper wires. Advances in technology led to the rapid replacement of coaxial cables by optical fibers, radio, and microwave transmission (both terrestrial and satellite-based). To ensure high-quality speech transmission over long distances, the outside plant is supplemented by electronic equipment such as amplifiers; these constitute transmission systems.

*Service definitions and boundaries.* Service definitions and "boundary" lines are of critical importance for telecommunications policy because they often determine which entities (private or public sector) can offer which services and on what terms and conditions (regulated or unregulated). They therefore have a profound influence on the structure of the telecommunications sector and on the economic and regulatory relationships between this sector and others that depend heavily on it. These include data processing, banking, electronic publishing, and electronic marketing services.

A complex combination of technological and economic forces has led to the introduction of a wide array of new telecommunications services and precipitated significant changes in the domestic and international telecommunications environment. These forces are exerting strong pressures on suppliers and policymakers alike to revisit traditional service definitions and the way industry is organized. More specifically, the proliferation of computer devices, and the convergence of communications and computer technologies, have created both the need and the opportunity for protocol conversion or computer-controlled networks linking computers with incompatible standards to electronic markets for securities,

commodities, foreign currency, and other financial transactions. Such services can be offered by entities unaffiliated with the transmission carrier (which in many countries is a single entity) and provided over leased lines that interconnect nodes of computers. Policymakers must delineate which services can be supplied independent of the traditional carrier, and those which the carrier must offer to new service providers.

The process of drawing the definitional lines of demarcation among the telecommunications services is based on widely divergent criteria and rationales. A cross-country comparison reveals at least three major approaches. The first focuses on technological factors; it seeks to distinguish among services on the basis of their underlying technological characteristics (for example, formats, protocols, and content of information). The second approach is based on the economic relationship between a specific service and the existing offerings of the monopoly provider (assuming the presence of a primary service supplier, as is the case in most countries). Finally, boundaries also may be drawn on the basis of legal and institutional considerations.

*Basic versus enhanced services.* One key dichotomy that has been the focus of regulatory attention in many countries is the distinction between basic and enhanced, or value-added, services. In the case of basic service, a pure transmission quality is offered over a communication path that is virtually transparent in terms of its interaction with customer-supplied information. The standard voice telephony (both fixed and mobile) is the prime example of such service. Enhanced service, by contrast, combines transmission with computer processing that either (a) acts on the format, content, code, protocol, or similar aspects of the transmitted information; (b) provides the subscriber with additional, different, or restructured information; or (c) involves subscriber interaction with stored information.

In addition to the basic-versus-enhanced dichotomy, two other definitional distinctions are now evolving in global telecommunications policymaking: (a) enhanced services versus informational and transactional services, and (b) facilities versus services.

*Enhanced versus informational and transactional services.* Informational and transactional services are based on the sale and dissemination of information and data-processing services to users. Because they also afford the users the capability of executing transactions, they are rapidly becoming interwoven with conventional banking, securities, and commodities businesses. The vendors of informational services offer a distinct electronic service and

hence are dependent on telecommunications for the conduct of their business. However, unlike many enhanced service suppliers, they generally do not compete directly with the providers of telecommunications services.

From the public policy perspective, the distinction between enhanced and informational services is of critical importance. Although there may be some similarity between these offerings, informational and transactional services have a distinctly different function than enhanced services: They market information or specific brokerage services through electronic means. To the extent that these services provide some communications capability to their customers, that capability is almost entirely incidental to their primary business purpose.

Given these basic differences, an optimal public policy would call for a separate regulatory classification for informational and transactional services; these services should not become entangled in policies and regulatory processes of prior approval that have been formulated for the communications-oriented enhanced services. Enhanced services, even if less stringently regulated than basic services, are still subject to a more structured scheme of regulation than traditionally has been the case with informational and transactional services.

*Facilities versus services.* The distinction between the provision of facilities and that of services is the subject of attention in almost every country that has focused on telecommunications policy. At one end of the policy spectrum, unrestricted entry in both facilities and services is permitted. At the other extreme, monopoly is retained in the provision of facilities, with some flexibility with respect to the offering of certain kinds of services on a competitive basis. Between these two approaches, there are numerous variations involving competition in either or both facilities and services.

#### *Demand for telecommunications services*

The telecommunications sector is part of a broader communications industry that includes the postal service, express freight carriers, and portions of the transportation sector. Telecommunications also can be viewed as a component of the information-processing industry. In this discussion we adopt a narrow definition of the industry by focusing on the provision of basic telephone service to consumers and businesses (which accounts for the bulk of the industry's activities), while we abstract from issues related to the provision of private lines or private networks to business users.

The demand for telecommunications services has a number of distinguishing characteristics (see Sharkey

1982, chapter 9). First, it has the strong periodic element of the peak-load model. Demand is much larger during the mornings of business days; remains heavy, though less so, during the afternoons of those days; and is lightest in the evenings, at night, and during weekends.

Second, demand has an important random element. The desire of a user to make telephone calls to some extent depends on random events. However, seasonal variation—a very important characteristic of demand in other public utilities—is not so marked in telecommunications. In most instances, the desire to communicate with others must be satisfied at the moment the demand is expressed—service must be supplied instantaneously at the lifting of a telephone instrument.

Third, telecommunications outputs cannot be stored, and capacity cannot be readily expanded or contracted (asynchronous communication modes such as facsimile, voice-mail, e-mail, and others are notable exceptions). Consequently, the variability in demand imposes on service providers the burden of maintaining capacity sufficient to meet the peak levels of expected demand. In off-peak periods there is necessarily a level of excess capacity. A low load factor (ratio between the average demand over a period and the peak demand) would be very costly in this industry in view of the high fixed costs underlying its technology. Thus, capacity in the telecommunications industry is necessarily related to peak demand rather than average demand.

Finally, a subscriber's demand for a communications service includes the demand for potential communications with every other user of the service. A communications service must therefore be supplied by means of a network in which information can be transmitted in both directions between any two subscribers.

These demand characteristics have important implications for the structure of supply of telecommunications outputs and in particular for the existence of natural monopoly. The wide divergence between peak and off-peak demand does not in itself lead directly to economies of scale or natural monopoly. As far as this factor alone is concerned, the requisite capacity could be provided just as efficiently by a large number of suppliers, with an average load factor corresponding to that of the single supplier. However, the periodic nature of demand does lead to natural monopoly. The capacity of each component in the network of a given supplier must be designed so as to handle the maximum demand expected for that component. If demand is fragmented among several suppliers, it is unlikely that the periodic profile of each supplier would coincide with the original profile of demand. But

if multiple suppliers face different peak periods, then it necessarily follows that their combined capacities would exceed the capacity of a single supplier.

Consider for example the case of a single supplier facing a demand profile with two peak periods: a business demand peak during the daytime and a residential peak in the early evening. If demand is fragmented so that business and residential subscribers are served by different suppliers, then their combined capacities substantially exceed the capacity of a single supplier. The excess capacity in the system is an economic waste, and the cost-minimizing structure would therefore call for a single supplier (natural monopoly).

In addition to its predictable periodic element, demand for telecommunications includes an important stochastic component. Stochastic demand creates various complications that typically are not considered in the deterministic case. The capacity of a telephone system is engineered to meet demand with the expectation that some calls will be blocked during certain busy times. The probability of a call not being completed because one or more of the relevant lines are busy, called the blocking probability, is an important aspect of the grade or quality of service. Blocking is determined by the amount of switching and transmission capacity installed in the network. More capacity implies fewer blocked calls and a higher grade of service.

The need for probabilistic engineering gives rise to significant economies of scale (see Waverman 1975). A useful measure of output (capacity) of a telecommunications system is the number of call seconds of message time available, whereas circuits are a measure of capital inputs. In a typical communications system, to ensure a probability of 0.99 that a call will be completed as dialed, the capacity of a single circuit is 40 call seconds per hour. If we assume that demand is characterized by Poisson arrival (that is, the probability of incoming calls arriving at the same time can be approximated by a Poisson distribution), the addition of a second circuit increases the capacity of the plant to 540 call seconds—an increase of more than 1,000 percent. As the number of circuits increases, the capacity of the system, and hence the probability that a caller will find an open line, also increase.

Returns to scale are more pronounced at higher grades of service, and economies of scale decline as the number of circuits increases (for example, doubling the number of circuits from 64 to 128 increases usable capacity by 120 percent, whereas doubling the number of circuits from one to two leads to an increase of more than 1,000 percent in usable capacity). Thus network fragmentation is

likely to impose substantial cost penalties at small levels of demand because a larger number of circuits would be required to maintain the same grade of service.

*Network, call, and congestion externalities*

The benefit that a subscriber derives from a communications service increases as others join the system.<sup>1</sup> This is the classic case of external economies in consumption—positive externalities are associated with the decision of an individual or a household to purchase access to the telephone network. These network externalities reflect the total benefits that accrue to other subscribers (businesses and individuals) because of their ability to communicate with the new subscriber. These benefits are spread very diffusely among other subscribers to the network, and most residential subscribers (unlike businesses) receive no direct compensation for the benefits experienced by other subscribers as a result of their purchase of network access. The total of these uncompensated benefits is the residential network externality.

This positive externality has fundamental policy implications. First, the principle that residential access prices warrant financial support is widely recognized. Since private consumption confers benefits on others, the positive externality warrants that the price be supported below production cost by the amount of the gain. It should be noted, however, that the presence of the subscriber externality calls for the underpricing of local access, not local usage. Second, it renders interconnection among separate networks very desirable.

In many instances telephone operators have been able to maintain the price of residential access below the marginal production cost of network access by supporting these costs with net revenues derived from business customers, toll services, and other premium offerings. Thus the operating entities themselves may partially internalize this externality if subsidizing access raises their profit as well as consumer welfare. In rural areas, by contrast, telephone operators may not be able to support residential access prices for two reasons. First, the geographic scarcity of rural subscribers renders the cost of supplying them with network access much larger than the costs faced by operators in urban areas. Second, rural operators face much lower revenues derived from services to business customers.

A second form of externality is associated with the use of a communications system. Communication is inherently a two-party or multiparty process, yet only one party (the one initiating the call) is charged. In general the recipient also benefits from the call, that is, call external-

ities are positive. However, the tentative consensus seems to be that the call externality is not nearly as important as that associated with access. Users who frequently communicate with one another tend to share costs cooperatively over time by sharing in the placing of calls.

An important negative externality between users arises from the possibility of congestion. A subscriber's attempt to make a call may be frustrated if the relevant lines are busy (blocked) due to the demands of other users. The cost is the wasted effort and the lost benefit of making the call at the desired time. This negative externality can be reduced by congestion pricing, that is, differentiated by peak and off-peak periods.

*Natural monopoly elements in local-exchange service*

It is generally conceded that the provision of local telephone service is a natural monopoly. The natural monopoly characteristics of local-exchange service can be attributed to four general sources.<sup>2</sup>

*Economies of scale and scope in the physical provision of basic services.* For individual network components, unit costs (that is, total costs per unit of traffic-handling capability) typically fall over a large range as rated capacity increases. The local distribution cable represents a clear example of this tendency to decreasing unit costs. A single wire pair is generally sufficient to fully handle the traffic requirements of individual subscribers. The needs of larger users can be met by increasing the number of copper pairs or by using a higher-capacity medium (for example, coaxial cable) while still relying on a sole duct and termination equipment. Having competing connections to each location and competing networks of switches within a single neighborhood, which account for a large share of the local exchange's embedded costs, would normally represent a costly and inefficient duplication of these facilities. It is therefore efficient to have a single supplier provide all the local cable connections within a particular neighborhood. In addition economies of scale in switch construction, operation, and maintenance imply that the least-cost way of serving local neighborhoods is through a single switch that meets a minimum efficient size criterion.

*Economies of scale in network planning and management.*

Economies of scale in network design and management arise primarily from the unique characteristics of telecommunications demand previously identified. A large network would normally be able to handle randomly varying demands more efficiently than a small one. The greater the number of users, the more likely it is that traffic will be evenly distributed over time, thus minimizing the eco-

conomic waste inherent in the excess of capacity over off-peak demand. Relative to several competing small networks, a single large network also could secure economies of network management and coordination (that is, efficiencies in managing the day-to-day flow of traffic over the installed facilities). Larger networks generally have more alternative routes between any two points and can achieve peak-load efficiencies by allocating demands among these routes. Finally, to the extent that network management entails certain fixed costs (that is, costs that do not vary with the size of the network above a certain threshold level), a single large entity could manage network resources more cheaply by spreading these costs over a larger user base.

*Network externalities.* The total gross benefit that society receives from the connection of a subscriber to the network is equal to the sum of values received by all the called and calling parties that use the connection. This benefit consists of two parts. The first part is received by the subscriber; the second corresponds to the external benefits that are received, in total, by all others who can communicate by telephone with the subscriber. If there are costs to transferring and monitoring calls between different company networks, then consumers clearly have incentives to join the network with the greatest number of existing members. Large networks, which minimize inter-network transfers, are clearly more efficient than small ones. Moreover, because the second component of the total benefit increases as the total number of existing network members increases, large networks have greater incentives to recruit new subscribers than do small ones.

*Advantages in raising capital.* Scale economies may also be realized in other parts of a large network's operations. An industry consisting of many small firms might be prone to unstable fluctuations and price wars and hence would not generally look attractive to potential investors. A single large network is likely to be able to attract additional capital at a lower cost and in larger quantities than smaller competing networks.

#### *Natural monopoly elements in interexchange service*

In the analysis of interexchange markets, four types of scale economies are generally considered (see Waverman 1989).

*Plant economies of scale.* Plant scale economies reflect the advantages of size inherent in transmitting messages from one point to another and are analogous to those realized at the plant level in the traditional process industries. Some of the individual components of the interexchange network display substantial economies of scale. The costs

per circuit-mile for coaxial cable, one of the two main transmission media in use today, fall sharply as capacity increases. Scale economies for optical fiber technologies are even greater than they are for coaxial cable. However, the investment costs in some of the other components of the typical terrestrial system entail negligible scale effects. To obtain an accurate assessment of the overall economies of scale in long-distance service, therefore, one must examine all the cost elements of the system.

There are three basic components of investment costs in a long-distance transmission system: (1) real estate and structures, including cable rights of way, buildings, land, and access roads; (2) outside equipment such as cable, microwave towers, and antennas; and (3) radio equipment, including transmitters, receivers, repeaters, and backup equipment. These three components comprise the "basic transmission" costs and are almost invariably included in all the reported studies of long-distance transmission investment costs. Two additional cost components, multiplex and switching equipment, are frequently omitted from these studies. Multiplex equipment imposes many individual telephone conversations on a single transmission.

Different types of terrestrial transmission systems (wire pairs, coaxial cable, microwave radio) display an overall downward trend in investment cost per circuit mile as capacity increases (several econometric studies confirm the presence of such economies). The costs of constructing a building needed to maintain the equipment do not, over broad ranges, increase in proportion to volume. In addition real estate costs exhibit certain indivisibilities (for example, one access road would be required irrespective of the capacity of the station). The property and radio costs are approximately constant up to about 1,000 circuits (see Brock 1981, p. 199). Scale economies in multiplexing are minor; multiplex costs vary in proportion to the system's capacity. As in multiplexing, economies of scale in digital switching systems are relatively minor.

In a typical system, basic transmission costs account for approximately 50 percent of the total investment requirements. Inasmuch as economies of scale are negligible in the other components of the long-distance system, the overall returns to scale in interexchange service should be significantly lower than those found in basic transmission. At low levels of capacity, total costs are dominated by property and radio, whereas at high levels they are dominated by multiplex equipment (for both terrestrial and satellite transmission). Given that multiplex costs are almost linearly related to capacity, scale

economies should diminish significantly as capacity increases. Indeed, various studies of microwave transmission confirm the presence of significant scale economies up to about 250 circuits, moderate economies between 250 and 1,000 circuits, and insignificant economies at levels above 1,000 circuits.

*System economies of scale.* Economies of scale at the system level are analogous to economies of the firm—they reflect the extent to which multiplant operation confers economies above and beyond those associated with operating a single plant of optimal scale. A large, multiplant telephone company may enjoy advantages due to: (a) alternative routing and network management; (b) the interdependence of investment and capacity decisions; (c) integration and common standards; (d) administrative and accounting procedures; and (e) design.

A large single supplier can reroute traffic flows with greater flexibility. In theory the cost efficiencies of routing flexibility could become available to competing firms if they agreed on provisions for accommodating traffic overflows. In practice, however, the cost of negotiating and enforcing these contracts is likely to severely limit their usefulness. Given the very large number of possible paths between terminal points (even in relatively small networks), a contractual relation for selecting alternative paths would certainly entail prolonged and costly negotiations.

When scale economies can be realized by expanding capacity in large, indivisible chunks, the carrying costs of excess capacity can be reduced and the opportunities for scale economies exploited more fully through coordinated investment, rather than with autarkic expansion by each separate plant. Also, in an integrated system with a single decisionmaking unit, all costs associated with revenue division and with the recording, accounting, and transaction of intercompany payments are reduced substantially. If segments of the network were owned and operated by different companies, identical standards would have to be accepted by all companies involved so that the systems could integrate with one another; without such an agreement the design of interfaces and equipment would be more difficult and costly.

*Economies of scope.* The telecommunications industry encompasses a number of different product offerings. Its output is readily divided at least into local and long-distance service and thus constitutes a classic multiproduct case. In addition to economies deriving from the size or scale of a firm's operations, cost savings may also result from simultaneous production of several different outputs in a single enterprise, as contrasted with their pro-

duction in isolation, each by its own specialized firm. That is, there may exist economies resulting from the scope of the firm's operations. Cost savings may result if one telephone company provides both local and long-distance service.

Two major factors underlie economies of scope in telecommunications: complementarity of equipment and transactions costs. As an example, consider the provision of an enhanced service. In principle, the functions required to provide such a service could be located at several points in the network: at the customer's premises, in dedicated facilities interconnected to the network, or in the central office itself. In many instances, locating the functions in the central office reduces the need for interface equipment and minimizes installation and maintenance costs. More generally, economies of scope can be obtained by standardizing the interfaces and equipment for complementary services.

*Contract economies of scale.* Contracting economies reflect the reduction in costs and risks that may result by organizing activities within the firm. There is a tendency for local-exchange monopolies and firms offering interexchange service to join together to avoid the costs of negotiation of the many contracts that make up the system and to avoid being charged the monopoly prices on some links. Internal organization attenuates the aggressive advocacy that epitomizes arm's length bargaining. Perhaps the most distinctive advantage of the vertically integrated firm, however, is the wider variety and greater sensitivity of control instruments that are available for enforcing intrafirm (in comparison with interfirm) activities. Internal rules are more efficient than interfirm contracts.

#### *The impact of changing technologies*

The telecommunications industry has undergone a real revolution in recent years. The sector has exhibited a high, continuous rate of productivity increase benefiting from rapid innovation in electronics, computers, materials, and processes. Conventional wisdom holds that changing technologies will ultimately undermine each of the monopoly forces within the industry. These new technologies involve both switching and transmission.

*Technological advance and costs.* The technological change in the telecommunications industry has been a part of the general electronic revolution. Many industry observers contend that a "convergence" of computers and communications systems—resulting in technological spillovers and economies of scope—is dramatically altering the economic characteristics of both industries.

However, casual empiricism suggests that, overall, quality-adjusted hardware costs dropped much less rapidly for communications systems than for computers. And within communications systems, it was transmission rather than switching costs that fell the most.

Cost reductions have been particularly impressive in the long-distance and traffic-sensitive portions of the market. Microwave technology combined with satellites, high-capacity fiber-optic cables, and improved multiplexing have vastly increased capacity and greatly reduced the cost of providing service. The cost of fiber-optic cables declined from \$10 per meter in 1975 to \$1.75 in 1980 and further to \$.60 per meter by 1985. Satellite earth stations fell in cost from \$2 million each in 1965 to \$30,000 in 1981 and to \$5,000 by 1986. The cost of providing long-distance channels decreased from about \$33 per circuit mile in the late 1950s to less than \$4 in the late 1970s (see Bolter, McConnaughey, and Kelsey 1990, chapter 5).

In switching, digital computer-driven equipment has reduced maintenance costs, and software innovation has expanded substantially the range of services. Within the network, these advances have allowed carriers to supply customers with specialized private networks composed of shared facilities under software control. Switching also has moved closer to the final user, with versatile private switches (PBXs) competing with local-exchange carrier switches to supply a wide range of office communications services.

Where usage is not concentrated, however, technological change has not had nearly the same impact on costs. The non-traffic-sensitive and customer-specific loop that connects every subscriber to the central office has not experienced anywhere near the technological change that has occurred in the long-distance and other traffic-sensitive portions of the industry. For low-volume nodes, the copper cable pairs continue to represent the least-cost technology. However, fiber-optic distribution and microwave bypass are becoming economically viable in large office buildings.

Advances in computer software and data storage technologies are facilitating significant process innovations throughout telecommunications. Long-distance connections are now established and calls routed dynamically, according to the availability of network links and switches. Furthermore, new low-power radio technologies promise to link subscribers via lightweight, vest-pocket telephones. Also, broadband networks, utilizing high-speed switching and fiber-optic links to the consumer, may ultimately provide a broad array of video, data, and personal communications services.

*Natural monopoly.* As noted above, the impact of technological change has been massive but uneven. Cost reductions have been particularly notable in long-distance transmission. Access technologies have advanced less rapidly. Although technological change has driven the entire industry, it has undermined natural monopoly in interexchange markets much more rapidly than in local-exchange service.

In local-exchange service, digital technology has been rapidly supplanting analog applications in switching. Digital switches, whether PBXs or central-office switches, are now available in an almost continuous range of sizes, from 20 to more than 10,000 lines. Systems for load balancing, billing, number changes, and other housekeeping functions are now available in automated form in most of these switches. Furthermore, the development of remote maintenance and housekeeping technologies implies that even relatively small firms can capture economies of scale in these overhead activities by centralizing them. As a result economies of scale in overhead activities are now probably small and certainly declining.

New transmission technologies at the local-exchange level include cable-based telephone access, cellular radio, and direct microwave links to local or long-distance switching nodes. Cable-based access exhibits economies of scale similar to those of wire-based access. Although cellular radio permits a more efficient use of the spectrum relative to existing mobile offerings, the problems of allocating scarce radio frequency still limit the number of allowed carriers (currently two per area). Finally, microwave links typically entail large fixed costs and large carrying capacities and are thus limited to relatively high-volume transmission routes. As such, they do not offer a viable option for linking individual users to the basic telephone network. Overall, despite the development of new transmission technologies and the concomitant increase in the number of potential competitors to the local-exchange carriers, economies of scale in local transmission and distribution have not been eliminated.

The impact of new technologies on network planning and management economies is more ambiguous. Advances in computer technology are likely to reduce the cost of planning and managing network resources. To the extent that the cost of such overhead activities is fixed, the relevant economies of scale are reduced. But greater computing power may provide more scope for active network management and increase the advantages enjoyed by the competitor with the most extensive network.

Similar countervailing forces may be at work regarding network externalities. New technologies are likely to

increase the ease with which calls can be monitored and transferred between networks. The disadvantage to a customer of joining a small network therefore may be offset by the ease of communicating with other networks. However, innovation is likely to appear first in large networks, where the external benefits of adopting the new technologies are large. Thus more rapid innovation may reinforce the advantages of large networks. The important point is that the impact of new technologies in reducing natural monopoly elements in telecommunications is likely to be far smaller in the areas of network externalities and network management economies than in the areas of basic service provision.

Rapid technological change is likely to accentuate the advantages of large networks in raising capital, and hence in that limited context it might actually reinforce the argument for natural monopoly in telecommunications. A generation ago the industry enjoyed a stable environment characterized by an orderly rate of technological change, narrow product substitutes, standard and compatible equipment, identifiable product and service boundaries, long and predictable product economic life, limited number of suppliers, and stable pricing arrangements. Today depreciation life is contracting, rate base accounting is being assaulted by accelerating technology, investment alternatives are proliferating, service boundaries are coalescing, and allocation is buffeted by market diversity and market segmentation. Technological change may therefore effectively increase the risk perceived by investors because (a) product innovations face an inherently uncertain reception from consumers; (b) stable pricing arrangements are no longer feasible in the face of rapid change; (c) a rapidly changing environment increases the likelihood that investments will be made in wrong technologies or generation of equipment; and (d) the proliferation of suppliers raises the risks of nonstandard and incompatible equipment. As a result of this instability, the difficulty of raising capital on acceptable terms is likely to increase and the advantages of having a single, stable service provider may become greater.

In summary, local-exchange telecommunications retains many of the characteristics of natural monopoly, even after the modifying impact of technological change is fully accounted for. Nevertheless, the emergence of competing access media has greatly increased potential entry or competition for the local-exchange market and is likely to effectively discipline incumbent behavior in this market as the incremental costs for these different media (of providing local-exchange access) converge over time. Furthermore, the dramatic shifts in the switching and

trunking technology are likely to significantly alter the way in which local services are provided; there will be an increasing substitution away from the relatively expensive service—access lines—and toward the relatively cheap service—switching and trunking. To the extent that remote switching and trunking are becoming viable substitutes for access lines, market power in the local-exchange market is likely to diminish substantially. Finally, as noted above, some of the most dramatic manifestations of the technological revolution in telecommunications have been in the traffic-sensitive portions of the business. Relatively high-volume routes within densely populated areas (for example, downtown business districts, industrial parks, large apartment buildings) now can be efficiently supplied by an entity that does not necessarily provide service in the intermediately surrounding geographic area.

The economies of scale and scope in basic service provision that still characterize local-exchange service in many areas appear to be far less important and perhaps nonexistent in the long-distance market. To begin with, basic transmission is the only component of the long-distance terrestrial system that exhibits any significant economies of scale. When the other cost components of the system (for example, multiplexing and switching) are also included, overall economies of scale appear to be substantially smaller. Microwave is conducive to multi-firm transmission facilities over many routes (economies of scale in microwave transmission are insignificant at levels above 1,000 voice-grade circuits). In addition, satellites present entirely different operating characteristics relative to cable technology. Under certain operating conditions and system configurations, satellite communications can make entry economically attractive on small scales between particular points. Fiber-optic technology, by contrast, exhibits significant economies of scale over almost the entire range of output. However, as noted above, technological change has led to impressive reductions in the cost of fiber-optic cables. As a result, the portion of the overall costs of interexchange service that is attributable to transmission has declined. Since the other components of the system (including maintenance and other housekeeping functions) exhibit insignificant scale effects, overall economies of scale are likely to be small.

### **Current issues in telecommunications policy**

Now that some of the underlying characteristics of the telecommunications industry have been explored, a number of the policy issues to which they give rise will be discussed. First, the case for deregulation is explored, with

a view to identifying appropriate regulatory reform strategies that recognize the potential market power of current local-exchange service and the largely competitive character of interexchange markets. The discussion then turns to a review of the alternatives to traditional cost-based and residual pricing regulation, which policymakers have begun to consider as a means of encouraging monopoly efficiency, stimulating technological innovation, protecting consumers, and reducing administrative costs. Structural issues are then considered: whether local and long-distance networks should be separated and what the policy should be toward entry and competition (actual or potential) in network operations.

Some of these issues are clearly related. A decision to permit entry into a market, for example, greatly constrains the price-setting abilities of the incumbent firms (and of the regulators), and pricing policies in turn influence the entry decisions of potential competitors. Also, price structures involving cross-subsidy are likely to be undermined if entry is allowed.

#### *Alternative strategies for regulatory reform*

The world is fast approaching a new Information Age in which a significant portion of productive global resources will be directed to collecting, analyzing, transmitting, and reporting information. Only sound regulatory policies that optimize efficient use of a country's telecommunications resources will allow continued innovation (especially by the private sector) and ensure that society does not suffer an unnecessary reduction in its economic welfare. Regulatory policies must ensure that society does not have to invest more resources than necessary in its telephone network and that investment in the network is allocated so as to maximize benefits.

*The case for deregulation.* In telecommunications a primary public interest goal is to control monopoly power and protect the consumer. For much of this century and in most countries, fairness and efficiency in telecommunications have been sought through the public utility form of governmental regulation. This paradigm has been expressly premised on the assumption that the industry constitutes a natural monopoly, and thus achieving efficiency requires that a private or public company be granted a monopoly for the provision of telecommunications service. The government-bestowed monopoly, however, creates strong incentives for overpricing and reduced output and can be used to leverage other markets through cross-subsidization and improper cost allocation between regulated and unregulated activities.

The regulatory framework is intended to guide the

market to a socially efficient solution by ensuring that the monopoly attains a cost-efficient scale while setting prices in a manner that does not abuse its market power. Thus, to prevent the reduced output of monopoly services, the public utility model strictly controls entry and exit, regulates price and the conditions of service, and imposes ubiquitous service obligations. To prevent the use of monopoly power granted by fiat to leverage other markets, this paradigm also controls the prices of regulated services and severely restricts the utility's participation in competitive activities.

Critics of monopolistic provision question the efficiency of this regulatory framework. They argue that government intervention in telecommunications through regulation and ownership has imposed significant direct and indirect (or opportunity) costs on society. The public utility paradigm allegedly exacts significant efficiency costs in resource allocation by (a) distorting prices with the use of average prices for groups of services rather than individual prices for the services in each group, shifting costs to future periods by using uneconomically slow depreciation rates, and cross-subsidizing local with long-distance service and residential with business customers; (b) distorting investment decisions and limiting private incentive to innovate with new technology; and worse, (c) affirmatively discouraging innovation that would render obsolete large quantities of embedded equipment included in the rate base.

Regulation also tends to deter price competition, provides only limited incentives to cut costs or increase managerial efficiency, and generally limits the choices available to consumers. Regulatory price ceilings prevent the supply of higher-quality offerings, while regulatory price floors discourage the supply of lower-quality, inexpensive options that many consumers would find attractive. Furthermore, regulation tends to react much more slowly than the marketplace to changing technological conditions and frequently limits the ability of market participants to respond quickly to changes in demand and supply. In addition regulation entails direct and indirect administrative costs. The experience of interexchange markets reveals that firms and other interested parties may be prepared to expend considerable resources as intervenors in the regulatory process. Finally, regulatory rate making not only leads to significant administrative costs but also is subject to serious practical difficulties, making terribly elusive the goal of keeping prices close to costs.

As discussed in the previous section, the interexchange markets are structurally competitive. The natural monop-

oly characteristics of long-distance services have been significantly modified by technological change. Local-exchange service, by contrast, continues to be a natural monopoly due to increasing returns to scale in network design and management, despite the development of new transmission and switching technologies. However, changing technologies are slowly undermining each of the natural monopoly forces, whereas recent economic and technological developments have greatly increased the cost of continued regulation. An optimal regulatory strategy in telecommunications should therefore seek to isolate the segments of the market still considered to involve technological natural monopoly—local telephone service—from segments that can no longer be taken to constitute natural monopolies, such as long-distance services and the provision of terminal equipment.

The network externality is the other form of market failure that prompted regulatory intervention in telecommunications. As noted, it is most significant at low levels of household telephone penetration, and in several countries it has been extremely important in the development of the system. If deregulation occurs, this form of market failure is likely to return and hence some other means of control must be identified.

*Promoting regulatory reform.* The primary objective of regulatory reform is to ensure that services are provided on a basis that is consistent with the goals of economic efficiency and social equity. Reforming the telecommunications industry can be pursued entirely within the context of monopolistic service provision.

The precise characteristics of regulatory reform obviously depend on the specific circumstances of individual countries—in particular, their regulatory background and their administrative and political institutions. In general, *regulatory reform aims at establishing a more arm's length relationship between the government and the service providers so as to ensure that they have the flexibility needed in a rapidly changing market environment.*

The traditional model for the telecommunications industry entails a single, state-owned service provider that has a monopoly on all aspects of the market and is solely responsible for all operational decisions. Under this model there are few, if any, competitive market elements and no real requirement for a separate regulatory function. Rate designs normally reflect social, rather than efficiency, objectives, subsidies are usually pervasive and well hidden, and telecommunications services often subsidize other sectors of the economy.

Perhaps the single most important element of any telecommunications regulatory reform strategy is the sep-

aration of operational activities from government oversight and regulatory activities. Such separation is necessary to ensure fair and impartial policy development, to insulate the telecommunications industry from short-term fluctuating political pressure, and to ultimately admit competitive entry.

Governments have a strong incentive to pursue policies with short-term benefits, even where these involve high long-term costs. Frequently, the financial targets imposed on public enterprises have been highly variable from year to year, creating uncertainties in the enterprises' planning. When budgetary constraints are tight, governments also can use public enterprises to advance political and social goals not directly related to those enterprises' main functions. The costs imposed on the telecommunications carrier by public-service obligations should be accounted for as carefully as possible and made publicly known. In addition, financial targets and other requirements should be set on a medium-term rather than an annual basis.

In some countries, efforts for regulatory reform could be unduly constrained by the inherited problems and may not be sufficient for ensuring optimal performance. Experience indicates that most governments do not easily and willingly forgo using the instruments they have in hand. As long as the publicly owned telecommunications operators comply with government requests, they may face little pressure to make efficient use of their resources. In those cases the only means to break bureaucratic inertia, political favoritism, and apathetic service delivery is through more fundamental and radical structural change—frequently privatization.

#### *Pricing policy*

*Historically, almost all economic regulation has had the same general approach to pricing. First, overall prices are set so that the firm covers its costs of doing business, including a reasonable return on its invested capital. The primary objective is to protect consumers from monopolistic exploitation; hence the focus is on controlling overall prices by controlling profits on invested capital. Second, regulators set the structure of prices, generally on the basis of perceptions of distributional equity. Three principles enter into these deliberations: prices should be nondiscriminatory, that is, they should be the same for similarly situated customers; service should be universally provided; and conflicts between these two principles should be reconciled through "value-of-service pricing."*

In telecommunications these principles produced the policy of residual pricing. The goals of universal service and

nondiscrimination dictated low residential prices. Value of service called for lower prices in rural areas. As a result, revenues for basic service were below cost, compensated by value-of-service pricing for other services. It should be noted that whereas costs were explicitly taken into account in setting the general price levels, they played no direct role in the determination of the structure of prices.

*The inefficiency of standard pricing policies.* Economic efficiency requires that services be priced at their marginal costs. A telecommunications system incurs two types of costs. The first is associated with connecting a subscriber to the network and is therefore non-traffic-sensitive. In contrast, a subscriber's usage of the system gives rise to traffic-sensitive costs that vary with the time and duration of connection, the distance traversed by the call, and whether the call is intra- or interexchange.

An efficient telecommunications pricing system would offer a two-part tariff to each subscriber. One part would be a fixed access charge, levied either as a lump sum or on a periodic basis, to recover the marginal non-traffic-sensitive costs of connecting the customer to the network. This fixed component would vary substantially among subscribers depending on their locations and other factors. The second component of the two-part tariff, related to traffic-sensitive costs, would vary with the subscriber's usage of the network and would reflect the mix and duration of intra- and interexchange calls and the times the calls were made.

A comparison of the standard telecommunications pricing policies with the above principles reveals three sources of inefficiency (see Kahn 1984). First, access and usage rates are generally averaged over a large number of subscribers, and as such they do not reflect individual geographic, temporal, or other factors that cause true access costs to vary across subscribers. Nor do local or long-distance rates reflect the large differences in the usage-sensitive costs of calls between persons in different locations. Second, typical local rates do not take into account the amount of local usage at peak hours when additional calling requires extra capacity. The standard practice of providing service on a flat-rate basis, with no charge per call or per minute, clearly leads to excessive local calling. Third, a significant portion of the costs of providing access to the network is normally recovered in charges for using the system despite the fact that those costs are largely independent of usage. As a result, the basic monthly service charge is generally low, encouraging consumers to become subscribers, or even to order second lines, when the value to them of that access is less than the cost to society of providing it.

*Inefficient pricing as an instrument of cross-subsidization.*

The standard inefficient telecommunications pricing practices are the consequence and instrument of a complex system of cross-subsidies between different subscriber groups. First, the costs of toll markets are artificially inflated above direct costs to provide a flow of revenues to local operating companies. Second, business subscribers, the predominant users of long-distance service, subsidize residential customers. But since businesses normally pass their costs on to their customers, residential telephone service is effectively being subsidized by a tax on all the purchases of goods and services produced by businesses. Third, urban subscribers subsidize customers in remote rural areas. Fourth, customers with a preference for making local calls during off-peak hours subsidize those with a preference for peak hours.

Three arguments relying on the economic characteristics of the telecommunications market and the relationship between toll and local service have been offered to justify the subsidization of local service by toll service: externalities, demand complementarity, and preserving service for the poor.

The simplest and most familiar argument in favor of the residential access subsidy is that subscription to the network yields external benefits. In principle, then, one could justify making heavy business users, who presumably benefit most from the system, subsidize the basic access charge so that they can continue reaching those who would otherwise drop off the network. It should be noted, however, that the presence of the subscriber externality argues for the underpricing of local access, not local usage. Consequently, a local measured service that allows for the distinction between access and local usage prices provides a better vehicle for the satisfaction of the subscriber externality than does a flat-rate local pricing scheme.

Another argument frequently advanced to justify the subsidization of local service—both access and usage—from toll is demand complementarity: Both access lines and local usage are required for the placement of toll calls (at least with current technologies). Thus, it is argued, proper costing would allocate part of these local costs to usage. It is true that an access line and a local connection to the toll switch are prerequisites of each toll call. Still, demand complementarity does not imply that separate markets should not exist for each of these elements of a typical toll call. In fact, optimality requires each of these to be costed and priced separately. Access costs are not caused by toll or any other traffic. To the extent that a toll call affects local traffic-sensitive costs, its price should

include a component that recovers the cost of transport to the toll switch. In principle there is no difference between a local call to another local subscriber and a local call to the point of presence of a toll carrier.

Along similar lines, it is asserted that customer access is not a service but merely a prerequisite to the provision of "real" telephone services, that is, to placing and receiving calls. But such an assertion is in itself merely semantic, and the inferences drawn from it are almost entirely fallacious. First, the defining characteristic of a service is that it is (or would be) demanded in its own right. By that criterion, access is clearly a service. Even if most customers were not interested in access in order to place calls, many would still want access if only to receive calls. Second, the relevant question is not whether access is a service but rather the efficient way to recover costs. In that context, two pertinent economic questions must be addressed. Is a separate, identifiable incremental cost associated with subscriber access? And does charging for access separately serve a purpose? The answer to both questions is definitely yes. The connection of a subscriber to the network entails scarce resource use even if the subscriber never uses the connection; and an important efficiency objective is served when consumers are confronted with prices that reflect the respective incremental costs to society of their taking more or less of each available service.

One important problem with the proposition that all marginal access costs should be recovered in the basic monthly charge is that purely cost-based prices would exclude many poor people from enjoying what has become a basic necessity in modern society. However, economic efficiency is not necessarily incompatible with the important social goal of helping the poor. The standard practice of holding down prices for all (including those who need such help the least) in order to help the poor ends up injuring almost everyone. If a subsidy for basic service is retained, it must be less haphazardly distributed and more tightly targeted at those who really need such help. It is very inefficient to lower tariffs for a large percentage of telephone users in order to help the small percentage of those who are disadvantaged.

*Second-best pricing.* Because of economies of scale in important parts of the telecommunications business, it is likely that even strict marginal-cost pricing would not provide telephone companies the revenues they require and to which they are entitled under traditional regulatory principles. This underrecovery is made worse by the presence of the subscriber externality. Prices would, therefore, have to somehow depart from marginal cost until the

revenue constraint is satisfied. It is important to recognize that not every departure from marginal-cost pricing is acceptable from the economic efficiency point of view. These departures ("second-best" pricing approaches) should be chosen so as to minimize the resulting welfare loss.

The first approximation to a resolution of this dilemma with minimum loss of economic efficiency is provided by Ramsey pricing. This pricing rule calls for markups above marginal costs that are inversely proportional to the elasticities of demand for the several services, to elicit the requisite increase in total net revenues. To the extent that prices must depart from marginal costs in at least some markets, if such departures are concentrated where demand curves are the steepest—the most inelastic—the welfare loss will be minimized.

It is immediately clear that the historic pattern of telephone pricing conflicts significantly with the above prescription. The largest markup above marginal cost is generally imposed in long-distance calling—the service whose demand is the most elastic—and the smallest markup (in fact, a negative markup) is imposed on access—the demand for which is the least elastic. Moreover a substantial portion of the total revenue requirements of subscriber plant costs is recovered from the small percentage of total calls represented by long-distance usage. This perverse pattern of markups can cause substantial losses in economic welfare—much more than the movement from first- to second-best pricing rules.

*Rate-of-return regulation versus price cap regulation.* Rate-of-return, or cost-based, regulation is the predominant form of price regulation in the telecommunications sector around the world. Regulators have been attracted to this mode of controlling the behavior of monopoly service providers because conceptually it seems fair to both the regulated firm and its customers. It permits the firm to earn sufficient revenues to cover its costs, including a fair rate of return on equity. It is also designed to protect consumers from the monopolistic pricing distortions that would normally arise if the monopolist could freely exercise its market power.

Experience has revealed, however, that even when rate-of-return regulation is executed correctly, it imposes nontrivial economic losses by creating perverse incentives associated with cost-plus contracting. Rate-of-return regulation has four major shortcomings. First, it does not give firms strong incentives for cost minimization since their costs are recovered in their rates. Second, this form of regulation does not encourage firms to be more efficient and innovative since it fails to distinguish increased

earnings attributable to increased efficiency from those attributable to the exercise of market power. Third, firms have incentives to inflate their rate base while shifting costs from services in which they face competition to those in which their market power permits them to recover revenues above the economic costs of providing the service (and hence to improperly cross-subsidize). And finally, the framework needed to support this mode of regulation is elaborate and often cumbersome, and its administrative costs are substantial and growing. During periods of inflationary buildup, the administrative cost problem becomes particularly pronounced because the regulated entities are forced to repeatedly seek interim rate relief.

An alternative set of regulatory strategies can be broadly classified as social contracts. Under the general strategy of social contract regulation, regulators first delimit a group of core activities that they continue to regulate and then stipulate a list of constraints that the regulated entity must agree to meet in the future; in exchange, regulators agree to detariff or deregulate entirely other competitive or nonessential services that the utility may offer. As long as no stipulated constraints are violated, the regulated entity may freely price any service; if it reduces costs, it may keep a share of the resulting profits (see Einhorn 1991, chapter 1; also see Brennan 1989 and Cabral and Riordan 1989).

Price caps represent a form of social contract regulation. Under price caps, aggregate index ceilings are placed on prespecified groups of services (called "baskets"); the regulated entity can freely price any service, so long as no index ceiling constraint is violated. Index ceilings are adjusted periodically to allow for expected cost inflation (easily observable changes in costs that are generally beyond the entity's control) and a precommitted rate of productivity improvement. The regulated entity retains any profits that may result from cost cutting or technological innovation.

The price cap model offers a number of advantages as compared to cost-based regulation. Under price cap regulation, the regulated entity has every incentive to minimize costs and adopt efficient technological improvements because any increases or decreases in its costs are not automatically passed through to consumers. The utility has no incentive or opportunity to strategically distort its reported cost data because costs do not enter directly into the price cap formula. Nor does the regulated entity have an incentive to expand its rate base uneconomically, because the price cap model specifies neither a rate base nor a maximum rate of return on

invested capital. Utilities have no opportunity to shift rate-based costs of competitive services on to their captive monopoly activities, and the administrative costs of regulation are reduced. In addition price caps might also offer consumers greater protection against sudden steep rate increases than cost-of-service regulation can provide. Finally, the price cap model could also reduce the resources that competitors feel compelled to commit as intervenors in the regulatory process.

Different interpretations of a price cap could offer consumers different degrees of protection against rate increases and give carriers different degrees of flexibility to change rate levels or even rate structures for capped services. At one extreme the cap requirement could be interpreted to impose a ceiling on the average rates of capped services overall. This interpretation would afford the carrier broad discretion to adjust its rate levels and structures for such services. All tariff provisions that altered the rate structure for a capped service but were revenue neutral would be presumed lawful. At the other extreme the cap requirement could mean a ceiling on the rate associated with each element of a service. Under this approach a tariff filing for a service that leaves the rate structure for that service unchanged, and also either lowers or leaves unchanged the charge for each of its rate elements, is eligible for streamlined regulation.

Between the two extremes are a large number of possibilities for defining the concept of price caps. For example, individual price caps could be imposed for certain services, whereas other services could be grouped together and subjected only to a limit on how much their rates could increase on average. For some groups of services, such a group rate constraint could be supplemented with one that limits the amount by which the rates of any individual service could rise. Clearly, the central issue is which interpretation of price cap strikes the best balance between the primary objectives of protecting consumers against unreasonable charges for services, and providing carriers with sufficient flexibility to introduce innovative services quickly and attain the most efficient mix of services their networks permit.

#### *Structural issues*

Structural regulation determines which firms are allowed to engage in which activities. An important form of structural regulation is functional separation, in which entities are prohibited from undertaking different activities simultaneously.

In the telecommunications industry structural regulation is often concerned with the extent to which firms

operating in one regulated market are permitted to enter others. One question that might be asked is, should a telecommunications entity be permitted to vertically integrate across manufacturing, interstate toll service, and local-exchange service?

*Vertical integration.* An important reason for vertical integration has to do with investment incentives. As asset specificity becomes more important, exchange relations take on a progressively stronger bilateral trading character. The reason is that parties to such trades have a stake in preserving the continuity of the relationship. At the same time, however, problems of adapting bilateral contracts to changing circumstances normally arise. Autonomous market contracting is thus supplanted by more complex forms of governance as asset specificity deepens. Some transactions may be removed from the market and organized internally instead. Vertical integration may be viewed then as a response to relatively high costs of market exchange (see Williamson 1971).

In an incentive sense, internal organization attenuates the aggressive advocacy that epitomizes arm's length bargaining. Even if interests are not perfectly harmonized, they are at least free of representations of a narrowly opportunistic sense. The most distinctive advantage to the firm is the wider variety of instruments at its disposal for enforcing intrafirm in comparison with interfirm activities.

The main argument against vertical integration is that it can be used strategically to achieve anticompetitive effects. Established firms may use vertical integration to increase finance requirements and thereby to discourage entry if potential entrants feel compelled to adopt the prevailing structure (and they frequently do). Moreover, firms may use the excess profits they realize in activities where they enjoy market power to finance aggressive behavior in markets where they face strong competition.

In the specific context of the telecommunications industry, regulators frequently must address whether separation should be imposed on (a) the activities of network operation and equipment manufacturing; (b) local, long-distance, and international operations; (c) fixed and mobile network operations; and (d) the activities of network operations and the retailing of services over the network. These questions generally entail a tradeoff between the cost-efficiency advantages and the anticompetitive disadvantages of integration. The case for allowing vertical integration, therefore, depends on the effectiveness of regulating conduct.

As long as the production of terminal equipment is not a natural monopoly and the equipment manufactured by different firms does not produce noise or other forms of

harm for the telephone network, regulatory policy should encourage access to local telephone systems on equal terms. Freedom of entry is preferable to arrangements that bar local telephone operators or other firms from manufacturing or selling telephone equipment. Vertical integration into equipment manufacturing, however, would enable the monopolist to evade regulation by transferring monopoly prices of services to the manufacturer through excessive equipment prices. Another concern is that the integrated monopolist might engage in economically unwarranted self-dealing by purchasing from its own affiliate even if competitors offered superior equipment at lower prices.

The issues involving local versus long-distance services are more complex, in part because of the problem of shared costs and the efficiency derived from coordinated operation of an integrated network (such an efficiency arises because a large, integrated operator commonly routes calls during busy periods through distant switching centers if nearer ones are operating at full capacity). This routing is only one of a variety of network-wide planning decisions that may make production less costly when local and long-distance operations are contained within one firm.

Dominant telecommunications carriers might use their market positions in basic transmission services to discriminate against other vendors' competitive offerings that rely on those basic network services. For example, dominant carriers could adopt network interconnection standards so as to prevent or limit competition from other carriers. Dominant operators could also improperly subsidize competitive services with revenues from regulated services.

Regulators generally face a very difficult, if not hopeless, task in preventing cross-subsidization of competitive services, discriminatory acts against competitors, and ultimately, evasion of regulatory objectives concerning captive monopoly customers. Given these emerging realities of regulatory practice, a policy of "quarantine" whereby the regulated monopoly carrier is prevented from participating in potentially competitive markets might be appropriate. In this context the separation of local and long-distance services represents an attractive structural option.

*Access pricing and interconnection issues.* Like the other network utilities, the telecommunications industry is characterized by transportation and distribution networks (transmission media and switching centers) linking upstream production with downstream consumption. As outlined earlier, substantial competition pervades many activities in the sector, while in other portions of the industry's operations, competition is weak or nonexistent, at least given the present state of technology. Thus

whereas interexchange service is now regarded as fully competitive, entry into this segment of the market requires access to subscribers in the local-exchange loop, which retains many of the characteristics of natural monopoly. These basic network elements, therefore, constitute essential inputs for the provision of many telecommunications services—inputs without which suppliers cannot hope to operate. As such, these monopoly services are referred to as “bottlenecks” or “essential services.”

Absent regulatory constraint, the holder of the bottleneck monopoly could repress competition by creating artificial handicaps for its rivals in the market for the final product sold to consumers. The monopolist could impose costs on its competitors by impeding their access to the bottleneck, thereby forcing them to raise their prices to cover their elevated costs, and thus weaken their ability to compete. The monopolist’s market success would be caused by neither greater efficiency nor better meeting consumers’ demands, but rather by the imposition of socially unnecessary and harmful costs on rivals and their customers. This is the fundamental complicating phenomenon hindering the deregulation of local-exchange service.

*The objectives of access pricing.* As technological change and deregulation reduce entry barriers in telecommunications, rival firms will seek to interconnect to the telecommunications network at a greater number of locations than in the past. At each interconnection point an access price will have to be determined. The terms of access should not distort the process by which prices are adapted to consumer preferences and demands for telecommunications services. Prices should be sufficiently high to be compensatory (at least cover the long-run incremental cost of the use of the network by the entrant) yet not so high as to preclude efficient operations by the entrant.

The primary challenge for public policy is to set a level and a structure of access prices that promote dynamic efficiency through efficient entry and investment decisions while enabling the incumbent firm to remain financially solvent and sustain the cost of social obligations (see Cave and Doyle 1994). Regulation should, therefore, ensure that there is sufficient pressure on the incumbent to operate in an efficient manner but that no unnecessary duplication of network construction takes place.

In the unpredictable and fast-changing technological and marketing environment of today’s telecommunications sector, it is difficult to predict what collection of basic network elements will prove to be essential to the efficient provision of some desired service by each supplier. Therefore the opportunities for competition to work

effectively and to bring innovative offerings to consumers would be enhanced by making available on an unbundled and nondiscriminatory basis any basic network element, or any collection of functions, needed by the entrant.

*Vertical structure and access pricing.* Monopoly control of bottleneck facilities can create irresistible incentives to behave anticompetitively and subsidize unregulated competitive activities from regulated monopoly activities. One potential remedy is vertical separation. Vertical unbundling generally reduces the incentive and ability to leverage market power and simplifies the access-pricing problem. It is the direct cost of providing access alone that is relevant for the correct choice of the interconnection charge. However, vertical separation may lead to the loss of important scope economies, as discussed earlier.

Vertical integration involves the incumbent owning the bottleneck facilities and competing against the entrants seeking access to them. In this case, in addition to the direct cost of access being relevant, there is also the opportunity cost incurred (all the potential earnings that the incumbent forgoes) when the entrant rather than the incumbent offers the service.

Whether the telecommunications industry should be vertically separated or vertically integrated depends on the extent of scope economies and on the costs of regulation. Clearly, when the bottleneck facilities are vertically isolated, the industry is easier to regulate since the opportunity cost to the input supplier does not arise. However, the extent of joint economies may be sufficiently strong to offset the additional burden of regulating access prices.

*The Baumol-Willig rule.* The Baumol-Willig rule for efficient components pricing sets the charge for the use by competitors of a component part of the facilities controlled by a vertically integrated firm (see Baumol and Sidak 1994). In practice the facility of interest is that which competitors cannot duplicate economically—the local distribution, that is, transmission from the local exchange to customers’ premises.

Economic efficiency requires the price of any product to be no lower than the product’s marginal cost or its average incremental cost. The pertinent marginal cost as well as the average incremental cost must include all opportunity costs the supplier incurs in providing the product. Opportunity cost refers to all potential earnings that the supplying firm forgoes by offering services to competitors that force it to relinquish business to those rivals. The Baumol-Willig rule states simply that the optimal fee for access to a monopolist’s bottleneck facilities is the sum of the direct incremental cost of permitting the competitor to use the facilities and the opportunity cost to the

monopolist of supplying this downstream competitor with access to those facilities.

The rule offers the prospect of success to entrants that can add efficiently to the supply of the final product, whereas it ensures that inefficient entrants are not made profitable by an implicit cross-subsidy extracted from the incumbent. Thus the rule always assigns the supplier's task to the firm that can do it most efficiently.

### **A future research agenda**

Since the early 1980s many countries (notably Japan, the United States, and several countries within the European Union) have undertaken substantive steps to liberalize their telecommunications markets. Most enjoyed fully developed telecommunications networks and had already made significant progress toward rate rebalancing by the time they embarked on their regulatory reform programs. Although market liberalization has generally led to significant improvements in service and performance, many policymakers question its appropriateness for developing countries.

Most developing-country governments have owned and operated the basic telecommunications infrastructure, which has had to compete for investment capital with other socially important infrastructure projects. As a general matter, there has been a significant shortfall between aggregate demand for central treasury capital and its supply. Some countries have even viewed the telecommunications sector as a source of funds, including hard currency, to support other governmental programs. Because of the shortage of investment capital, most of these countries have highly underdeveloped telecommunications networks. In addition, the telecommunications revenues of developing countries often are dependent largely on high prices for international services. High international rates are considered the most expedient and least disruptive way to generate revenues, especially as compared to higher local, residential service rates. Practically, revenues from international services hold down local rates while permitting the expansion and affordability of the telecommunications infrastructure.

There is little doubt that a policy of open entry would ensure businesses a broad array of telecommunications services and competitive prices. Large businesses generate high demand for telecommunications services, especially the international services whose prices are above their underlying costs.

Pursuing an open entry policy in such a context would raise serious economic issues. No empirical evidence exists that a public network can be developed from a low

level of penetration to an appropriate level of "universal availability" while simultaneously permitting competitive entry for providers of network infrastructure and basic services. It is not clear that any firm would target the less attractive market segments until significant rate rebalancing is implemented. This lack of empirical evidence is due to the fact that countries have introduced competition only after most of the population was served by a fully developed network. Without exception, these countries used internal price subsidies to develop their public network systems. Subsidies were directed from international services to domestic, long distance to local, business to residential, and urban to rural.

Within a developing country, a successful open entry policy would depend on a firm's ability to serve both the lucrative, telecommunications-intensive, large-customer market and the residential and rural markets. Assuming, however, that the firm was obligated to provide some level of "universal service," it is not clear that the firm could generate sufficient revenues to finance the development of the public network infrastructure while other entities compete for the more lucrative segments of the market. Competitive pressures would drive down the margins available from services sold to large customers.

The above arguments suggest that there exists a serious policy tension between basic network development and market liberalization. Indeed, in privatizing their telecommunications industries in the last few years, several developing countries have granted an "exclusivity period" during which the privatized entities have been protected from competition—ostensibly to finance network expansion with retained profits. However, opponents of this approach argue persuasively that the rapid technological change of the last few years has largely eliminated the need for exclusivity in order to finance network development, since new technologies are offering low-cost alternatives to the traditional fixed-loop network. The perceived tension between network development and market liberalization, and the extent to which it has been eliminated by technological change, is an important policy issue (especially for countries undergoing privatization) that needs further empirical analysis.

Another important issue that requires further analysis relates to the relative advantages of the price cap model as compared with cost-based regulation. Although on efficiency grounds the price cap model is clearly a preferred method of controlling monopoly behavior, it shifts most of the risk to the regulated entity—as opposed to cost-based regulation, which shifts most of the risk to the consumers. Therefore the immediate application of price cap

regulation might not be appropriate in countries that seek to attract substantial investment in their telecommunications infrastructure.

### Notes

1. Externalities in communications are discussed in Rohlfs 1974, Squire 1973, and Willig 1979.
2. For an illuminating analysis of the natural monopoly characteristics of local exchange service, see Greenwald and Sharkey 1989.

### References

- Baumol, William J., and J. Gregory Sidak. 1994. "The Pricing of Inputs Sold to Competitors." *Yale Journal on Economic Regulation* 11(1): 171–202.
- Bolter, Walter G., James W. McConnaughey, and Fred J. Kelsey. 1990. *Telecommunications Policy for the 1990s and Beyond*. Armonk, New York: M. E. Sharpe, Inc.
- Brennan, Timothy J. 1989. "Regulating by Capping Prices." *Journal of Regulatory Economics* 1(2): 133–47.
- Brock, Gerald W. 1981. *The Telecommunications Industry: The Dynamics of Market Structure*. Cambridge, Mass.: Harvard University Press.
- Cabral, Luis M. B., and Michael H. Riordan. 1989. "Incentives for Cost Reduction Under Price Cap Regulation." *Journal of Regulatory Economics* 1(2):93–102.
- Cave, Martin, and Chris Doyle. 1994. "Access Pricing in Network Utilities in Theory and Practice." *Utilities Policy* 4: 181–91.
- Einhorn, Michael A., ed. 1991. *Price Caps and Incentive Regulation in Telecommunications*. Boston, Mass.: Kluwer Academic Publishers.
- Greenwald, Bruce C., and William W. Sharkey. 1989. "The Economics of Deregulation of Local-exchange Telecommunications." Bellcore Economics Discussion Paper 56. June.
- Kahn, Alfred E. 1984. "The Road to More Intelligent Telephone Pricing." *Yale Journal on Regulation* 1(2): 139–57.
- Rohlfs, J. 1974. "A Theory of Interdependent Demand for a Communications Service." *Bell Journal of Economics* 5(1): 16–37.
- Sharkey, William W. 1982. *The Theory of Natural Monopoly*. Cambridge: Cambridge University Press.
- Squire, Lyn. 1973. "Some Aspects of Optimal Pricing for Telecommunications." *Bell Journal of Economics* 4(2): 515–25.
- Waverman, Leonard. 1975. "The Regulation of Intercity Telecommunications." In Almarin Phillips, ed., *Promoting Competition in Regulated Markets*. Washington, D.C.: The Brookings Institution.
- . 1989. "U.S. Interexchange Competition." In R. Crandall and K. Flamm, eds., *Changing the Rules: Technological Change, International Competition, and the Regulation in Communications*. Washington, D.C.: The Brookings Institution.
- Williamson, Oliver E. 1971. "The Vertical Integration of Production: Market Failure Considerations." *American Economic Review* 61(2): 112–23.
- Willig, Robert D. 1979. "The Theory of Network Access Pricing." In Harry M. Trebing, ed., *Issues in Public Utility Regulation*. East Lansing: Michigan State University.

# Competition and regulation in the railroad industry

Ioannis N. Kessides and Robert D. Willig

The rail industry has been one of the most extensively regulated sectors in the economy (see Friedlaender 1969; Keeler 1983). Price, entry, exit, financial structure, accounting methods, vertical relations, and operating rules have all been subject to some form of governmental control. The public utility paradigm of governmental regulation has been applied to the rail industry on the assumption that the industry's economic characteristics preclude competitive organization and any need for market responsiveness.

Over the past three decades, however, economists and policymakers have become increasingly critical of the traditional public utility model (see Friedlaender 1971; Levin 1978, 1981a; and Boyer 1987). It has become common wisdom that there is often effective competition in the relevant economic markets in which rail carriers seek to meet demand. It is also generally agreed that governmental restrictions on the structure and conduct of firms in the rail industry impose considerable costs on society. The misallocation of freight traffic among competing transport modes, excess capacity, excessive operating costs, and poor investment decisions are often the result of misguided regulatory policies. Regulatory controls have, therefore, been held responsible in large part for the poor financial condition of the railroads, for the deterioration of the rail plant, for the suppression and delay of cost-reducing innovations, and for the mediocre quality of rail service.

The purpose of this chapter is to suggest a set of principles for restructuring railroad regulation, and indeed for restructuring the orientation of railroad entities, for the sake of the public interest. We focus first on the economic characteristics of the rail industry and their implications for the design of efficient regulatory policy. Then we apply powerful sets of analytic tools to clarify the relevant principles for reform. Much can be learned about policies to promote the public interest, from an understanding of market demands for the industry's products and the

nature of the productive techniques available to the industry's firms. Indeed, before the implications of policies aimed at rate regulation or infrastructure investments can be fully assessed, a full understanding of the nature of technology, costs, and demand facing the rail industry is required. The role of the government in relation to market behavior should therefore be based explicitly on the underlying economic characteristics of the industry and the technological conditions of its production.

We hope to impart the following message: Recent developments in industrial organization analysis and in regulatory practice call for a major reorientation of public policy toward railroads. We suggest in this chapter a set of principles to guide such a reorientation.

## Public policy issues in the rail industry

The economic characteristics of the rail industry make it a natural target for government intervention, yet also render it particularly difficult to regulate in the public interest. The old regulatory systems failed to handle the central regulatory problem arising in railroads and certain other major industries (for example, telecommunications, electric power, and postal services): the mixture of competition and monopoly elements in supply. Indeed, in these industries, just as in the railroad industry over the years, regulation has stifled competition in the provision of services, restricted the benefits of economies of scope, retarded innovation, and fostered inefficient service. Regulation thus has harmed the public interest while protecting it from the exploitation of monopoly power (see Willig and Baumol 1987). The first-best lesson of the perfect competition model, calling for prices to be set equal to marginal costs, has no doubt contributed to the common regulatory ethos that seeks to equate price to some measure of cost. This doctrine has frequently been used where it is completely inappropriate and without logical foundation—that is, in cases where prices should be based on demand as well as cost considerations.

This section focuses on the central pricing issues involved in the partial deregulation of railroad rates.<sup>1</sup> It articulates principles to guide regulatory oversight of the rate setting of unsubsidized railroads—principles that are both consistent with economic analysis and essential for the protection of the public interest. Public interest regulatory oversight of railroad pricing involves two basic issues. The first of these is the adequacy of revenues, the criteria by which adequacy can be judged, and the means by which it can be achieved. The second issue is the choice of rates that are both consistent with adequate revenues and best for the public interest.

In a regime of deregulation, one of the key elements in protecting the public interest is the elimination of any residual regulation that effectively prevents the rail network from achieving financial viability. The public will hardly be well served by a set of regulatory rules that makes it impossible for the railroads to compete in the financial marketplace. The rail network that resulted would become increasingly deteriorated and obsolete, and cumulative abandonment of service would become the prevailing practice.

In determining prices for the outputs of multiproduct railroad firms, regulators have long faced a number of difficult issues that flow inevitably from the basic economic characteristics of the industry discussed above. The endemic economies of scale and scope imply that straightforward measures of costs cannot be used to dictate pricing. Economies of scale imply that marginal cost pricing, absent subsidy to the firm or multipart tariffs, will not allow the firm to break even. Further, because the shared costs that are a concomitant of economies of scope cannot be unambiguously identified with individual products, any rule selected to associate shared costs with individual services will be arbitrary. Such arbitrary measures as fully distributed (or “fully allocated”) costs, therefore, cannot substitute for marginal cost measures as decision rules for proper pricing. The misguided search for a purely cost-based substitute rule relies inappropriately on the model of perfect competition for guidance on regulation.

A system of rate regulation in which costs are apportioned on any basis other than demand is inappropriate because it is highly unlikely that prices set will permit railroads to achieve an adequate rate of return. Moreover, such a method leads to serious inefficiency by discouraging innovation and by generating prices that are too high to attract competitive traffic. The absence of competitive traffic in turn severely restricts the amount of services delivered by railroads and thus produces still higher rates

for the remaining traffic (see Braeutigan 1977; Kahn 1988).

By contrast, there are sound pricing principles that promote economic efficiency while simultaneously removing impediments to adequate returns for carriers. These principles can be applied in a practical fashion to assess the reasonableness of those rates that are judged to require continued regulatory oversight (see Braeutigan 1979, 1984). The principles lead to demand-differentiated prices, sometimes referred to as Ramsey prices, which apportion all unattributable fixed and common costs of the railroad among its services on the basis of the values of those services to consumers, mathematically expressed as their elasticities of demand. Economically efficient differential pricing combines cost and demand factors in an optimal manner, by pricing each service at a markup over marginal costs that is inversely related to the elasticity of demand for that service. The resulting set of rates encourages the purchase of more rail transportation services by more shippers than would be the case with fully distributed cost-based pricing, thereby creating a larger traffic base over which unattributable costs can be apportioned and lowering prices for shippers generally. Ramsey pricing maximizes the opportunity for rail carriers to earn an adequate rate of return on capital and fosters innovation and efficiency in the provision of rail transportation services by rewarding carriers that achieve cost reductions.

Economically efficient differential pricing is entirely consistent with the hallmark of deregulation: that market forces, rather than regulation, should control the rates for transportation services. Thus, when a particular type of traffic is subject to competition, direct or indirect, regulatory intervention is unjustified. Furthermore, so long as a railroad's earnings fall short of its cost of capital, the need for regulatory constraints on any of that carrier's rates is minimal, and to the extent such a constraint prevents the carrier from earning an adequate return in the future, it is contrary to the public interest. By definition, there is no danger that such a carrier is receiving excessive overall profits derived from market power or any other cause. In addition, if the rate for any service supplied by a railroad not yet earning adequate revenues overall is held down by regulation below the level that consumers of that service are prepared to pay rather than do without the service, then, in the long run, even those consumers will be harmed. The carrier will find it unprofitable to invest the necessary replacement and maintenance capital, causing a deterioration in, and ultimate withdrawal of, the service.

*The proper criterion for adequacy of revenues*

Since avoiding impairment of financial viability plays so crucial a role in any rational program of rate regulation, it is important to describe the criterion by which financial viability can be judged. What information is required to determine when a firm's revenues are adequate to cover its pertinent costs? While the answer would appear to be obvious, the history of regulation demonstrates rather forcefully that it is in fact widely misunderstood. The basic issue is that the cost of the firm's capital, including any capital it has generated internally, must always be included in these calculations.

The logic of this criterion is straightforward. Revenues are defined to be adequate when they are just sufficient to enable the firm to attract the capital needed for maintenance, replacement, modernization, and whatever expansion is justified by demand conditions. If revenues are below this level, the deterioration and eventual disappearance of the service are inevitable.

Adequate revenues are those that provide a rate of return on net investment equal to the current cost of capital (that is, the level of return available on alternative investments). This is the revenue level necessary for a railroad to compete equally with other firms for the financing needed to maintain, replace, modernize, and, where appropriate, expand its facilities and services. If railroads cannot earn the fair market rate of return, their ability to both retain existing investments and obtain new capital will be impaired, because the existing and prospective funds could be invested elsewhere at a more attractive rate of return. Indeed, the market for funds is one of the most competitive in the economy. There is no escaping the following principles that determine the adequacy of revenues:

- The firm's overall rate of return must be equal to the returns currently earned by the typical firm with similar risks elsewhere in the economy. Otherwise, the required funds will be denied to it.
- The adequacy of a firm's revenues can be judged only by comparison with the earnings of firms outside of regulated industry. If the regulated industry's earnings are compared with the market value of the firm's equity, the market prices of those securities will automatically adjust themselves downward to match any act by the regulator that restricts the earnings of the firm below a compensatory rate of return. Such a comparison will thus appear to justify any earnings restrictions, no matter how inappropriate.
- In determining the firm's revenue requirements for financial viability, the rate of return obtained by compar-

ison with other industries must be applied to a rate base that covers the economic replacement cost (under regulation) of all facilities. (Suitably updated historic costs can be utilized instead of replacement costs if the allowed rate is expressed in nominal terms.)

- With the rate base determined in this way and the rate of return on that rate base equal to the cost of capital, as given by earnings prevailing elsewhere in the economy, the result will be the total net earnings figure that can appropriately be considered to be adequate for the railroad to compete successfully in the capital market.
- This earnings figure must not be applied as a rigid ceiling. Doing so would make it impossible for railroads to earn this figure over the long run, since they would be precluded from making up for any revenue shortfalls resulting from temporary downward fluctuations in demand for their services.

To make sense economically, prices must never be incompatible with this earnings level. Of course, no prices can guarantee that a railroad will earn adequate returns overall. If demands for its services are insufficient, operations are conducted wastefully, or services are poor, even appropriate prices cannot be expected to lead to profitable operation. But once the railroads are permitted to charge appropriate prices in a competitive environment, the regulatory impediments to financial viability will have been eliminated. It then will be up to the railroads to take advantage of the opportunity through economic operations, quality service, and effective marketing.

*The regulatory problem*

Indivisibilities, pervasive economies of scale and scope, high costs of entry, and small-numbers competition in the railroad industry are all consistent with the likely persistence of prices in excess of marginal cost. However, while scale economies go hand in hand with natural monopoly, a railroad may or may not have the price-setting discretion that characterizes the textbook monopolist. It all depends on whether the activities characterized by economies of scale and scope are shielded from other sources of competition in the relevant market and whether there are protective barriers to entry.

In the railroad industry, extensive capital sums must be sunk in way and structures and in a variety of ancillary facilities to create new rail lines. The sunk cost and longevity of railroad capital may suggest that the railroad industry cannot be conceived to be contestable. However, railroad services are far more contestable than these impediments to rail entry would suggest because other modes of transportation, such as trucking and water

carriage, often exert strong competitive pressures on the rates charged for shipment of a wide variety of commodities.<sup>2</sup>

The basic patterns of railroad regulation, established many decades ago in wholly different market conditions, are simply obsolete. Their premise was that railroads had a collective monopoly, or near-monopoly, in land transport. This condition disappeared long ago, if it ever existed. Nearly every sphere of rail freight service now faces intense competition. Rival products and rival sources of supply (including trucks, barges, and alternative rail routes) are likely to impose effective competitive constraints on many, if not most, rail activities. In those activities where there is no evidence that the railroad holds a position of market dominance, the industry should be offered freedom in pricing. Still, there remain instances in which the competitive checks of intramodal, intermodal, geographic, and product competition are weak or nonexistent. There is an understandable apprehension that in such cases market forces may not be relied on to prevent excessive pricing. The resulting monopoly power is the basic justification for the regulation of rail rates and earnings and defines the basic task with which regulation must grapple.

Before discussing the appropriate means to deal with this issue, however, it must be emphasized that, in practice, effective competition can assume a variety of subtle forms. Therefore, one must never proceed in haste to undermine the workings of the market through special intervention. Railroads do not face only the competition of trucks and barges. For example, oil and natural gas shipped by pipeline competes with coal shipped by rail; since coal shipment is profitable to the railroads, the competition of petroleum products limits the price railroads can charge for carrying coal. Also, the market served by one railroad may compete for the coal with a market served by another. This situation, too, can keep rates in line.

#### *The cost allocation problem*

The presence of substantial economies of scale and scope in the railroad industry creates several problems for government regulation. Perhaps the most troubling is the fact that it is impossible to allocate, in any nonarbitrary way, a share of fixed and common costs to any one of a railroad's many activities. There is simply no way to subdivide those costs in a mechanical fashion that is unique and has any foundation in economic logic.

In practice, regulatory authorities historically have determined tariffs based on so-called fully distributed (or

allocated) costs. Under this method regulators do (somehow) allocate shared production costs to individual services. Each service is then required to generate revenues that will cover all the costs associated with that service. Although it is often argued that there is no sound economic rationale for fully distributed cost pricing, this practice obviously has economic consequences.

Traditionally, regulatory proceedings have focused on three types of fully distributed cost rules. The first of these is the distribution of shared costs on the basis of a common measure of utilization, such as gross ton-miles. Under this approach, termed the relative output method, shared costs are allocated in proportion to the number of units of output of each service. A second approach is the allocation of shared costs in proportion to the costs that can be directly attributed to the various services. This attributable cost method has also been traditionally used by many unregulated firms in their allocation of overhead costs. A third scheme requires allocation of shared costs in proportion to the gross revenues generated by each service. This gross revenue approach has been frequently used to allocate overhead costs between freight and passenger services.

In addition to costs that are directly attributable, a service may also be assigned a portion of those costs that cannot be clearly associated with any one service. Railroad track, for example, is used in the transport of many kinds of freight. Shared costs may therefore constitute a large portion of total costs. Thus, the method of allocating shared costs may significantly influence the rate required for any particular service.

#### *The problems in fully allocated cost pricing*

Fully distributed cost pricing rules suffer from several disabilities:

- Since fully distributed costs bear no direct relationship to marginal costs, there is no basis in economic efficiency for fully distributed cost pricing.
- On grounds of economic efficiency, it may sometimes be desirable to set a price for some service so that the revenues it generates do not cover its fully distributed costs.
- Because the determination of fully distributed costs is arbitrary, there is no economic basis for concluding that a service is being subsidized by other services if its revenues are less than its fully distributed costs.
- Fully distributed cost pricing is anticompetitive since it prevents a supplier from offering a service at a proposed tariff less than a fully distributed cost price, particularly if the proposed tariff exceeds the marginal cost of providing the service. In addition, there is circular reasoning

behind the fully distributed cost practice. Tariffs that are determined to be “appropriate” at a given time may depend on the existing levels of output or revenues, which in turn depend on previous tariffs. Thus, fully distributed costs may depend on the acceptance of a prior tariff structure.

- The most serious defect of fully distributed costs as a basis for rate determination is that they do not necessarily measure marginal cost responsibility in a causal sense—they are costs that are averaged by an arbitrary method. They do not measure by what amount costs would be increased if additional quantities of any particular service were taken, nor do they measure by what amount costs would be reduced if the service were correspondingly curtailed. Also, being apportionments of historical costs, even when they do accurately reflect historical responsibility for the incurrence of these costs among the respective users, they do not provide a reliable measure of what will happen to costs in the future if particular portions of the business are expanded or dropped.
- Finally, the fully allocated cost criterion completely neglects any demand data. Even if based on “relative use” as measured in tons or ton-miles, it cannot capture the role of demand, which economic analysis has shown to be vital in the choice of optimal prices. Even the best-intentioned of fully allocated cost standards must employ some rigid criterion to allocate the portion of a railroad’s total costs that is not directly attributable to any one of its services in particular. But no such fixed allocation criterion can possibly reflect the subtleties, fine structure, and changes in patterns of demand for the railroad’s services that are induced by external developments and that clearly call for adjustments in its prices. This, of course, is true not only of a standard fully allocated cost approach, but of any rigid formula that bases future prices on cost data of the past and thus cannot take account of changes in demand.

It may seem paradoxical that fully allocated cost criteria, which are apparently designed to ensure that all costs are covered by revenues, can in fact preclude rail carriers from achieving financial viability. The reason is that ceilings based on fully allocated costs are set so that unattributable costs are divided in an arbitrary manner among all types of traffic. Then, for these costs to be recovered, all types of traffic must actually move at the rates that include the arbitrary cost allocations. But traffic with transport value that is below average for its tons, ton-miles, or other allocator will not move by rail at those rates. That is, any service whose demand is insufficient to cover its allocated share of total cost at the fully allocated

cost-determined price will have a revenue shortfall that fully allocated cost ceilings will prevent other services from making up. Consequently, if the unattributable costs are substantial, and if the values of rail services vary substantially, fully allocated cost rate ceilings will preclude attainment of adequate revenues.

The effects of fully allocated cost pricing on the utilization of transport resources are equally pernicious. In doing their best to earn adequate revenues despite the handicap imposed by fully allocated cost rate ceilings, rail carriers will be unable to preserve traffic whose value to the shipper exceeds its attributable cost but which falls sufficiently far below fully allocated cost. True, in the absence of fully allocated cost regulation, any such traffic could contribute revenues that exceed the costs that it causes and would provide social benefits greater than social costs. But with fully allocated cost rate ceilings, this traffic will reduce the net revenues of the rail carrier and will thus not be compensatory. The reason is that this traffic will be assigned its portion of unattributable costs on the basis of its tons, ton-miles, or some other arbitrary allocator, thereby reducing the share of those costs allocated to other traffic with higher value, and consequently reducing the ceiling and the rates on that traffic.

Fully allocated cost rate ceilings may also stifle the incentives of railroads to innovate and compete. A rail carrier cannot be expected to invest in new facilities, research and development, and marketing activities designed to elicit new traffic if the financial gains from the new traffic are counterbalanced by induced decreases in the ceilings on the rates charged to pre-existing traffic. Similarly, a rail carrier could not be expected to compete for freight by offering low rates if the necessary markups were much below the arbitrary allocations of unattributable costs; if it did so, it would never earn adequate revenues because its gain from the low-rated traffic would be outweighed by the induced decrease in the ceilings applied to more highly rated traffic.

#### *Long-run marginal cost and pricing efficiency*

The indivisibilities, economies of joint production, and high fixed costs that make small-numbers competition in the railroad industry an inevitable consequence also render the traditional measure of static deadweight loss incomplete as a welfare indicator. A regime of marginal cost pricing would eliminate the deadweight loss. But marginal cost pricing is a questionable regulatory objective, since the railroads would incur substantial losses. If the regulator attempts to force rates to equal marginal costs, overall revenues will fall short of overall costs.

Without subsidy, reduction of the short-run welfare loss to zero would cause long-run deterioration of the industry's capital stock. If revenues are to cover total costs, rates for rail systems that are characterized by scale economies must generally lie above the costs economically attributable to individual services.

It should also be noted that the use of long-run marginal cost to measure pricing efficiency frequently leads to misguided rules that could force the railroad into a pattern of behavior in conflict with the dictates of the market. Indeed, the rigid requirement that each rate always cover the long-run marginal cost of service is tantamount to a prescription of pricing inefficiency for railroads. Moreover, such a misguided decision would be likely to impose a heavy penalty on the public by sometimes depriving it of a valuable service at a price it is willing to pay and that also best serves the interests of the company—namely, a price that lies between long-run and short-run marginal cost.

The role of a cost floor as a measure of efficiency is to determine whether the railroad would be better off without the traffic in question. There are two basic reasons why it will often be appropriate for a price to lie below the corresponding long-run marginal cost. First, many investment decisions that were entirely rational and appropriate when they were made will subsequently be affected by unexpected developments. Such eventualities may cast a shadow over the future of the service that utilizes the investment. A railroad is always better served by carrying any and all traffic that can cover its short-run avoidable costs and contribute to its fixed and common costs, than by abandoning the service. The test of efficient pricing above short-run avoidable costs is whether the railroad is pricing in accordance with market demand. So long as the revenue-inadequate railroad is charging profit-maximizing rates, it is necessarily pricing efficiently; if the price maximizes the service's contribution to company profits, clearly no other price can conceivably bring that service closer to being compensatory in the long run.

The second reason why efficient prices will often fall short of long-run marginal cost affects even services whose financial viability is absolutely clear. Whether a railroad will be able in the long run to earn revenues that are sufficient to cover the replacement cost of a particular service or a group of services depends on the level of demand over time. The rail industry is strongly affected by business fluctuations in the economy, and demand for individual rail services and groups of services can and does vary widely over time. Even a service whose financial viability is absolutely clear will certainly encounter years in

which business is good and other years in which business conditions are poor. In the less prosperous years, the firm's earnings will often fall short of long-run marginal cost because market conditions permit no alternative. Of course, the shortfall will then be made up during the prosperous periods. In this manner then, the firm will in the long run meet its revenue requirements. But to insist that prices always cover long-run marginal costs is effectively to undermine the market pricing process and, very likely, even the viability of the service. Doing so would clearly distort the intertemporal pattern of usage of the service and thereby reduce economic efficiency. In addition, innovation and improvements in operating efficiency over time could potentially reduce costs and enhance contribution. A rule that assumed assets would not be replaced simply because current revenues from a particular service were depressed would remove any incentive or ability to respond to upswings in demand or make improvements in efficiency that would otherwise permit the service to continue.

The long-run marginal cost should never be used mechanically as a rigid minimum cost floor in the pricing of an operating railroad. At the same time, it should be emphasized that the long-run marginal cost cannot serve legitimately to establish the level of efficient pricing above short-run costs at any point in time. Instead, efficient rates will always have to be consistent with demand. This is true regardless of whether a railroad has market dominance over a particular service or has achieved adequacy of revenues. The demand for each service always helps to determine the contribution that service should make to the railroad's overall costs if its behavior is to comport with the requirements of economic efficiency.

#### *Economically efficient pricing*

If there were no need for enterprises to be financially self-supporting, an ideally efficient allocation of society's resources would exist if the price of each good or service were equal to its marginal cost. At such prices, consumers elect to purchase all units of goods and services that yield them benefits larger than the costs of providing them. And, in response to such prices, consumers avoid purchasing units that yield them benefits smaller than the costs of providing them. As a result, the economy misses no opportunity to allocate resources to uses where they yield benefits greater than costs, and no resources are allocated to uses with benefits lower than costs.

In industries without substantial fixed costs, competition tends to result in prices that approximate marginal or incremental costs. In the railroad industry, however, the

prevalence of large fixed and common costs makes it impossible for the supply of rail services to become financially self-supporting with marginal cost pricing. The financial infeasibility of marginal cost pricing rules out any sensible mechanical or formula-based procedure for regulatory determination of rates. In particular, compensatory rates cannot be determined by the regulator on the basis of cost data alone since the financial viability of any price depends also on the quantity of rail services customers are willing to buy at that price. This is true because there is no correlation between demand considerations and any cost accounting convention.

Allocation of fixed and common costs in accord with any non-demand-based apportionment rule will almost invariably produce inconsistencies with the patterns of shipper demands. Some rates will be too low, and consequently the railroad will receive less than the optimal contribution from those services. Other rates will be too high, so that the railroad will earn either less than the optimal contribution or no contribution at all. In short, in a multiproduct industry with uncongested fixed and common costs, the pricing of individual services on the basis of any cost allocation is contrary to the interests of both the operating entities and the shipping public. Rational determination of prices must be based on both cost and demand conditions to permit adequacy of revenues and achieve efficiency.

#### *Demand-based differential pricing*

Non-demand-based cost apportionment methods do not necessarily reflect the railroad's ability (or inability) to impose the assigned allocations and cover its costs. Thus, they frequently overassign or underassign the carrier's unattributable costs to particular services. If a carrier sought to apply fully distributed cost pricing to all its traffic, it would lose that portion of the traffic for which demand could not support the price assigned. In that event, the remaining shippers would be saddled with a larger portion of the carrier's unattributable costs since they would no longer share those costs with the lost traffic.

Ramsey prices, in contrast, apportion all of the railroad's unattributable fixed and common costs among its services on the basis of their demand characteristics. Each service is priced at a markup over marginal cost that is inversely related to the elasticity of demand for that service. Services whose demands are highly elastic are assigned prices that are very close to their marginal costs, whereas services whose demands are very inelastic are priced well above those costs. The magnitude of these

markups among all services must be sufficiently high to earn net revenues that cover fixed and common costs and, hence, achieve revenue adequacy.

The logic of this inverse elasticity rule and its implied allocation of unattributable costs is quite simple. The elasticity of demand provides a quantitative interpretation of the traditional concept of value of service, which has played an important role in public utility pricing. Consumers who place relatively high value on a service will have demands for it that are relatively inelastic, and vice versa. If a rise in the price of a service would lead to no significant reduction in quantity demanded (that is, if demand is inelastic) then the service must be worth at least the higher price to its consumers, that is, the value of the service must be high. Conversely, if a rise in the price of a service would lead consumers to curtail their demand substantially (that is, if demand is quite elastic), then the service must be worth little more to its consumers than the original price, that is, the value of the service must be low.

In view of this correspondence between value of service and demand elasticity, the inverse elasticity rule of Ramsey pricing can be restated in terms of a familiar and long-used principle in railroad pricing. Services with relatively high values to their consumers should contribute relatively large net revenues to the coverage of unattributable, fixed, and common costs. Thus, the implicit allocation of unattributable costs should be based on value of service rather than any pro rata sharing or other arbitrary method. All factors that influence a rail carrier's elasticities of demand are relevant for the carrier's Ramsey prices. These factors may include the value of the commodity shipped, intermodal competition, intramodal competition, interport competition, and the substitutability of other commodities for the one shipped at its destination. Value of service is therefore properly construed as a market concept. It refers to the value of the rail carrier's service with all demand factors considered and generally cannot be evaluated by such measures as the ratio of a commodity's price to its weight alone.

#### *The efficiency and equity of Ramsey pricing*

Under Ramsey pricing, the "nonmarginal" portion of total costs (that is, the total cost less the marginal cost of each service multiplied by the quantity of the service provided) is apportioned on the basis of demand. Equivalently, the nonmarginal portion of total costs is the shortfall between total costs and the revenues that would accrue from pricing each service at the level of its marginal cost. In the presence of economies of scale, this

shortfall is positive. Ramsey prices, therefore, deviate from marginal costs only to the extent necessary to provide adequate revenues. They thus permit the railroad to achieve the goal of revenue adequacy with less sacrifice of economic welfare than does marginal cost pricing.

Increases above marginal cost in the price of an elastic service cause much traffic to be lost—traffic that would generate net benefits because it is valued above the cost it causes. However, less traffic is lost when the price of an inelastic service is raised, and the traffic that is curtailed is the least-valued portion. Consequently, when prices must be elevated above marginal costs to cover unattributable costs, it is economically efficient to increase the prices of inelastic services more than the prices of elastic ones. Such Ramsey prices are, on average, the lowest consistent with financial viability. As long as the price charged to the price-elastic service exceeds its incremental cost, the service is contributing to the carrier's overhead costs. Thus, Ramsey pricing principles benefit all shippers by establishing a set of rates that encourages the purchase of more transportation services by more shippers than would artificial prices based on fully distributed cost. By creating a larger traffic base over which unattributable costs can be apportioned, Ramsey pricing also benefits the so-called captive shippers; the expansion of rail traffic represents an increase in the flow of commodities to their markets at lower transportation costs. As a result, social productivity is enhanced, and more consumers can obtain more of the goods they desire at lower costs of supply.

Since Ramsey prices are based on the relative values of the different services, they may seem to approximate the solution of the profit-maximizing monopolist, sometimes loosely described as "charging what the market will bear." However, only the firm's necessary costs, including the cost of capital, are covered by Ramsey prices. Monopoly prices, by contrast, are controlled by no such constraint. Ramsey prices therefore are very different both qualitatively and quantitatively from monopoly prices.

It should also be emphasized that Ramsey prices are equitable. First, they are nondiscriminatory in the sense that services with similar economic characteristics have similar prices, regardless of the commodities shipped, the route, or the identity of the shipper. That is, two different services with the same elasticities of demand will be priced at the same percentage markups above marginal costs. And two different services with the same marginal costs and demand elasticities will bear identical Ramsey prices. Second, while the Ramsey prices of different services are different proportions of the services' marginal

costs, the burdens from these necessary markups that are borne by the consumers have roughly the same proportion to their respective values of service.

#### *The stand-alone cost constraint*

Ramsey pricing requires that both the marginal cost and the elasticity of demand be quantified for every movement in the carrier's system—which is all but impossible to do with any degree of accuracy. The amount of data and the analysis required are overwhelming. Thus, while the Ramsey formula is useful as a theoretical guideline for rate determination, a valid criticism is that its application would be administratively difficult and burdensome.

The Ramsey pricing rule has also been criticized because it does not constrain the railroad's pricing of traffic over which it possesses market dominance and thus fails to protect captive shippers. In addition, although Ramsey pricing minimizes the static welfare cost of the revenue adequacy constraint, output levels still are less than they would be if rates were set at marginal costs. This situation results in economic inefficiency because the value of the lost output to the shipper is greater than the value of the resources saved by reducing output. Under these conditions, it may be feasible for the parties to negotiate a contract with incentive clauses, volume-sensitive pricing, or two-part pricing, which would leave both parties better off than at the flat Ramsey price and consequently be yet more desirable for the public interest.

The critical issue from the standpoint of efficiency is the criterion used to set the ceiling on rates where there is market dominance. As noted above, rate ceilings derived from fully distributed costs are inimical to the public interest. Economically rational ceilings are obtainable from the stand-alone cost. The stand-alone cost of serving any captive shipper or group of shippers that benefit from sharing joint and common costs is the cost of serving that shipper or group of shippers alone, as if the shipper or its group were isolated from the railroad's other customers. A rate calculated by the stand-alone cost methodology represents the theoretical maximum rate that a railroad could levy on shippers without substantial diversion of traffic to a hypothetical competing service. Thus, the stand-alone cost criterion serves as a surrogate for competition and leads to a simulated competitive price. The competing service could be either a shipper providing rail service for itself or a third party competing with the incumbent railroad for the traffic. In either case, the stand-alone cost represents the minimum cost of a possibly hypothetical alternative to the service provided by the incumbent railroad.

*Stand-alone costs: protection against excessive rates*

The stand-alone cost test rules out the possibility of abuse of monopoly power by enforcing a competitive standard on railroad rates. The hallmark of monopoly power is the elevation of the price of a service above the costs at which competitors could provide that service. The stand-alone cost test makes that impossible and imposes the same ceilings on rates for any traffic over which the railroad is dominant that the market would impose if it were subject to either active or potential competition. In the long run, in contestable markets, no group of shippers would agree to pay a carrier more for their transportation services than it would cost them to produce these services for themselves or more than it would cost a competitor to supply the services to them. In the short run, a rail carrier facing either active or potential effective competition could not obtain revenues from a group of shippers that exceeded their stand-alone costs, because those shippers could then be profitably served by a competitor charging lower rates. Thus, the stand-alone cost test affords shippers the same protection that effective competition would provide.

Clearly, the stand-alone cost is unnecessary and inappropriate where there is competition. In a competitive market, the price set by competitors (reflecting current costs of service) will set a market ceiling. If only potential competition exists, the regulatory test is still unnecessary; if the rates charged by the existing carrier exceeded stand-alone costs, that fact would constitute an invitation to entry by the potential competitors. However, for any shippers that are truly captive, which is to say that the rail carrier faces no effective direct, indirect, or potential competition for their freight, the stand-alone cost does provide an economically rational ceiling.

No regulatory ceiling is needed to act as a surrogate for active or potential competition from a mode that can operate through the market. Market pressures will enforce the stand-alone cost ceiling since no one will be able to sell at a higher price. Yet another consideration reduces further the likelihood that regulators will have to intervene, except on the rarest occasions, to enforce stand-alone cost ceilings on rates. This consideration stems logically from the very concept of stand-alone cost. If the rates for any service exceed those necessary to cover stand-alone cost, that fact by itself invites the sort of competition that automatically prevents the continuation of such excessive rates.

The stand-alone cost test does not apply, and cannot be made to apply without disastrous consequences, if railroads are denied the freedom to abandon unremunerative facilities or services. Without such freedom, a railroad

cannot earn adequate revenues if it is constrained by stand-alone cost ceilings on rates in the potentially remunerative portions of its activities. For this reason, any public policy that limits the freedom of railroads to curtail unremunerative services must also provide public funds to help defray the costs of those services.

The stand-alone cost ensures the equitable treatment of all of a railroad's shippers. By requiring each service or each group of services supplied by a rail carrier to contribute revenues less than stand-alone costs, the test ensures that each shares in the benefits derived from the economies of scope resulting from simultaneity of production. Thus, each shipper is guaranteed some benefit from the revenue that the carrier collects from others. The stand-alone cost offers assurance to each shipper that it will be better off with the existing rates than it would be if it had to fend for itself, as it would have to do in the long run if the rail carrier were denied adequate rates.

If the price paid by a shipper is no greater than the stand-alone cost of that service, then that price cannot possibly contribute to the cost of any facility from which the shipper derives no benefit. This must be true because the stand-alone cost of any facility used by a shipper includes only the (replacement) cost of those facilities after subtraction of any contributions made by any other railroad customers toward the cost of these services. Thus, together, all the customers that share the use of some facilities will provide revenue contributions that do not exceed the costs of the facilities they use. There will be no excess that the railroad can use to defray the cost of unused facilities. The stand-alone cost test therefore precludes cross-subsidies among the railroad's different customer groups.

The absence of cross-subsidies under the stand-alone cost test is an appropriate and accepted criterion of equity in the treatment of shippers. Cross-subsidies are properly of public policy concern because they generally lead to a misallocation of resources by encouraging inefficient investment. For the shippers, cross-subsidies may be of concern because they are perceived as unfair. If payments of one group of shippers help make up for shortfalls in payments by another, the first group might well believe it is being forced to cross-subsidize the second. Yet mere payment of a relatively higher rate is not evidence of a cross-subsidy where fixed and common costs must be covered. Rather, a cross-subsidy in an economic sense can occur only if a shipper (or a group of shippers) pays more than the total cost of serving it alone. If no shipper pays more than that amount, differences in their rates simply reflect differing contributions to the common costs of the system, not cross-subsidies.

Imposing stand-alone cost as a rate ceiling is a form of incentive regulation that avoids introducing distortionary incentives to the railroad with respect to its operations and costing decisions. Since the stand-alone cost is the cost of service by a hypothetical entrant that offers alternatives to the shippers at issue, it is not determined by any of the costs actually incurred by the regulated railroad.<sup>3</sup> Consequently, under the system of stand-alone cost rate ceilings, a railroad has no incentive to pad or otherwise increase its expenditures for the purpose of relaxing a regulatory constraint. Further, since the ceilings apply only to services over which the railroad has monopoly power, they do not interfere with the railroad's incentives to pursue aggressively additional traffic and other new business opportunities. Finally, while stand-alone costs may be calculated on the basis of detailed engineering studies and judgments, it is significant to note that they are consistent with the "price caps" that are becoming so popular today, inasmuch as they can be periodically updated on the basis of net measures of inflation and changes in productivity.

#### *Efficient pricing and regulatory control*

For prices to be efficient, they must reflect implicitly all of the interdependencies that characterize a rail network. This could be taken to imply that to institute efficient prices for one segment of a railroad's activities (that requires regulatory oversight), it would also be necessary to regulate the prices for all of the railroad's other services. Convincing evidence that such a conclusion is unfounded is provided by the workings of the free market in unregulated industries. In such industries, although there exists no authority that coordinates pricing decisions, compatible and efficient prices nonetheless emerge, their consistency ensured by the forces of competition. This is precisely why free and unplanned markets perform so effectively in comparison with those operated by central planners, despite the latter's alleged ability to take interdependencies into account.

It is for this reason that no regulatory control need be exercised over the rates of competitive services. Here, efficient prices are automatically imposed by the market, and regulatory intervention can only impede the efficiency of the process of rate determination and resource allocation. Also, relatively little control need be exercised over rates set by a carrier whose revenues are still short of adequacy. If total revenue is not yet adequate, the best rates in terms of the long-run public interest are those that maximize the railroad's net revenues—Ramsey prices. Any railroad with inadequate revenues has powerful incentives to select such rates. In such a case, the railroad

as a whole possesses no monopoly power that offers it excessive profits; for individual services for which competition is inadequate, the stand-alone cost test provides the requisite protection to shippers. Under these conditions, there is no possibility of unfair competition through cross-subsidy, with noncompetitive rates increased in order to permit noncompensatory prices in competitive markets. For where the railroad's overall revenues are inadequate, any internally subsidized service must be a drain on the railroad's already insufficient revenues and therefore self-destructive. Thus, where overall revenues are inadequate, only the stand-alone cost test need ever be used in the regulatory oversight of rate setting.

More than this minimal regulatory scrutiny may conceivably be required if (and only if) a railroad is in a position to earn revenues that are more than adequate. Here, there is at least the hypothetical possibility that high prices for one service will be traded off for price reductions in another. Consequently, it may be desirable to devote regulatory attention to prices for services sold on markets from which competition, direct or indirect, actual or potential, is absent. Yet even here, the railroads have incentives to select the efficient Ramsey prices. That is, the interests of the railroads are still likely to be served best by the prices that best serve the public interest—although it must be admitted that the incentives to select Ramsey prices are apt to be somewhat less powerful than those in the prevailing case of insufficient revenues.

There is one principal source of incentives for a carrier capable of earning adequate revenues to adopt efficient pricing, even though its net revenues are constrained by regulation to cover only its capital costs and no more. Such a rail carrier is motivated, perhaps more than other firms in similar circumstances, to maintain its traffic base and to guard against substantial diversion of its traffic to suppliers already in operation or to potential competitors. This is because a large portion of a rail carrier's capital stock is nonfungible, or sunk, so that significant losses of traffic would cause losses of revenue far greater than the costs that would thereby be saved. Consequently, a rail carrier with adequate revenues has a particularly compelling incentive to set rates in a manner that will discourage defections of shippers and market erosion to competing suppliers of transportation services, in both the short and the long run. It may be clear intuitively that among the pricing policies that generate adequate revenues, Ramsey pricing most effectively discourages such defections and market erosion. This is true simply because at any one time the Ramsey prices yield shippers the greatest total net benefits possible from prices that

yield adequate revenues, and therefore offer shippers the smallest feasible inducement to divert their traffic.

In sum, regulation need not take on the overwhelming task of controlling all of a railroad's rates, simply to ensure an appropriate choice of prices in those circumscribed arenas requiring regulatory attention. Elsewhere, the forces of competition and the self-interest of the railroads constitute powerful mechanisms that can do the job efficiently and automatically using the crucial demand information possessed by the railroads—which is certain to be more complete and more accurate than any demand data a regulatory agency could hope to assemble.

#### *Contestability and the scope and structure of regulation*

Contestability is an apt benchmark for the railroad industry. By contrast, the familiar benchmark of perfect competition is neither attainable nor desirable for the railroad industry, in which economies of scale and scope are substantial. In this industry, attempts to approximate perfect competition may in fact be highly inefficient and contrary to the public interest. In any case, the theory of contestable markets demonstrates quite clearly that neither large size nor small number of firms necessarily means that markets need function unsatisfactorily. Indeed, a variety of market forms far removed from perfect competition may perform well for the public interest so long as such markets are structurally contestable. If an industry is contestable, it is best left alone without government interference, even if it is composed of a very small number of large firms. Impediments to entry and exit, not concentration or scale of operations alone, are a primary source of interference with the public interest workings of the invisible hand.

Contestability focuses increased attention on entry barriers and their defining characteristics. High fixed costs and the consequent economies of scale, for example, have traditionally been considered as impediments to entry; contestability analysis shows, however, that they need not permit excessive profits or prices or any of the other manifestations usually associated with market power. It is the presence of sunk costs rather than economies of scale that is of vital importance for market performance.

The theory of contestability offers an improved set of rules to be followed by regulators in those cases in which their intervention is appropriate. In addition, it provides economically sound criteria for distinguishing between cases in which intervention by the public sector is warranted and those in which it is not. The theory of contestability is the framework from which the following precepts for railroad regulation (discussed above) were derived:

- Permit a private sector railroad to have freedom of pricing and operations on services that face effective competition in the relevant market, whether from other railroads, other transportation modes, other origins, other destinations, or other commodities.
- Permit a railroad to set prices that are responsive to differences in demands, as well as to differences in marginal costs, and further to enter into voluntary contracts with shippers that have individualized terms, conditions, commitments, and compensation mechanisms.
- Constrain the prices that a railroad sets to captive shippers over whom the railroad has monopoly power, by the stand-alone costs of the shipper's service (or by a comparison of the revenues and stand-alone costs associated with any larger group of shippers' services) and by the stipulation that the railroad's prices do not generate earnings that persistently exceed the railroad's replacement costs, including a competitive return on capital.

In addition, contestability is a fruitful framework for the analysis of issues pertaining to the vertical structure of an industry. For one thing, in a perfectly contestable market, survival against potential competition requires a firm to undertake efficient vertical relationships and to structure itself efficiently along vertical as well as horizontal and conglomerate dimensions. For another, contestability theory suggests consideration of the idea of separating firms vertically in order to segregate the portions that need regulation from those that do not because of their degrees of competition or contestability.

This idea emerges from the application of contestability theory to regulatory policy where sunk costs are not pervasive in an industry but centered in a particular sector of its operations, such as the track, way, and structures in railroading. By isolating the activities with which the heavy sunk costs are associated, their need for regulation can be quarantined. By placing relations with the remainder of the industry at arm's length, to the extent permitted by economies of scope, it may be possible to leave the operations of the bulk of the industry safely to the free market, permitting open entry and more flexible pricing, and to draw a regulatory net over only the segment of the activities that are inextricably associated with heavy sunk costs. Thus, contestability suggests a flexible, case-by-case regulatory approach.

#### **Options for vertical railway restructuring**

The historical model of railway operations is the monolithic organization: A single entity controls all facilities and operating and administrative functions and determines what services to provide to significantly captive

markets. This railway is an integrated entity that owns and operates its own facilities and vehicles. Typically, the monolithic entity lacks financial incentives and desegregated information on profitability, and is (at best) production-oriented, unresponsive to market demands for services, and hierarchical (if not bloated) in organizational architecture.

*The need for restructuring*

Although no one would deliberately choose the monolithic railway structure from the standpoint of public interest, it has nevertheless been chosen all too often, either for private interests in monopoly control or for the political benefits that could be collected and disbursed through a state-owned monolithic railway. It is predictable that a state-owned railway enterprise would fail to be responsive to the needs of shippers and would instead be politically responsive, at the expense of providing efficient operations and a stimulus to the economy (see Willig 1994). It is equally predictable that a privately owned railway that was exposed to excessively controlling and economically arbitrary regulation would also lack incentives for efficiency and market responsiveness. Financial deficits would be a natural consequence, as the railway entity failed to attract traffic from alternative modes and geography, as it expended inefficiently on costs, and as it allowed its facilities to suffer from deferred maintenance and replacement.

The conditions that generated the monolithic railway model no longer exist in most countries, and governments have had to consider fundamental restructuring of both the railway entity itself and the relationship between the railway and the state. The objectives for such restructuring have properly included injecting more innovative and efficient management, reducing railway deficits and burdens of public subsidies, increasing competition with other transport modes, and improving responsiveness to the needs of emergent capitalist enterprises.

Four generic options can be identified for the vertical restructuring of railways, addressing the set of relationships between the railway entity and other transportation entities (both rail and nonrail), the markets served, and the functions performed. These functions include ownership, improvement and maintenance of the fixed facilities, control of operations such as dispatching and freight classification, train movement, equipment provision and maintenance, marketing, and financial control and accountability.

*Option 1: Lines of business organization.* Railway entities can be reorganized and accorded financial responsibility

for lines of business to foster comprehensive business planning, market-sensitive and cost-sensitive decisions, and greater responsiveness to demand for various services. British Rail, for example, has divided itself into five lines of business that are financially accountable to top management and that “purchase” service by contract from an operating department that is organized along a matrix of regional and functional lines. By so doing, British Rail hopes to give commercial sectors a profitability objective and to give noncommercial lines of business incentives to reduce their losses.

*Option 2: Competitive access.* Under this option competing railway companies would have exclusive control over some trackage but also have (and give) the right of competitive access over the trackage of (to) other companies. Some forms of competitive access include joint terminal agreements and conferrals of trackage rights, whereby one railway obtains the right to use the freight-handling facilities or line haul tracks of another railway at a particular location or along a particular route. A further characteristic of this option is arrangements for interlining traffic that is handed off between distinct railroad entities, in their preference sometimes to utilization of trackage rights. In the United States, railroads do a great deal of interlining under terms that are largely unregulated, perform reciprocal switching under terms that are subject to regulation, and exercise trackage rights that are sometimes freely negotiated and that sometimes result from regulatory mandates (that were mostly put into place in the context of settlements of disputes over rail mergers).

*Option 3: The “wholesaler.”* Under this option the railway entity would own and operate the fixed facility and perform all operations on behalf of marketing entities that would be the “retailers.” The railway itself would only haul trains, but it would do no marketing to shippers. In Australia, for example, freight forwarders function as retailers using the state railways’ “wholesale” services. These forwarders provide multimodal transport and conduct a deregulated trucking business. They control their own rail terminal and yard operations and negotiate on the open market with the railways to charter unit trains with agreed-upon service specifications. This structure permits competition among efficient intermodal “retailers” to flourish, despite a state or private monopoly on railway ownership.

*Option 4: The “toll rail” enterprise.* Under this option the entire fixed facility, except for exclusive facilities, would be the property and responsibility of one owner. There could be one or more authorized users, which

would pay tolls for use of the facility. This approach differs from the competitive access approach (option 2) in the following respect: Under the toll rail approach, separate entities provide the fixed facility and conduct operations, whereas under the competitive access approach, more than one entity operates in a given market over a particular fixed facility. Sweden has implemented a separation of fixed facility from operating functions since 1988. The United Kingdom recently moved in this direction by establishing a separate entity to hold and manage the rail system's assets associated with the track and road bed. And the European Union has articulated a policy principle that urges its members to move in the direction of separating rail operations from fixed facilities.

It is clear today that a railroad organized and controlled according to the monolithic model must be restructured in order to contribute best to the economy and to avoid being a significant impediment to growth and prosperity, becoming responsive to shipper needs and demands, as well as to marketplace opportunities for innovation. One key element of restructuring is to develop internal organizations of rail entities that provide managerial incentives, information, and decision-making decentralization that contribute to efficiency, market responsiveness, and fiscal responsibility. Thus, option 1 is certainly crucial for restructuring, whatever else is also entailed. It should be recognized that although an internally restructured railroad enterprise may show lower technical operating efficiency by some traditional measures (for example, coach-kilometers per locomotive-kilometer), it may succeed in making each service more responsive to customers' needs and willingness to pay. Economic productivity and customers' interests are best promoted by minimum total logistics costs, not the lowest railway rates accompanied by minimum service quality.

Another key element of restructuring is to unleash the forces of competition to the fullest extent. It is difficult to predict what are efficient and market-responsive vertical relationships and combinations of logistical roles among various rail entities, truckers, barge operators, port operators, warehouses, forwarders, and other players. The U.S. experience confirms what theory predicts: Decentralized, market-oriented decisionmaking that is freed from excessive regulatory control and energized by market incentives is the surest means of finding and implementing efficient, innovative solutions to the problems posed by transportation needs (see, for example, Baumol and Willig 1987).

Options 2, 3, and 4 are approaches to restructuring that have the potential for bringing more competition and more market decisionmaking into the domain of railroading and its vertical relations. Which of these options is the best choice is a complex policy decision with many important dimensions that must be considered. The analyses in the two subsections that follow may clarify some of the important considerations.

#### *Analysis of structural separation*

Options 3 and 4, which separate ownership of facilities from other rail functions such as train operations and marketing, have generated much attention of late and deserve serious analysis. These options have considerable appeal because they seem to mitigate the difficult problems blocking comprehensive rail deregulation that are associated with the roadbed costs, which are largely sunk. Fixed costs are large because of the infrastructure (track and stations, for example) that must be provided before any trains can run on a route. Because duplication of infrastructure would generally be inefficient, natural monopoly cost conditions characterize physical network provision. These fixed infrastructure costs are largely sunk because the assets are of minimal value for other purposes. For example, embankments and cuttings, the rail formation, and the platforms are fixed in place and committed irreversibly to a specified market. The sunk nature of infrastructure costs creates significant entry barriers, especially where natural monopoly conditions also exist.

The cost conditions relating to the operation of services on the physical network, by contrast, may be more consistent with active and potential competition. To operate a service it is necessary (at least) to have trains, staff, support, and rights-of-way. Although there are inevitably some sunk costs in hiring staff and buying or leasing rolling stock, they are small in relation to the massive sunk costs of establishing network infrastructure. Locomotives and freight cars constitute capital on wheels, and most of their cost might be easily and quickly recovered by rolling them to other markets.

If ownership of track and trains were separated—with the track assets held by the government, by a consortium of the operators, or by a regulated private entity—there might be vigorous active and potential competition over railway services provided by operators with equal access to the utilization of the roadbed. There would be no need to regulate these operators, who would have all the powerful incentives that accompany competition to be efficient and responsive to the needs of shippers and a growing entrepreneurial economy.

Several links in this chain of policy reasoning, however, may be inapplicable or wrong in a given set of realistic circumstances:

- The provision of many innovative and market-responsive rail services may require specific investment in infrastructure, such as maintaining or upgrading way and structure facilities, constructing loading and transshipment facilities, and building spurs of track to reach a shipper's location. It may be difficult and inefficient for any operator (or retailer) to coordinate, as necessary, with the infrastructure monopoly (or wholesaler) entity, especially if their incentives with respect to investment behavior are not in harmony. The investment incentives of the infrastructure monopolist (or wholesaler) will, of course, depend critically on whether it is a state-owned entity or, if it is in the private sector, on the character of its regulation.
- Efficient, safe, and delay-minimizing utilization of track and yard facilities by trains, cars, and shipments requires close coordination in accordance with priorities that are driven by considerations of both operations and shipper sensitivities. Rival operators (or retailers) will compete vigorously and acrimoniously over scarce or congested infrastructure facilities (or wholesaler services). Constantly sorting out their claims will be important for the overall efficient and responsive operation of the rail system. This task would be difficult enough for an unintegrated system with a monopoly infrastructure entity; it seems virtually impossible to accomplish efficiently where there exist rules against discrimination and infrastructure (or wholesale service) pricing that is either tightly regulated or, for a state enterprise, politicized.
- The freight hauling operations on all or part of the rail system in question may well constitute a natural monopoly, even when disintegrated from the infrastructure. The economies of scale and scope that arise from running long trains, from blocking many different shippers' freight in classification yards, and from efficient utilization of yard facilities, crew, and rolling stock are all associated with operations rather than infrastructure. Consequently, a separated operations firm may be a monopoly, and it may have considerable market power unless potential competition is a powerful force.
- For potential competition to be powerful, an entering operator must perceive that significant sunk investments in rolling stock and in specialized facilities can be avoided. Locomotives and freight cars may indeed be an example of capital on wheels so long as they can be transported to alternative points of gainful utilization without substantial costs. While this is likely to be the case for services provided in the middle of a landmass with a rich rail

network ready to accommodate the cars, it may not be the case for more specialized cars or for a more isolated market. Also, the entering operator may not have yard, loading, car maintenance, or spur facilities available unless it makes new and significant sunk investments. For these to be available on equal terms with the incumbent operator, it must be the case that the infrastructure entity made the needed investment as part of its role in the system. But the greater the entrepreneurship and risk-taking investment that the infrastructure entity (or the wholesaler, under that option) must undertake, the less is gained by the separation, since the infrastructure (or wholesaler) entity is either a state-owned or a tightly regulated private sector monopoly.

- Efficient pricing to cover replacement costs is made more difficult by separation. Where economies of scale are important, efficient pricing to cover replacement costs requires that shipments of different commodities on different origin-destination routes bear prices with different relationships to marginal costs. If it is the case that the operator (or retailer) firms can readily evade price discrimination by the infrastructure entity (or wholesaler)—so that different prices cannot be collected by the infrastructure entity (or wholesaler) for facility utilization (or for wholesale service utilization) by different shippers of different commodities—then it will be difficult if not impossible for the costs of the infrastructure to be defrayed by Ramsey prices. At the extreme, a regulated infrastructure (or wholesaler) entity charging competitive operators (or retailers) an equal price for each ton or each ton-mile of freight that utilizes each of its facilities is, in essence, recreating a system in which prices are set according to fully allocated costs. As discussed above, such pricing can be a prescription for inefficiency and financial disaster.

Thus, it is clear that separation of operations from infrastructure in a railroad system is no panacea for regulatory problems. Instead, as a policy direction, it must be compared with the leading alternative.

#### *Analysis of competitive access*

Option 2 is most clearly distinguished from the separation options just discussed by the fact that the competitive access option allows integrated operations by the rail entity. It is superficially easy, albeit mistaken, to identify an integrated carrier with the case of the monolithic carrier, because it is tempting to jump to the conclusion that an integrated carrier would make it difficult for other entities to participate in its business. This option implies a requirement that the integrated carrier make its facilities

available to other entities on a “fair and equal basis.” However, if the integrated carrier has strong incentives to keep other entities out, it is unclear how effective such equal access mandates are likely to be. In the United States the rail industry, like other regulated industries (for example, gas pipelines, telecommunications, and electric power), has seen many disputes with claims of “unfair” and “unreasonable” exclusion from a carrier’s facilities, despite rules of equal access.

Thus, an assessment of this option must include an analysis of the incentives of the integrated carrier to accommodate others wishing to participate, and able to participate efficiently, in the provision of service. If the integrated carrier is regulated in a fashion that permits the carrier to charge higher prices to captive shippers if it does more of the business, then the carrier clearly would have incentives to exclude other participants.<sup>4</sup> Likewise, if the integrated carrier is constrained by regulation in the amount it can earn from the portion of service it provides when it does cooperate with another entity, then the carrier has incentives to undermine or avoid efficient cooperation in order to enlarge its portion of service.<sup>5</sup> In addition, the integrated carrier would be motivated to exclude an efficient participant if by so doing the carrier would weaken, in a predatory manner, the competitive impact of that entity in another market. Under classic rate-of-return regulation or under a system of regulated “divisions” specifying what an integrated carrier can earn from a cooperative movement—both features of U.S. rail regulation at one time—an integrated carrier does have incentives to undermine efficient cooperation.

In sharp contrast, under the regulatory system that has been described above as serving the public interest well, an integrated carrier would generally have a real profit motive to cooperate with an efficient participant in its business. Here, it is not “divisions” that are specified by regulation, even on the service provided to a captive shipper. Instead, the described stand-alone cost rate ceiling applies to the price charged to the shipper, and cooperation with an efficient entity enlarges the pot of returns available from the service, enabling more money rather than less to be earned by the integrated carrier. Consequently, except for the rare possibility of predation, an integrated carrier would have ordinary business incentives to find and cooperate with efficient participants in its business and to negotiate mutually beneficial terms. This is just a railroad version of business “make-or-buy” decisions in other industries.

Despite the prevalence of efficient incentives on the

part of integrated carriers under the form of regulation described here, it is useful and wise to augment the system of regulation with a fallback set of standards to apply should disputes about predation through competitive access arise. In short, an integrated carrier that possesses a “bottleneck”—a facility without which the complainant cannot reasonably offer its services to the shipper—should not exclude the complainant by refusing an agreement that would be fully compensatory of all its costs, including opportunity costs.<sup>6</sup> For example, if another carrier, or an operator, sought to participate in a freight movement that represented new business for the integrated carrier, then it is to be expected that the latter would negotiate in good faith and not exclude the other entity if an agreement could be found that would at least cover the incremental costs of the integrated carrier. If another carrier sought to handle some freight part of the way that the integrated carrier would otherwise handle itself, then it is to be expected that the integrated carrier would accept an agreement that earned it a larger net contribution of revenues above incremental costs than it would earn if it handled the freight without the other participant. Here, the contribution that the integrated carrier would earn on its own is part of the opportunity costs it faces from cooperating with the other participant. These same principles apply to interlining, trackage rights, car hire, or any other form of cooperation or participation through the use of a bottleneck.

“Efficient component pricing,” or “parity pricing,” are alternate names that have been given to the principle that an integrated carrier should offer the services of its bottleneck at a price that yields it the same contribution that it would earn from performing the end user’s service itself. Behavior consistent with this pricing of bottleneck services, or more generally with the antipredation rule just articulated, leads to efficient vertical relations and is thereby consistent with nonpredatory incentives under the regulatory system we have described. Such pricing of bottleneck facilities does not place additional competitive pressure on pricing to shippers, since it is based on the contribution that could be earned from the shipper’s service at the extant shipper’s price. However, it does generate incentives for efficient combinations of transport services to make it to the market; it does provide quality and cost competition among potential and actual participants for the role of being part of the efficient combination; and it does help to ensure that those with efficient innovations in logistics or in marketing of transport services will be able to work with carriers to implement their ideas.

*Separation versus competitive access*

The primary virtue of separation as a policy option is that it may permit active or potential competition to reign among rail operators or retailers—with corresponding assurance of efficient selection among them for provision of their services at efficient prices. At best, separation will accomplish this end, but it will leave unresolved the difficulties with regulating the provision of the services of the infrastructure, or bottleneck, assets of the railroad network. Prices charged to shippers will be at least the sum of the competitive prices for the services of the operators (or retailers) and the regulated prices for the services of the infrastructure entity (or wholesaler). They are unlikely to be fully Ramsey-efficient prices for the coverage of replacement costs because of the difficulties of reflecting shippers' differences in demands in the prices charged for infrastructure services. At the same time, separation may create serious coordination problems, loss of economies of scope, and otherwise unnecessary transactions costs. In addition, rail operators may not face effective active and potential competition, undermining the potential for realizing the primary benefit of the option.

The competitive access option could also be fraught with problems when the incentives of bottleneck holders are adverse to efficiency and competition. A variety of solutions to competitive access problems have arisen in industries seeking to replace regulation with competition. Typical examples include mandatory interconnections with competitors and line-of-business restrictions in the telecommunications industry, “unbundling” of the transportation and energy components of price in natural gas markets, and equal access to marketing channels (for example, computer reservations systems) in the airline industry. In designing rules that govern vertical relationships among competitors formerly subject to economic controls, regulators must address a common basic problem: How to implement pricing and terms of access by “nonintegrated competitors” to the restricted portions of the network so that competition on the merits will work to ensure that the efficient alternatives successfully participate in the provision of end users' services. The compensation for and terms of access should not distort the process by which prices are adapted to consumer preferences and demands for transportation service. Prices should be sufficiently high to be compensatory to the “landlord” railroad yet not so high as to preclude efficient operations by the “tenant” railroad. Where incentives are significantly adverse to these goals, experience teaches that rules are too easily evaded and disputes seemingly never-ending.

It is thus fortunate that under rail regulation that focuses on the levels of rates charged to shippers—rather than on other prices, such as those charged for access to bottleneck services—incentives are generally for the promotion of efficient vertical relationships. As a result, if integration is permitted under this system of price regulation, then the outcomes are predictably consistent with efficient participation by the integrated carrier and by other, nonintegrated carriers as well, on terms that permit compensatory support for the efficient participants. Further, prices to shippers can be selected in accordance with Ramsey efficiency, even as they are constrained by regulation where the carrier has monopoly power. Moreover, unlike the virtues of separation, the efficiency of the outcomes of competitive access does not depend on the absence of economies of scope, on the absence of coordination problems without integration, and on the competitiveness or contestability of rail operations.

Separation of track assets from operations is likely to be a particularly attractive option where a dense and extensive rail network permits many operators to function and to provide both active and potential competition to each other. Another favorable factor is a mature and well-developed set of fixed facilities, so that there is relatively little extent to the domain of new infrastructure investments, where incentive problems are more likely to arise. Where this factor does not apply, it will be important for regulation of the infrastructure entity to permit it to enter into medium- or long-term contracts with shippers or with operators that themselves have contracts with shippers, so that the risks and rewards from investments can be efficiently shared by shippers, operators, and the infrastructure entity. The impediments to Ramsey pricing that separation might cause would be rendered insignificant to the extent that the infrastructure entity does not attempt to recover its sunk capital costs from “tolls” levied on traffic. If the infrastructure entity is expected to seek recovery of its replacement costs, then it should be permitted and even encouraged to implement forms of price discrimination that help to bring shippers' prices in line with principles of Ramsey efficiency. Finally, there may well be circumstances where a monolithic railway system cannot be converted to one with functioning competitive access because of embedded business culture and entrenched management. Here, the act of separation is so revolutionary that it may unsettle the business culture in a productive fashion and force reassignments of management that permit implementation of the necessary internal reorganizations of responsibilities, roles, incentives, and information flows.

### Concluding remarks

This chapter has outlined a set of principles that together add up to a program for restructuring the relationships between government and railroad entities. These principles point toward a great deal of reliance on market forces to shape prices and logistics of services. At the same time, the principles include economically appropriate protections for any captive shippers and for any carriers that may be excluded or foreclosed from participation for anti-competitive reasons.

On the subject of restructuring, we have pointed out that internal managerial reforms are necessary, as are policies that address railway vertical relationships. The two leading candidates—separation of track from operations in different business entities, and incentives and fallback rules for competitive access—were compared on several dimensions, and their relative levels of appeal were found to depend on a variety of characteristics of the business environment.

Restructuring along the lines suggested here, providing a greater emphasis on marketing effectiveness, can be expected to result in a more profitable railway that is better able to cover the costs of its commercial services. Any noncommercial services that are needed should be carried out on the basis of an explicit agreement between the railway and government that views public service obligations as a business relationship between a customer (government) and the contract supplier (railway). This agreement would help to ensure that noncommercial services are more effective in fulfilling public policy objectives, remove an insuperable drain on revenues that would condemn the railroad to insufficient investment, and eliminate cross-subsidies that hamper the railroad from maintaining its competitive position against other modes.

### Appendix A

#### Technology and the structure of railroad costs

The output of the rail industry is multidimensional. Railroad firms produce different types of transportation services for different users at different origins and destinations at different times and at different levels of quality. The mix of output and shipment characteristics can have a major impact on the costs of any given firm. For example, railroads specializing in coal traffic have very different cost characteristics than those specializing in movements of general manufactured commodities.

The most striking feature of the cost structure of railroads is the high incidence of costs that cannot properly

be attributed to any particular service at any particular point in time. That is, a significant portion of costs are incurred on behalf of several activities and do not vary with the amount of the service provided. These unattributable expenditures reflect both joint and common costs. Common costs are costs shared by two or more services in variable proportion. For example, a terminal represents a common cost; it is used by different services in varying proportions. More generally, the outlay on track and way and structures between points A and B is a common cost for all movements of whatever commodities are shipped between A and B over that route. Joint costs are costs shared by two or more services in fixed proportions. A backhaul movement is the classic railroad example.

The structure of railroad costs has important implications for the competitive structure of rail markets. It is sometimes mistakenly inferred from statistical evidence of constant returns to firm size that a competitive equilibrium with marginal cost prices covering total costs would be sustainable in the rail industry. Such reasoning neglects the critical fact that indivisibilities in rail technology make increasing returns to scale in total costs endemic and competition among a limited number of players inevitable. A rail link between two points requires lumpy investment in way and structures with associated highly significant economies of traffic density. Unit costs fall with output, letting all factors of production vary on a given route or route structure.

#### *Fixed and variable costs*

A fixed cost is one that is necessary to provide a service or group of services, and whose magnitude does not vary with changes in the quantity of a service provided. For example, if a railroad is to run between points A and B, a minimum outlay on track and roadbed must be incurred, even if the trains run virtually empty. Even in the longest run, the roadbed cost cannot be reduced to a negligible level if service is provided. Similarly, a loading facility may be necessary to transport coal efficiently between points A and B, but its cost may be unchanged if the amount of coal transported is doubled or halved. Common costs are often fixed. For example, the basic portion of the outlays on track and way and structures between A and B may be both fixed and common costs.

Fixed and common costs are quite different from variable costs. Economists employ two fundamental cost concepts in defining variable costs—marginal cost and incremental cost. The marginal cost of a service is the additional cost that would be incurred to supply an additional unit, or the saving in total cost by supplying one less

unit. As such, the marginal cost of a rail service is the per-unit opportunity cost to the rail carrier of the level of a service's volume. The term "opportunity cost" refers to the value a resource could contribute if it were used in some alternative occupation rather than the one to which it is currently assigned by the railroad. Thus, marginal cost is similar in meaning to unit incremental cost and to the true economic variable cost. However, its definition makes clear that marginal cost should include the traffic-sensitive costs of capital facilities that are fungible and economically attributable to the service, as well as the more obvious cost components such as fuel, labor, and traffic-sensitive maintenance and replacement costs.

For example, locomotives and other rolling stock used to provide a particular rail service have a significant opportunity cost for a rail carrier. If not utilized to supply the service, they could instead be gainfully utilized elsewhere in the rail network, by either the rail carrier at issue or another carrier. Assuming that at least some carriers do not have excess supplies of the equipment in question, or their functional equivalents, a decrease in the quantity supplied of the service would release equipment that could decrease or delay the need to lease or purchase stock for replacement or expansion. Consequently, it follows that the opportunity cost of the rolling stock is its replacement cost, at the current cost of capital. Thus, the marginal cost of a given service includes the costs of fungible capital goods that are utilized, such as locomotives and other rolling stock, at the current cost of capital for the period of time during which they were so employed.

Of course, the marginal cost of a service also includes the wear and tear on capital assets and the required maintenance expenses that the supply of the service causes. (For example, it is clear that the passage of rail traffic causes wear and tear on track, ties, and ballast, which in turn shortens the lives of the assets. Consequently, one element of the marginal cost of rail traffic arises from the hastening of the time in which the assets it utilizes must be replaced—the present discounted value of the capital cost of the assets' value—over the time period that their needed replacements are advanced.) However, the costs of facilities that are fixed or common are not included in the service's marginal costs.

The incremental cost of a service is the cost per unit of service necessary to provide the entire service, or the cost avoided by not providing the service, given all the other services supplied. The term "avoidable cost" is also used to describe the cost per unit of service that could be avoided by not providing a particular service.

The important conceptual point here is that a railroad's total costs are composed of some costs that vary with the amount of a particular service provided and of others that do not. This is obvious enough, but considerable confusion is often engendered when the additional point is made that in the long run virtually all fixed and common costs can be varied. The reason is simply that in the long run virtually all assets must be renewed or replaced. At the date when the decision regarding renewal or replacement of the fixed factors of production required to supply a service or group of services is under consideration, the costs involved are incremental to that service or group of services. If the decision were made to discontinue those services, the costs would not be incurred.

This obviously does not mean that there is no economic distinction between variable costs and fixed and common costs. What it does mean is that the perspective of the decisionmaker is very important. When a railroad is making decisions regarding the incremental costs of adding a particular service (or the avoidable costs of eliminating a service) given existing capacity, the short-run variable costs of service will include only the additional costs of production imposed by that service. Rarely will the short-run variable costs of service include the full measure of long-run fixed costs. In contrast, when a railroad is making the long-term decision whether it is economic to replace a portion of its rail network (or to make an entirely new addition to its network), the long-run variable costs of the service or services the railroad plans to offer will include all the fixed costs that will become sunk (that is, irreversible for a significant period of time) once they are incurred.

#### *Sunk costs*

Long-run fixed costs are those costs that are not reduced, even in the long run, by decreases in output (see Baumol and Willig 1981; and Baumol, Panzar, and Willig 1988). But such costs can be eliminated in the long run by total cessation of production. Sunk costs, in contrast, are costs that (in some short or intermediate run) cannot be eliminated, even by total cessation of production. As such, once committed, sunk costs are no longer a portion of the opportunity cost of production.

Sunk costs need not be fixed, and even more important, fixed costs need not be sunk. To operate with current production techniques, a railroad requires at least a locomotive and one car, the costs of which must be included among its fixed costs. Yet because they constitute capital on wheels, most of their cost can easily and quickly be recovered by rolling them to another market,

should the railroad's management decide (and be permitted) to close down the line in question. Thus, little or none of this portion of fixed cost is sunk, in contrast to the roadbed cost, which typically is sunk. While bridges, ballast, rails, and ties can also be moved from one route to another, they can be moved only at considerable expense.

The distinction between sunk and fixed cost is not a mere technological quibble. It makes a substantial difference for the design of appropriate public policy if the costs of the firms in an industry include the one rather than the other. Sunk costs contribute to entry barriers which, as is well known, can give rise to monopoly profits, resource misallocation, and inefficiencies. By contrast, fixed costs neither constitute barriers to entry nor entail the misallocation problems to which entry barriers lead. Fixed costs are not, and do not raise, entry barriers unless they also happen to be sunk.

#### *Economies of scale and scope*

The issue of whether a firm's total costs will be recovered from prices that are equal to the firm's marginal costs of supply is logically equivalent to the question of whether the firm's operations are characterized by economies of scale, or, in alternative terminology, increasing returns to scale.

For multiproduct railroad firms, economies of production could exist due to either the level of supply of all the firm's outputs (economies of scale) or the breadth of the set of services supplied (economies of scope). Economies of scale are exhibited where equiproportionate changes in the levels of all services provided would require a less-than-proportionate change in the level of efficient costs. In addition to economies deriving from the size or scale of a firm's operations, cost savings may also result from simultaneous production of several different outputs in a single enterprise, as contrasted with their production in isolation, each by its own specialized firm. That is, there may exist economies resulting from the scope of the firm's operations.

Substantial economies of scale in the provision of some rail services, whether focused on particular routes or types of freight, result from the heavy fixed costs associated with rail operations. To transport even small amounts of freight, a railroad must generally incur the costs of track, right-of-way, locomotive power, crew, and certain facilities. These costs do not rise proportionately with traffic volume. As more traffic uses a section of a roadway, very few additional fixed costs are incurred, and the extant costs are spread over more traffic. A single track line can

handle large amounts of traffic before a full second track must be added or advanced signaling systems installed. Scheduled trains can be made longer to accommodate more cars on the same origin-to-destination route without proportional increases in the costs of locomotive power and crew. The more freight that is scheduled to traverse the same route, the larger can be the preblocked movements, with correspondingly less reclassification yard activities and time needed, and with more opportunities to run efficient through-train service. In short, additions to the levels of rail services supplied do not entail proportionate additions to the levels of expenditure required for fixed plant, for equipment investment, and for operating expenses. This is precisely the hallmark of economies of scale. Fixed costs, of both the sunk and fungible varieties, per ton of freight fall as traffic volume increases. Cost efficiencies therefore may be associated with provision of rail services by a single carrier.

Another advantage of firm integration in the rail industry arises from potential economies of length of haul. With fixed terminal expenses, longer hauls normally imply lower costs per mile. In the presence of such economies, a railroad with an integrated nationwide system will sometimes have a cost advantage over competitors that make and accept interline shipments to and from other railroads.

Increased firm size may convey cost advantages because of specialization and massed-reserves economies. A large firm may employ a more richly specialized array of accounting, finance, marketing, engineering, research, and legal talent than a smaller competitor. This specialized talent may be reflected in lower administrative costs, higher productivity, or both. The large firm can amass its cash balance reserves and spread production, market, and financial risks over a larger volume of activity. The diversification of the portfolio of transportation services offered by a large railroad creates an overall system risk factor that could be substantially less than the risk associated with investment in just one of those services.

A large railroad firm with an integrated network may also realize economies in equipment investment. In general, railroads attempt to minimize the need for new equipment purchases by using equipment interchangeably throughout the system. When cars and locomotives are needed at some shipping point, the railroad can immediately send them out of the most convenient distribution center. Operations with assigned equipment require more switching than those that draw their equipment from common pools. In addition, the ability to use locomotives interchangeably reduces the number of reserve locomotives needed to protect against equipment

failures, repairs, and inspection. A larger railroad firm can, therefore, obtain the same degree of protection at lower cost relative to total capacity carrying costs.

Another pertinent feature of the railroad industry is that substantial economies of scope result from the common costs of rail operations. Outlays on rails, ties, rights-of-way, yard facilities, locomotion, and train crews are among the many common costs of rail operations incurred in carrying a variety of types of freight between a variety of origins and destinations. These shared costs confer economies of scope on carriers offering a multiplicity of transportation services: A carrier that provides an array of services can do so at a lower total cost than a set of carriers producing each service separately.

*Economies of size versus economies of scale and density*

The overall size of a railroad is likely to be quite independent of the amount of traffic that travels on any of its routes. That is, a large firm may have short or long hauls and high or low traffic densities between different points. There has been serious confusion about economies of size and economies of scale and density, and a concomitant failure to specify clearly which is being measured (see Harris 1977). Economies of scale are carefully defined to refer to a long-run average cost curve that declines as the quantity of the firm's output of a given collection of services increases.

Comparing the average costs of railroads that have different sizes of route networks, as many have done, does not provide information directly relevant to economies of scale, because such railroads do not supply different amounts of a given collection of services. Instead, they likely offer quite different collections of services as a result of their different route mileage and architecture. The correct and relevant measure in railroading is the extent of scale economies that relates to traffic volume on each route, rather than to traffic volume over an entire and possibly growing system. To emphasize this point, these economies have come to be termed economies of density. Thus, the critical determinant in pricing and (dis)investment policies is whether there are economies of density. It is therefore important to assess the degree to which unit costs decline as output increases while holding the route system, or miles of rail line, constant. A small firm with high traffic density could potentially have lower average costs than a large firm with low density.

Economies of density are normally attributed to declining average capital costs. However, the provision of rail service entails more than simply installed capacity; it includes minimal (and often indivisible) amounts of crew,

engines, maintenance, and other variables. Indeed, recent empirical studies indicate that the maintenance of way and structure and transportation expense (mainly fuel and crew wages) account for a significant portion of the estimated economies of density. Approximately two-thirds of these economies are due to variations in unit operating costs per route-mile.

Under significant economies of density, the cost-minimizing market structure for a given route might call for a single firm—that is, the route would be a natural monopoly. In the absence of any other scale economies, the national railway system could be made up of a large number of small firms, each with a local monopoly. Alternatively, if there were substantial economies of firm size without economies of traffic density, it would be economic to have a number of integrated nationwide railroads that competed on all their routes. However, with economies of density, and with economies of scope, and with some economies of end-to-end long hauls, the cost-effective structure of the rail industry is likely to be characterized by very few firms.

*Empirical evidence on scale economies*

There are at least two approaches to measuring cost-scale relationships in the rail industry. The first way is to use the expertise of those with intimate knowledge of railroad operations in ascertaining whether the costly inputs required to supply rail services must be expanded proportionately to accommodate increases in the quantities of services provided. This is known as the engineering approach. The second approach, statistical cost analysis, is to estimate econometrically the relationship between railroad costs and the levels of rail services provided. There is no conflict between the conclusions reached by using these two different approaches in the railroad industry. Both indicate quite clearly that railroad operations are characterized by increasing returns to scale and that the recovery of railroad costs consequently requires that prices exceed marginal costs.

The first approach has been followed by a long succession of industry observers, who have provided a knowledgeable overview of how economies of scale arise in rail operations. First, economies are created for the system as a whole by operations that are directly common to all traffic, such as network planning and management. If network management and control (for example, billing, payroll, systemwide insurance, and other housekeeping functions) involves a fixed cost regardless of network size (above a certain threshold), these costs will be spread over a larger user base in a larger integrated rail system.

Similar integration economies arise in communications and dispatching activities and from increases in work-force specialization within the repair facilities of larger systems. Finally, large railroads benefit from capital-raising and other pecuniary economies (for example, price concessions from suppliers). Indeed, this appears to be one of the most persistent advantages of firm size, with small incremental capital cost savings enjoyed out to very large scales. However, the capital-raising economies of scale are also associated with real resource savings. Negotiating a loan or a new stock issue or obtaining necessary regulatory clearances entails transaction costs, some portions of which are nearly fixed. Clearly, the larger the issue, the lower those costs per unit of capital raised.

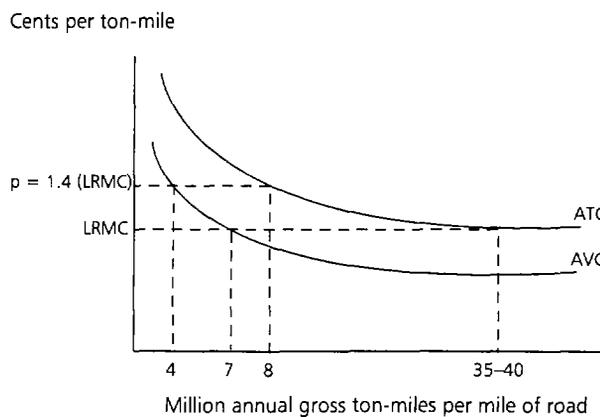
Second, the integration of the railroad system permits economies that directly benefit some traffic and indirectly benefit other system activities. Most ancillary plant (for example, storage and marshaling yards, sidings, switches, and fueling and repair stations) can be utilized by more and more shippers without causing a corresponding increase in the amount of investment required. A coal shipper might need a storage and marshaling yard to hold its cars until a trainload volume is accumulated. If a mine produces only 20 carloads a day and holds them until 100 cars are available, a yard that could store and switch 100 cars would be required. However, on an independent operation basis, only 20 percent of the yard would be utilized in the first day, 40 percent in the second, 60 percent on the third, and so on. Yet a railroad that connected with more mines might receive 20 cars a day from each of five mines and send a trainload every day. The railroad would still need only a 100-car yard, but it would have five times as many cars to share in the coverage of the investment and operating costs of the yard.

Similarly, a full siding is necessary if each day only one train will meet one other train coming in the opposite direction. The same size siding would be necessary if four trains were meeting four other trains at the same place. Crossing protection must be built and maintained in a densely populated area whether the railroad sends one train a day or three trains a day over the track at the crossing. The same is true for switches, fueling stations, and all other fixed plant investment. Once the plant is installed, a railroad can utilize it far more heavily with very little additional fixed investment cost. Also, a train of 40 cars needs a crew of the same size as a train of 60 cars. The ability to marshal cars of different shippers into a larger train also cuts other operating expenses. The engine power necessary for a longer, heavier train is not commensurate with the additional cars that have been added.

The statistical or econometric approach to analyzing railroad economies of scale also has a long history. This history is rife with academic controversy and with steadily improving research methods. For example, some econometric studies, because they were founded on arbitrary allocations of costs between freight and passenger services, found no evidence of rail economies of scale. Other studies failed to distinguish economies of scale from possible economies stemming from the geographic extent of a railroad's operations. Such studies, finding that railroads covering more territory do not necessarily enjoy lower costs per ton-mile of freight, incorrectly concluded that increasing returns to scale are absent.

Recent econometric studies conducted in the United States have avoided these pitfalls, and their important conclusions warrant discussion here. First, most of the rail system is subject to increasing returns to scale and has elements of natural monopoly, whether considered in a single-product or a multiproduct setting. Second, as figure 13.A1 indicates, although unit costs decline sharply with density, at some point between 35 million and 40 million annual gross ton-miles per route-mile, depending on the commodity mix, the cost curve flattens out and a large part of the traffic in the system flows over this range of flat (constant) costs. This range (flat part) represents the level of minimum efficient density, which one can think of as the capacity of a single track between two points, the fundamental indivisibility in the rail cost structure. Higher traffic density can be served at approximately constant cost by adding segments of parallel track and signaling devices. Third, for very short-haul, terminal-oriented railroads, the long-run cost curve

**Figure 13.A1 Unit total costs, operating costs, and traffic density**



LRMC: Long-run marginal cost  
 ATC: Average total cost  
 AVC: Average variable cost  
 Source: Levin 1981a.

seems to flatten out much sooner (at less than 2 million net ton-miles per route-mile). Fourth, there are considerable economies for longer hauls.

Overall, these studies establish the presence of substantial economies of scale in the freight operations of railroads. They indicate that pricing at short- and long-run marginal costs would recover less than 80 percent of total long-run costs. Also, high-density traffic seems to exhaust the economies of scale experienced at lower densities, but significant diseconomies of scale do not occur as densities grow larger. Consequently, since all railroads have relatively low-density traffic on many segments, and since most traffic flows on low-density track while it is gathered and distributed, rail services exhibit substantial economies of scale overall. As a result, prices set at marginal costs would leave uncovered a substantial portion of total efficient railroad costs.

## Appendix B

### Rail costs, profitability, and structural changes

Most of the statistical and econometric studies estimating rail costs and production functions suffer from two fundamental weaknesses. First, they generally fail to differentiate between way-and-structures capital, which is a measure of the quantity and quality of the capital utilized in the roadbed, and track, which in addition to being a proxy for the roadbed capital is also a measure of common carrier obligations to haul commodities. Second, they generally fail to take into account the effect on costs of the route network and to differentiate between high-density, fully utilized track and light-density, underutilized track (see Friedlaender and Spady 1980).

Way-and-structures capital is a measure of the capital utilized in the roadbed and as such should be treated as a conventional factor of production. An increase in the fixed factor, way-and-structures capital, should lead to a reduction in other factors and hence a reduction in variable costs. In contrast, general track and low-density track should be treated as technological variables that affect the costs of the railroad firm in a way that is not necessarily associated with conventional production theory. An increase in low-density route-miles or total track represents an increase in common carrier obligations and should therefore be associated with increases in expenditures on other factors of production.

A *ceteris paribus* reduction in way-and-structures capital will reduce the quality of the existing track and hence lead to cost increases by requiring increased amounts of variable factors. This is to say that more money must be

spent on equipment maintenance and train crews as the quality of the roadbed deteriorates and speeds are reduced. Similarly, a *ceteris paribus* reduction in track will be correlated not only with a reduction in common carrier obligations and improvements in the quality of the existing track, but also with increases in its utilization. The first two considerations will tend to reduce costs whereas the latter will tend to increase them, making the impact of reduced track somewhat ambiguous. Reduction in low-density track, in contrast, will reduce common carrier obligations and their associated costs and will therefore tend to generate cost savings.

### *Railroad costs and infrastructure variables*

To assess the possible savings that would accrue from policies aimed at changing the railroad infrastructure, it is important to quantify the impact on rail costs of changing the three main infrastructure variables—the amount of way-and-structures capital, general track, and low-density track.

*Ceteris paribus* increases in way-and-structures capital will raise the amount of capital embodied in each mile of track and thus lead to reductions in variable costs. Indeed, econometric estimates by Friedlaender and Spady (1980) reveal that a 10 percent increase in way-and-structures capital leads to more than a 4 percent decrease in variable costs, consisting of decreases of 11 percent in equipment usage, 3 percent in general labor, 3 percent in yard and switching labor, 2 percent in on-train labor, and 0.6 percent in fuel and materials. These estimates seem to indicate that the main effect of an increase in way-and-structures capital is to decrease equipment requirements, with somewhat lesser savings in the labor categories. This confirms the intuition that the source of the savings in variable costs that result from an increase in way-and-structures capital is train speeds.

*Ceteris paribus* reductions in light-density track are correlated with increases in the amount of capital embodied per mile of track and reductions in the proportion of low-density mileage; both of these factors should be associated with cost reductions. Econometric estimates (Friedlaender and Spady 1980) indicate that a 10 percent reduction of low-density route-mileage would reduce total variable costs by approximately 3 percent. This reduction comes about through reductions in yard and switching labor costs of somewhat more than 4 percent, in general labor and equipment expenditures of somewhat more than 3 percent, and in fuel and materials expenditures of less than 1 percent. Thus, the primary savings arising from the abandonment of low-density line

are concentrated in transportation and switching categories associated with moving trains over lightly utilized track.

Finally, *ceteris paribus* reductions in general track are correlated not only with increases in capital embodied per mile of track, but also with increases in the proportion of low-density track. While the first factor should tend to reduce costs, the second should increase them. Econometric estimates (Friedlaender and Spady 1980) reveal that a 10 percent reduction in general track or route-miles leads only to a reduction of total costs of less than 1 percent. In terms of factor utilization, reductions in general route-miles lead to sizable reductions in equipment and materials expenditures but increases in labor expenditures. Thus, as the same volume of traffic is moved over a smaller network, increased expenditures on labor and switching are required, whereas savings on fuel and equipment are achieved.

#### *Low-density lines and profitability*

Rail costs are quite sensitive to changes in way-and-structures capital and in light-density route-miles but not to changes in general route-miles. A change in general track or route-miles without a concomitant change in low-density route-miles has a small impact on variable costs but a significant effect on factor intensities. What distinguishes the provision of low-density service from that of general network expansion is the greater labor intensity of the former. Thus, efforts to adjust amounts of way-and-structures capital through roadbed maintenance or to abandon light-density lines are likely to have a rather large impact on costs, whereas the abandonment of general track per se will lead to relatively few economies.

Econometric estimates (Friedlaender and Spady 1980) reveal quite clearly that low-density lines are a significant drain on railroad profitability and seriously impede the attainment of static and dynamic efficiency in the industry. The avoidable losses recoverable by abandonment appear to be quite significant. In addition, the burden of excess capacity seems to have a dynamic impact on efficiency. The abandonment of low-density lines stimulates the formation of new capital on the high-density portions of the rail network. First, since abandonment reduces the need for cross-subsidization, rates on the high-density lines are permitted to fall toward marginal cost. The lower rates attract additional traffic and thus raise the level of desired capital. Second, the abandonment of low-density lines lowers the cost of capital to rail firms by improving their long-run profitability and reducing the risk of bankruptcy.

#### Notes

1. The authors acknowledge their debt to the thinking and writings of William J. Baumol on many of the subjects covered in this section. A summary of some of this material can be found in Baumol and Willig 1987.
2. It should be noted that in many instances the relevant competition is not just on the route invoked in the rail movement but also on alternative routes that offer economic substitute services for the shipper. For example, a manufacturer may find it equally desirable to ship output to two very different places for the purposes of sale, and will choose the option with the least expensive transportation.
3. This important property of stand-alone cost is not significantly undermined by the practice of determining stand-alone cost in a fashion that may provide guidance or even a model of the actual railroad. While these operations may provide guidance or even a model for the operations of the stand-alone railroad, the stand-alone cost need not reflect the same decisions as those made by the incumbent, especially if they lead to unnecessarily high costs.
4. One example of this effect arises under rate-base rate of return regulation, as was understood by Averch and Johnson (1962) in their seminal paper.
5. For a more complete discussion of these cases, see Ordovery, Sykes, and Willig 1985.
6. This standard was first developed in Ordovery and Willig 1981.

#### References

- Averch, Harvey, and Leland L. Johnson. 1962. "Behavior of the Firm under Regulatory Constraint." *American Economic Review* 52: 1058-59.
- Baumol, William J., and Robert D. Willig. 1981. "Fixed Costs, Sunk Costs, Entry Barriers, and Sustainability of Monopoly." *Quarterly Journal of Economics* 95: 405-31.
- . 1987. "Railroad Deregulation: Using Competition as a Guide." *Regulation* 2 (1): 28-36.
- Baumol, William J., John C. Panzar, and Robert D. Willig. 1988. *Contestable Markets and the Theory of Industry Structure*. San Diego: Harcourt Brace Jovanovich.
- Boyer, Kenneth D. 1987. "The Costs of Price Regulation: Lessons from Rail Regulation." *Rand Journal of Economics* 18: 408-16.
- Braeutigan, Ronald R. 1979. "Optimal Pricing with Intermodal Competition." *American Economic Review* 69: 38-49.
- . 1984. "Socially Optimal Pricing with Rivalry and Economies of Scale." *Rand Journal of Economics* 15: 127-35.
- . 1987. "An Analysis of Fully Distributed Cost Pricing in Regulated Industries." *Bell Journal of Economics* 1: 182-95.
- Friedlaender, Ann F. 1969. *The Dilemma of Freight Transport Regulation*. Washington, D.C.: The Brookings Institution.
- . 1971. "The Social Costs of Regulating the Railroads." *American Economic Review* 61: 226-34.

## REGULATORY POLICIES AND REFORM: A COMPARATIVE PERSPECTIVE

- Friedlaender, Ann F., and Richard H. Spady. 1980. *Freight Transport Regulation*. Cambridge, Mass.: The MIT Press.
- Harris, Robert G. 1977. "Economies of Traffic Density in the Rail Freight Industry." *Bell Journal of Economics* 8: 556-64.
- Kahn, Alfred E. 1988. *The Economics of Regulation*. Cambridge, Mass.: The MIT Press.
- Keeler, Theodore E. 1983. *Railroad, Freight and Public Policy*. Washington, D.C.: The Brookings Institution.
- Levin, Richard C. 1978. "Allocation in Surface Freight Transportation: Does Rate Regulation Matter?" *Bell Journal of Economics* 9: 18-45.
- . 1981a. "Exit Barriers and Railroad Investment." In Gary Fromm, ed., *Studies in Public Regulation*. Cambridge, Mass.: The MIT Press.
- . 1981b. "Railroad Rates, Profitability, and Welfare Under Deregulation." *Bell Journal of Economics* 12: 1-26.
- Ordovery, Janusz A., and Robert D. Willig. 1981. "An Economic Definition of Predation: Pricing and Product Innovation." *Yale Law Journal* 90: 1-44.
- Ordovery, Janusz A., Alan O. Sykes, and Robert D. Willig. 1985. "Non-Price Anticompetitive Behavior by Dominant Firms Toward the Producers of Complementary Products." In F. M. Fisher, ed., *Antitrust and Regulation*. Cambridge, Mass.: The MIT Press.
- Tye, William B. 1987. "The Voluntary Negotiations Approach to Rail Competitive Access in the Transition to Deregulation." *The Antitrust Bulletin* 32: 415-50.
- Willig, Robert D. 1994. "Public versus Regulated Private Enterprise." In World Bank, *Proceedings of the World Bank Annual Conference on Development Economics 1993*. Washington, D.C.
- Willig, Robert D., and William J. Baumol. 1987. "Using Competition as a Guide." *Regulation* 1: 28-35.

# Regulatory policies and reform in the electricity supply industry

David M. Newbery

In the electricity supply industry, high-tension transmission and low-tension distribution systems are natural monopolies, although generation and supply (that is, contracting with and billing customers) are potentially competitive. In most countries the industry is vertically integrated at the regional or national level; in many of these countries the industry is under public ownership. Some countries have investor-owned utilities, and many have mixed systems of public, municipal, and state ownership (particularly of the high-tension grid). Where the natural monopoly elements are privately owned, they are invariably regulated. The forms of regulation range widely, and different countries exhibit apparently very different systems.

Until the 1980s the system of ownership was normally taken as historically given and not questioned. Most industrial countries experienced rapid growth in electricity demand after the Second World War yet were able to finance the huge investment program needed to meet demand and modernize generation and transmission, with gradual improvements in reliability and typically falling real prices for electricity. Such success muted public criticism, at least until the oil shocks of the 1970s and fears about nuclear power shortly after.

In developing countries the electricity supply industry has been almost invariably under state control, and international organizations, including the World Bank, have provided substantial absolute amounts for (although rather modest shares of) the investment requirements. Ten percent of official development finance went into the power sector in 1984–91 (World Bank 1994, table A.4). Whereas nearly a quarter of total public investment in a sample of middle-income countries in the 1980s was for power (World Bank 1994, figures 1.1, 1.2), power accounted for about 15 percent of total World Bank lending up to 1991 (World Bank 1993, p. 12). Of the average annual power sector investment of about \$80 billion in developing countries during 1984–89, official aid

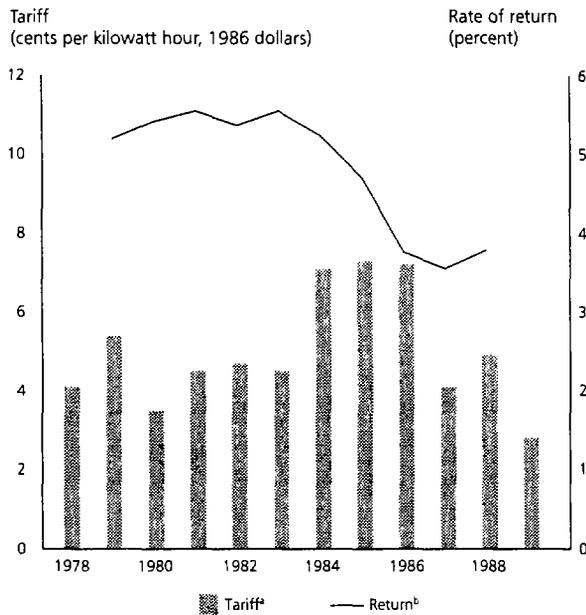
financed about 10 percent and the World Bank itself about 3 percent (World Bank 1994, tables A.3, A.4).

Performance was frequently unimpressive, particularly in the high-inflation period after the oil shocks of the 1970s. Because prices were normally below long-run marginal cost, often despite excess demand, investment could not be adequately financed out of profits, as in many industrial countries. As shown in figure 14.1, average real power tariffs declined to below \$0.04 per kilowatt hour (1986 constant U.S. dollars) for a sample of 60 World Bank member countries in 1989. The rate of return on revalued net fixed assets also declined to below 4 percent for a sample of 360 actual financial rates of returns recorded for 57 World Bank member countries (World Bank 1993), well below the 10 percent rate of return normally used as the test discount rate by international agencies. Only 60 percent of power sector costs were covered by revenue (Besant-Jones 1993), and self-financing ratios fell to only 12 percent of investment requirements in 1991 (World Bank 1993, 12).<sup>1</sup> Newbery (1992) noted similar problems for Asian countries.

Figure 14.2 shows the magnitude of the problem. Underpricing electricity resulted in a heavy fiscal burden estimated at \$90 billion annually (World Bank 1994, table 6.7), or about 7 percent of total government revenues in developing countries—larger than annual power investment requirements of about \$80 billion—while technical inefficiencies caused true economic losses of nearly \$30 billion annually.

By the late 1980s it was becoming clear that this situation was financially infeasible, not just for the utilities but also for governments, particularly in Latin America. There were calls for more fundamental reform, often associated with arguments for privatization, as a potentially dramatic way of solving the problem of poor financial and economic performance. The next section considers the pressures for reform in industrial and developing countries.

**Figure 14.1 Tariffs and returns of electricity supply industries in World Bank countries**

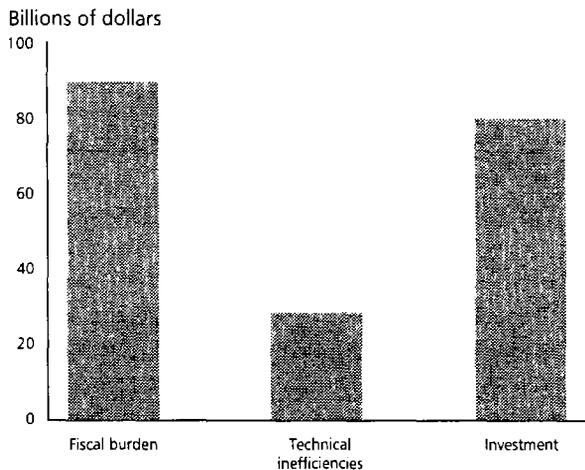


a. Data are for 60 countries.  
 b. Data are for 57 countries.  
 Source: World Bank 1993.

**Pressures for reform**

Why did pressures for reforming the electricity supply industry emerge in industrial countries? Briefly, the old systems of financing large, capital-intensive generation plants came under strain for a variety of reasons. Oil price shocks caused an initial rethinking in the 1970s, making nuclear power potentially attractive. But the fall in demand growth rates reduced the demand for new plants, and macroeconomic shocks and inflation in the late 1970s

**Figure 14.2 Inefficiencies in electricity supply in developing countries**



Source: World Bank 1994.

escalated the real cost of large investment projects, certainly in the United States and the United Kingdom, and probably elsewhere as well. These problems, combined with growing public unease, tightening safety standards, and prudential reviews in the United States made further investment in nuclear power unattractive in most countries. Excess capacity in generation introduced a wedge between the efficient price of electricity (short-run marginal cost) and the long-run marginal cost, on which pricing and investment decisions were often based. Ideas of deregulation, third-party access to the transmission system, and the development of low-cost, high-efficiency combined-cycle gas turbines and combined heat and power units opened possibilities of bypass of utilities by generators in the United States, and these reforms and innovations highlighted the tension between efficient and cost-recovering tariffs. Excess capacity also gave governments or utility regulators a breathing space in which to contemplate regulatory reforms without prejudicing investment and expansion, neither of which were needed.

These strains were visible in several industrial countries, but the main impetus for radical reform came from the United Kingdom, which had embarked on a program of privatizing nationalized industries and utilities, starting with British Telecom in 1984, followed by British Gas as well as many less contentious examples where no natural monopoly was at stake.<sup>2</sup> Both British Telecom and British Gas were sold as regulated monopolies, and there was widespread criticism that privatization merely transferred monopolies from the public to the private sector, with little obvious benefit and considerable risk to consumers and voters.

The British government had decided to privatize the electricity supply industry as well, and considered it a political imperative that the industry not be sold as a monopoly. The industry was radically restructured, a task of monumental complexity given the tight schedule and the lack of models to follow. Had the government realized the enormity of the task, it might well have balked, but masterful delegation combined with an efficient and dedicated civil service (coupled with the handsomely paid expertise of merchant banks and consultants) delivered the desired reforms. The most dramatic element was the deintegration of the industry—separating the natural monopolies from the potentially competitive parts—and the creation of a spot market for bulk power. That privatization worked, and delivered considerable efficiency gains in generation within a short time, undoubtedly convinced many observers in other countries that reform of the electricity supply industry was not only feasible but potentially very attractive.

The main question facing each country was whether reform was needed and, if so, what form it should take. Other pressures reinforced this questioning. The European Commission, as part of its drive to create the single market of the European Union, was attracted to third-party access to network industries, for which the English model of a bulk power market served as an obvious paradigm. This arrangement could no longer be dismissed as unworkable and now had to be argued against by reluctant continental electricity supply industries and their governments. The potential gains from international trade in Nordic power created similar pressures for market solutions, while the systemic reform in Central and Eastern Europe required utility regulation to be created to replace state tutelage and as a possible precursor for privatization. Given the apparently large book value these industries represented for increasingly fiscally pressed states, privatization was more and more attractive.

Developing countries were facing the rather different problem of underpricing, fiscal strain, and the resulting inability to self-finance noted above. Not only did this situation have adverse impacts on the electricity supply industry itself, but the fiscal drain, at 7 percent of gross domestic product (GDP), exceeded the general budget deficit in most countries. A study by the Asian Development Bank (1988), summarized by Kohli (1987), argued that increases in public indebtedness since 1975 derived in large part from the failure of public enterprises, of which the electricity supply industry was a large part, to generate adequate profits to cover their investment demands.<sup>3</sup> The diagnosis of the inefficiencies of public enterprise suggested two types of remedy: privatization or specified improvements in the performance of enterprises remaining in the public sector.

Even in countries where privatization of the whole electricity supply industry is not an issue, the need for efficient investment has led governments to consider ways to involve the private sector, with the possible additional advantage of attracting direct foreign investment without increasing government-guaranteed debt. The extent to which a utility can finance investment out of retained funds depends on the real rate of return earned on its capital, its gearing, and the cost of financing its existing debt. If the real rate of return on capital (at written-down current replacement cost) is  $r$ , the real cost of borrowing is  $i$ , the debt-capital ratio  $\gamma$ , and the rate of growth of required capacity is  $g$ , the fraction of new investment that can be financed can be no higher than  $(r - i\gamma)/g$  (assuming that the existing debt can be maintained constant in real terms).<sup>4</sup> In a fast-growing economy with high gearing, this

fraction might be less than unity. For example, if  $r = 10$  percent,  $i = 5$  percent,  $\gamma = 80$  percent, and  $g = 12$  percent, only half of new investment can be financed out of retained profit (although in this case the debt-equity ratio will steadily fall). If gearing remains constant, the fraction of profits required for investment falls to  $(1-\gamma)g/(r - i\gamma)$ , or using the previous numbers, only 40 percent (but in this case 80 percent of new investment would be financed by borrowing).

The magnitude of this investment demand could be substantial: China alone expects to spend \$120 billion in 1994–2004 to meet its forecasted power needs. Much of this investment will be foreign financed, and although the Chinese government clearly hopes that investors will take much of the construction and performance risk, investors view the proposed form of price regulation as unattractive, especially given the difficulties of hedging against exchange rate risk.<sup>5</sup> If, as many countries would prefer, this investment is financed by private capital without government guarantees (and thus not counted as official debt), robust systems of regulation will be required where current regulation provides inadequate safeguards to private investors. In most cases such systems will have to be either created (in the case of entirely state-owned industries) or reformed.

#### *The choice of regulatory system*

Regulation will be required whenever private equity investment is involved. Even if the entire industry remains publicly owned, it may beneficially be corporatized and subject to explicit regulation, rather than remain subject to implicit regulation.

Several questions must be answered to design a good system of regulation: How will tariffs be set and investment financed? Should the industry be maintained as a vertically integrated whole, or should it be deintegrated into competitive and natural monopoly parts? If it is vertically integrated, should third-party access be allowed, and if so, how should access charges be set and regulated? Should there be a bulk electricity market, and if so, for whom and on what terms? If the industry is deintegrated, should regulation be confined to the natural monopoly parts of transmission and distribution, with no regulation of the competitive parts? Should transmission be combined with distribution, with generation, or kept separate? Should the industry be vertically integrated on regional lines, or would a single transmission grid for the whole country be preferable? Is it preferable for all, a part, or none of the industry to be in public ownership? Given answers to these structural questions, what form should

the regulatory institutions take, with what powers and systems of modification and appeal?

The fact that the electricity supply industry in different countries exhibits a wide range of answers to these questions prompts another question: To what extent is the choice of regulatory systems constrained by physical endowments, politics, the constitution, and the legal system? If, as will be argued, the regulatory system in any country has evolved a structure of institutional supports to ensure its stability and predictability, what has changed to allow the system to be substantially altered?

The transition to a market economy in Central and Eastern Europe creates opportunities for a radical reconsideration of the whole system of economic regulation, particularly because many countries in this region have placed heavy emphasis on privatization and foreign investment. For such countries the lessons of experience elsewhere may be especially helpful. Privatization in industrializing countries raises similar questions. Options for system redesign occur rarely, and the choices made at such a critical juncture will cast long shadows. Subsequent changes may be far more difficult.

This chapter attempts to answer the questions posed above and to draw lessons from the experiences of the English electricity supply industry,<sup>6</sup> which provides a good example of restructuring and privatization, from other industrial countries in Europe and North America, and from Latin America, which parallel in many ways the English experience.

#### *Natural monopoly and competitive elements*

A natural monopoly arises when a single firm can provide the range of goods or services at lower total cost than can a set of firms. This cost condition is not itself sufficient to justify preserving the industry as a regulated monopoly; the cost advantage would have to be sufficient to justify the additional costs of regulating the resulting monopoly. More competitive structures may raise production costs but reduce regulatory costs or allocative inefficiencies sufficiently to provide the service at lower total social cost. Several countries are actively encouraging duplication of telephone lines to customers, for example, in the hope that the resulting competition will reduce the overall cost of service.

Berg and Tschirhart (1988) cite Farrer's (1902) catalog of typical characteristics of natural monopolies. A natural monopoly (a) is capital intensive and of minimum economic scale; (b) produces nonstorable output with fluctuating demand; (c) has locational specificity, which creates location rents; (d) supplies necessities or is essen-

tial for the community; and (e) has direct connections to customers. Some of these attributes contribute directly to the likelihood that a single firm will have lower supply costs within a well-defined area. If the output can be stored, or readily shipped to dispersed customers, then the market increases in size and may be able to sustain more than one firm at minimum economic scale. The combination of necessity and direct connection implies large potential exploitative power by the producer, ensuring that regulation or public ownership will be politically inevitable.

It is clear that electricity closely fulfills these conditions. Because electricity cannot be readily stored, supply must be continuously adjusted to varying demand. At one extreme each customer could have his own generator, but the spare capacity required to meet peak demands would be excessively costly, as would having an inefficiently small plant. As a result numerous consumers are supplied by each utility through a distribution system. Up to some limit the larger the number of consumers served, the lower the average operating and capital costs, because smaller proportionate reserve margins are required, larger generating stations with lower running costs can be built, and the benefits of scheduling stations of differing variable costs in a merit order can be realized. These scheduling advantages require central dispatch, while the full advantages of an integrated system require coordination between investment in generation and transmission.

#### *The demand for regulation*

Network natural monopolies like the electricity supply industry must inevitably be subject to social control in a democracy. These social pressures also were visible in the former soviet-type economies and in many other political systems. Scale economies, particularly in distribution, provide the network owner with considerable market power, whereas the nonstorability of the supply, consumer's dependence on the supplier, and the essential nature of the service all conspire to generate large social and political demands that the supplier not abuse his market power. Local or central governments have therefore always stood ready to require suppliers to guarantee access on fair terms. The fact that suppliers need rights of way provides the leverage enabling governments to impose an obligation to supply.

There is a counterpart to this demand for regulation, because the electricity supply industry is capital intensive, and its assets are durable, long-lived, and immovable. The political demands for access and "fair" or nonexploitative prices mean that investors, after they have sunk their cap-

ital, must expect that they will be limited in the prices they can charge and subject to possibly onerous obligations to supply and to guarantee security, stability, and safety. If investors are to be induced to invest, they need reassurance that future prices will be set at a sufficiently remunerative level to justify their investment. Once the capital has been sunk, the risk is that the balance of advantage will shift toward those arguing for lower and possibly unremunerative prices. There are numerous examples of countries failing to adequately index the prices of public utilities in periods of inflation. The problem can be posed more sharply. Why should anyone sink money into an asset that cannot be moved and will not pay for itself for many years? Investors would have to be confident that they had secure title to the returns and that the returns would be sufficiently attractive.

Durable investments thus require the rule of law, specifically, the law of property, which is a public good provided by the state. If the state exists primarily to enforce the rights of property owners, then there is no problem. But by the time electricity became important, the state represented a wider range of interests and needed to balance the claims of property against those of workers, voters, and consumers. The resulting tensions weakened property rights, because the coercive power of the state could be used not only to enforce laws, but also to regulate economic activity, impose taxes, and even to expropriate property.

If the industry is to be successfully privately financed, regulation must credibly satisfy the demands of both consumers and investors. Some countries, notably Germany and the United States, have managed to solve this problem, but many have failed. If it is not possible to create an efficient and credible system of regulation, public ownership will be the only alternative. Indeed, the simplest explanation for Short's (1984) observation that most network utilities exhibiting natural monopoly are in the public sector is that it was not possible to devise a satisfactory and credible system of regulation that would both attract finance and deliver the service at lower cost in the private sector.

Each jurisdiction must therefore find a solution to the basic problems of reassuring consumers and investors (who may be the taxpayers), though not all solutions will be equally satisfactory to both groups. A good system of regulation will command the support of consumers, will provide sufficiently remunerative prices to enable investment to be financed, and will do so at low cost, which in turn means that investors have confidence that their investment will be able to cover its financial costs. In turn

investors will coordinate investment in transmission and generation to secure the least-cost expansion of the system consistent with adequate security against system failures, fuel shortages, and price shocks. Because electricity is vital to production, governments will also have to be convinced that electricity supply will be under adequate domestic control in times of international tension or conflict.

How can the regulatory system be designed to reassure private investors? The experiences documented below illustrate various solutions, which fall into two main types. One solution is to provide constitutional guarantees to a fair rate of return, as in the United States, upheld either by an independent legal system that protects property rights or by creating sufficiently independent regulatory agencies supported by appeal procedures to guard against expropriatory behavior. Designing such regulatory systems is critical to the success of attempts to privatize the electricity supply industry, a fact that is often inadequately appreciated in Central and Eastern Europe (Newbery 1994a).

The second solution is a regulatory compact in which the costs to the government of intervening to impose tighter regulation outweigh the benefits in terms of lower prices and short-run voter support. Many continental electricity supply industries have evolved systems of essentially self-regulation in which prices are kept remunerative but not exploitive and supply and quality are satisfactory, so that the government has little obvious reason to intervene. This protection against intervention may be strengthened by the division of responsibility between the various tiers of government (central and local, or state and federal), as it may also be if the government itself relies on consensus (as in a coalition) that would be disturbed by intervention. Intervention may be deterred if it is appreciated that the consequences of intervention would have high economic cost.

This last point can be illustrated by telecommunications, which many countries have found the most attractive industry to privatize early. Because telecoms have highly durable and specific investments, investors cannot recover their sunk cost and move elsewhere. The investment is very capital intensive, at about \$2,000 per new line, and the operating costs are low relative to the capital cost. The investor's worry is that once the investment has been made, the regulator or government may wish to lower prices and transfer rents to domestic subscribers. Because operating costs are so low relative to total costs, this option would seem to be politically attractive to the host country. What might deter the regulator or govern-

ment from expropriatory tariffs or overly tight price regulation?

Gilbert and Newbery (1988, 1994) have argued that the efficacy of regulatory regimes can be studied as a repeated game between the utility and the regulator, provided that the regulator operates under a stable constitutional regime. The constitution together with the laws under which the industry is to be regulated and privatized lay down the rules of the game and the expectations of the players. If the regulator deviates from these rules or expectations, the utility may retaliate, at the least by not investing further but possibly by more costly and immediate actions. The regulator will then weigh the costs of this retaliation against the benefit of lower prices and be dissuaded from deviating if the costs are too high relative to the gains.

These costs will be high if the country needs to sustain a high rate of investment in telecoms and if foreign expertise is required for further investment. If the foreign investor pulls out because of justified dissatisfaction with regulation, other companies will be reluctant to risk a similar fate. The country then will be forced to stop telecoms investment or will have to spend large sums acquiring the indigenous expertise for autarkic expansion.<sup>7</sup> The retaliation costs will be even higher if foreign expertise is needed to operate the system on a daily basis or foreign cooperation needed to interconnect with foreign networks, because in these situations the host country not only loses future expansion but risks the whole current system. Modern telecoms systems rely heavily on software programs to manage the switches and route the calls, and the threat of erasing (or not adequately maintaining) this software would be an even more costly form of retaliation. In fact the costs to the country of alienating a major foreign telecoms investor would seem to be so large relative to the benefits that all parties should be confident that it will not happen. Therefore, there ought to be little to worry about, unless regulation is politicized and political tenure unstable.<sup>8</sup>

#### *Regulatory solutions*

The history of the electricity supply industry in different countries illustrates the variety of solutions that have been found to this problem. The solutions available to any jurisdiction are constrained by politics, history, endowments, technology, and the state of the economy. The solutions fall into three main types: the industry can be entirely publicly owned, and hence directly subject to political control and access to funds; entirely private but regulated either explicitly or implicitly; or a mixed system

in which the private sector is implicitly controlled by the potential of the remaining publicly owned system to take over its function. In addition the regulatory system can be local, regional, or national.

Henney (1992) compared the way in which 11 European electricity supply industries coordinate dispatch and investment in generation and transmission. The simplest structure is a publicly owned national monopoly such as existed in Belgium, France, Italy, Portugal, and the United Kingdom before 1990. Austria, the Netherlands, and Spain have deintegrated industries to varying extents but have cooperative power pools that arrange dispatch in cost-merit order. In the Netherlands the four regionally based generation companies own the grid/dispatch company; in Spain the grid company is under public control; and in Austria the grid is owned by the national company with the ultimate obligation to supply. Coordinating investment in Austria is rather decentralized, whereas in the Netherlands the industry draws up plans subject to government approval, and in Spain the government determines the investment plan.

Germany and Switzerland have far more complex and fragmented structures, reflecting their federal structures and the lack of major restructurings associated with nationalization (many other European electricity supply industries were nationalized after the Second World War). The Scandinavian electricity supply industries are under mixed public and private ownership, are largely self-regulating, and have achieved coordination by cooperation and negotiation, reflecting the prevailing spirit of polite cooperation and competition of Nordic societies.

Historically, the electricity supply industry emerged in the last quarter of the nineteenth century, before central governments had any experience or inclination to become involved in productive activities (except armaments). The railways had already found it necessary to obtain rights of way and with them both enabling legislation and regulation (Foster 1993). Central governments therefore often defined the terms under which suppliers could hold franchises with rights to create distribution systems, though the scale of early electricity undertakings made them naturally subject to local government oversight. Not surprisingly, early undertakings were frequently either municipally owned or franchised by the local government.

This form of regulation was fine so long as generating stations were small compared to local demand. It became clear very early, however, that there were substantial economies in building larger units and serving larger market areas using higher-tension transmission systems. The main problem to solve was how to transfer responsibility

for electricity supply from the local level to an authority covering a sufficiently large number of consumers to reap these economies of scale while preserving satisfactory representation of local interests. Resolving this problem first required solving the problem of coordinating investment in generation and transmission to secure least-cost delivery of electricity. The key was the creation of an integrated transmission system within some area, which in turn had responsibility for dispatching power stations in merit order, thus securing the least-cost generation of electricity.

### **Lessons from history**

The main differences to be observed in types of regulation across countries stem from the different solutions countries found to the problem of breaking out of the constraints of the local municipal-based undertakings. The British story is perhaps the most dramatic in the variety of structural reforms that have characterized its evolution. Other countries have typically adopted a more evolutionary approach to regulation, although several have found nationalization necessary to achieve the required structure to support subsequent coordination in investment and operation. The history of the British industry is therefore discussed below in more detail than others because it brings out more clearly the conflicting interests that must be balanced by any regulatory solution.

#### *Britain*

The history of the British electricity industry can be divided into four phases.<sup>9</sup> Until 1926 the industry was decentralized and uncoordinated, with generation under both private and municipal ownership subject to loose regulation laid down by statute. The creation of the Central Electricity Board as a public corporation in 1926, set up to build the high-tension grid, marked the start of the second phase. During this phase the industry reaped some of the benefits of coordination by public ownership of part of the natural monopoly element, with mixed ownership in generation and distribution. Nationalization in 1947 was the only way to overcome the problems of rationalizing municipally and privately owned local distribution companies and to resolve the problems created by the maturing of the franchises of private undertakings. The industry was successfully restructured, but the system of regulation was less satisfactory, reflecting an inefficient equilibrium that only privatization appeared capable of upsetting. Although privatization in 1990 resulted in substantial changes in the structure and operation of the industry, it raised again the critical question

of whether the benefits of increased competition outweighed the difficulties of achieving the benefits of coordination.

These coordination benefits are very clear from the earlier history of the industry (Hannah 1979). Before 1914 the large number of locally restricted producers faced a fundamental problem: cheap electricity at a price low enough to create adequate demand required integrated distribution and large generating stations under single ownership as natural monopolies. Existing municipal undertakings could not expand into neighboring jurisdictions and would not permit private generators to take them over. Relations between the public and private sector were perhaps more strained than in other countries, and the debate over public ownership more vigorous and polarized. These shortcomings became apparent during the First World War, and in 1917 a Reconstruction Committee recommended that the 600 undertakings be replaced by large power plants in 16 districts, a step that might halve the cost of power. This recommendation failed because rationalization would require either powers of compulsory purchase or nationalization, neither of which was politically acceptable.

A subsequent inquiry in 1925 produced a damning indictment of the power of local interests to block technical improvements. The committee argued for a national grid and suggested an ingenious compromise to the conflict between public and private interests: the Central Electricity Board should build and operate the grid, and existing companies should build and operate stations and distribute power locally. New investment would be coordinated by the board, as would dispatch. Such a proposal was presented to Parliament and bitterly opposed in 1926, although no private assets were to be transferred to public ownership. The proposal finally passed in December with the Labour Party's support after the General Strike of 1926. After passage of the 1926 act, supply expanded rapidly: between 1929 and 1935 output of public-supply undertakings increased by 70 percent, despite the Depression. Given the capital-intensive nature of electricity supply, only an estimated 48 percent could be financed out of profits. There appeared to be no difficulty in raising capital for what was a prosperous and rapidly expanding regulated monopoly.

If the Central Electricity Board was a notable success, the hope that the numerous distribution companies would voluntarily agree to merge and coordinate their activities was a disappointing failure. The political debate between the Conservatives who argued for voluntary mergers and those who argued for enforced reorganiza-

tion under public ownership was suspended during the Second World War. Originally, private generators had been granted 42-year franchises; these were maturing and could be acquired by municipalities. During the war the franchises were put on ice, but after the war the incoming government was faced with the choice of either nationalization to impose a sensible, coordinated distribution system or increased fragmentation among municipalities, which seemed incapable of rational cooperation. Public ownership at the national level was thus a superior alternative to public ownership at the municipal level.

One might conclude that nationalization was forced on the industry by the initial franchising provisions and that the Conservative Party was happy to acquiesce in the forced reorganization, although the party's individual members might have been unwilling to underwrite nationalization. The industry was nationalized by the Labour government in 1947, and for most of the post-war period the Central Electricity Generating Board operated all generation and transmission in England and Wales as a vertically integrated statutory monopoly.

The main weakness of nationalization was that it had no clear objective to guide its policy once it had achieved the initial task of rationalizing the industry, which occurred rapidly in the post-war reconstruction period. This failure to specify clear objectives is symptomatic of a deeper problem. Public ownership inevitably allowed the various interest groups a stage on which to influence outcomes and thus ruled out the pursuit of any simple, single objective. These interest groups included not only the management and unions within the industry but also those of the coal industry, whose fate was inextricably linked with that of electricity.

Coal supplied 80 percent of the fuel for generation in 1960, and although this share fell slightly over time, in 1990 more than two-thirds of electricity was still generated from coal and power generation took 80 percent of the output of British Coal. The nationalized industries dominated the Trades Unions Congress, which in turn had close links with the Labour Party that had nationalized these industries. Domestic coal has thus been protected against imported coal and heavily subsidized, particularly in 1974–92, when subsidies averaged 19 percent of the sales revenue of the Central Electricity Generating Board. Major energy users take one-third of power and have successfully argued for lower electricity prices to match foreign competition. Finally, electricity is essential for every household and voter in the country. Electricity prices and investment demands both have macroeconomic significance and have at various times

been constrained by the government's fiscal position. Prices have been held down to slow inflation and investment curtailed to protect the budget, with adverse effects on the industry. Given all these pressures to hold down electricity prices in the face of high prices for inputs, it is not surprising that the average real rate of return in the whole period of public ownership was only 2.5 percent, well below the average in U.K. manufacturing.

Energy policy was shaped not only by the demands of the employees, suppliers, and consumers but also by the technical characteristics of electricity. Fuel costs are roughly half of total generation costs, and because electricity is nonstorable, security of fuel supply is critical. Britain has thus chosen to favor indigenous coal rather than imported and often cheaper oil. The 1956 Suez Crisis revealed the insecurity of oil supplies and was responsible for accelerating the ambitious and ill-fated nuclear construction program to diversify fuel supply while reducing import dependence. The oil shocks of the 1970s created added concerns about security and further entangled energy policy with foreign policy. Yergin (1992) argued convincingly that the geopolitics of oil made this entanglement inevitable for any major oil-importing power such as Britain.

The General Strike of 1926, and the miners strikes of 1974 and 1984, demonstrated that indigenous fuel supply did not automatically ensure security of supply and prompted repeated attempts to diversify away from coal. At the political level defeating the miners strike in 1984 was a key part of weakening the Trades Unions, thereby altering the balance of political power in favor of the incumbent Conservative Party.

Over 1947–90 the balance of power shifted between the different interest groups, depending on external circumstances such as the oil shocks, Suez, and domestic priorities such as inflation or strikes. Given the varying political importance of these objectives at different times, it is hard to see how specifying the pursuit of a simple objective such as "minimize the (social) costs of meeting demand" would have been feasible. It is therefore hardly surprising that the regulatory framework implicit in public ownership failed the test of economic efficiency.

The Conservative government under Margaret Thatcher had a variety of motives for privatization, and one extremely telling argument for considering the privatization of the electricity industry: such industries operated under private ownership with apparent success in a number of European countries, and certainly in the United States. There was a growing belief that the large nationalized industries, of which the Central Electricity

Generating Board was an excellent example, were inflexible, bureaucratic, secretive, and largely out of political control. The government could commission studies, audit the nationalized industries, and subject them to the searching inquiries of the Parliamentary Select Committees, but it had few sanctions short of denying them access to investment funds or resisting requests to raise tariffs. Such negative sanctions merely increased the inflexibility of the organization and did little to promote an aggressive and competitive industry. Privatization therefore held the considerable attraction of upsetting this very unsatisfactory politico-economic equilibrium and undermining the monopoly power of the coal miners and other nationalized unions.

The government had already successfully privatized British Telecom and British Gas, the second as a monopoly and the first as virtually a monopoly (the tiny competitor, Mercury, was protected from further entry for five years). In both cases aggressive and tight regulation was required to change the corporate culture, and even then the change came slowly. There was growing dissatisfaction with a concept of privatization that transferred public monopolies intact to private ownership, and mounting evidence that competition rather than ownership was the decisive factor in improving economic performance.<sup>10</sup> It was therefore argued that to disturb the unsatisfactory politico-economic equilibrium in an industry as prone to such a variety of pressures as electricity, it was essential to dismember or deintegrate the industry. The separate stages of the previously vertically integrated industry, it was suggested, should be forced to operate in full public view in the marketplace rather than in the obscurity of committee rooms. It was to be one of the most ambitious attempts anywhere to introduce competition into an industry normally considered to be a natural vertically integrated monopoly.

The British government in July 1989 legislated the resulting privatization of the Central Electricity Generating Board and the distribution companies of England and Wales (Scotland and Northern Ireland came later) in the Electricity Act. The board was divided into four parts: high-tension transmission was assigned to the National Grid Company, the fossil-fueled generators were split between PowerGen and National Power, and nuclear power stations were retained in public ownership in Nuclear Electric. All four were vested as public limited companies on March 31, 1990. At the same time the 12 local distribution companies, now to be known as the regional electricity companies, were vested; and the National Grid Company was transferred to the joint ownership of the regional electricity companies, which were

sold to the public in December 1990. Sixty percent of the shares in National Power and PowerGen were subsequently sold to the public in March 1991.

The new structure introduced in March 1990 thus divided the process of electricity supply into four activities: generation, transmission, distribution, and supply.<sup>11</sup> Generation accounts for about two-thirds of the industry's costs, transmission for 10 percent, distribution for 20 percent, and supply for the remaining 5 percent. Supply is further subdivided into sales to a franchise market of smaller customers, restricted to the local regional electric company, and a nonfranchise market of large customers, which can be served by any company acting as a private, or second-tier, supplier. Transmission and distribution as natural monopolies were to be regulated by the Office of Electricity Regulation, but the government argued that there was no natural monopoly in generation providing there was freedom of entry. Such freedom was guaranteed, and the generators therefore were subject not to detailed regulation but to the threat of competition from new entrants, as well as actual competition from each other and with imports from France and Scotland (about 9 percent of the total).

#### *Germany*

Early supply in Germany faced the same difficulties as in other countries of small scale: high costs, and low load factors for lighting, making electricity appear too expensive for power use.<sup>12</sup> As a result large firms supplied power from self-generation (over 80 percent of total supply as late as 1913). Initially, most electricity utilities were privately financed, but the increasing importance and profitability of electricity encouraged municipal participation. The advantages of exploiting cheap coal mines and hydro power and the benefits of economies of scale encouraged larger generating stations supplying over long distances. By the end of the First World War, the distribution system covered the whole of Germany, providing reserve capacity rather than power pooling. The pressure for increased concentration led to mergers, resulting in mixed public-private enterprises, and by the 1920s the present structure of the industry had been largely determined. The present structure in the former West Germany has three types of firms operating at the national, regional, and local levels. Eight enterprises produce and transmit high-voltage electricity interregionally; 41 regional suppliers intermediate between these producers and local suppliers, while producing one-quarter of total supply; and about 1,000 local suppliers serve final demand. The high-voltage grid was jointly operated by nine regional companies by 1930.

The federal structure in Germany appears to have allowed a greater diversity of forms of regulation, and there was a conscious attempt in some states to create public or mixed enterprises to compete with otherwise dominant private companies. The advantage of expansion—driving down costs with scale, allowing the undercutting of rivals, which in many industries would have led to very concentrated ownership structures—was thus impeded by regional public interests. As in Britain the tension between public and private ownership and the failure to secure the potential benefits of wider area coordination created pressures toward nationalization at the federal level, but the combined power of private industries and the federal states was able to resist this pressure. One measure of the potential costs of such a highly fragmented system is that as late as 1961, 38 percent of total electricity supply was produced by industrial self-generation. Another measure is that in 1991 the maximal inter-regional difference in high-voltage energy prices was more than 40 percent (Müller and Stahl forthcoming). The number of public utilities decreased from about 16,000 in 1933 to 3,000 in 1955 and to 1,000 in 1987.

Coordination and regulation are devolved to the federal states, which are able to take advantage of yardstick comparisons between companies in determining justifiable cost-based prices. Investments must be approved and licensed, which reduces the risk of inefficient or duplicative investment but does not guarantee least-cost expansion. The equilibrium is one in which the security of investment to owners is high, and as a result of the ability to pass through costs to final consumers, especially domestic customers, the cost of finance therefore low. The municipalities with their ownership stakes support this system because they are able to participate in profits, which can be used to finance other local services. Entry is difficult, and there is no third-party access to transmission, restricting competition from potential suppliers. Competition from gas is muted by common ownership in the energy companies.

The system is well designed to finance the rapid expansion that has characterized the industry until recently. In periods of rapid demand growth and low reserve margins, coordination inefficiencies are probably low, providing new investment takes place in large, efficient stations. With lower growth in demand and the emergence of excess capacity, the regulatory system is put under some stress because high profits are no longer required to finance investment. The increasingly evident disparity in the costs of German coal and French nuclear electricity,

not to mention imported gas, implies that some customers in Germany could be supplied at considerably lower prices than those currently charged. Whether these pressures will be sufficient to force third-party access to the grid, with the potential for consequent rationalization in tariff structures and in generation, remains to be seen (Newbery forthcoming).

The main guarantees provided to investors by the German system of regulation lie in the diversity of interests that must be satisfied to maintain support, ensuring that efficient companies will be able to prosper. But their security is further buttressed by the cartelized structure of the industry and its exemption from much competition law. Germany has been successful in evolving a system that continued to meet the demands of its customers without radical reform—or perhaps because it would have required nationalization to carry through substantial reforms, which was more difficult in a federal system. At a deeper level the electricity supply industry exhibits many of the characteristics of the German social market system that has evolved to facilitate long-term relationships and reduce risks created by short-run competitive market pressures. The system of devolved regulation and restrictions on competition is largely insulated from federal government pressure under the *Ordnungspolitik* system set up after the last war, facilitating the cooperation, coordination, and investor security needed for investment in a fragmented industry, at the possible cost of lower competitive pressure.

#### *Scandinavia*

The Scandinavian countries, particularly Norway and Sweden, differ from most of Europe in the importance of nonfossil fuel power.<sup>13</sup> Norway is 99 percent hydro. Sweden is 50 percent hydro and 46 percent nuclear. Finland is 21 percent hydro and about 40 percent nuclear. Denmark, by contrast, is almost wholly fossil fuel-dependent. The Scandinavian model is characterized by cooperation and self-regulation between both public and private producers and, until very recently, gradual evolution rather than any periods of restructuring. The grid in each country is owned by a state company: in Sweden, Vattenfall; in Finland, Ivo; in Iceland, Landsvirkjun; and in Norway, Statkraft. In Denmark the more than 100 distributors formed coordinating boards, ELSAM and Elkraft, which operate generation and transmission systems in the two geographically separate systems. Thus all Scandinavian countries have evolved systems for dispatch and for coordinating power transmission and generation investment, unlike Germany and the United States.

The development of the Swedish system is instructive. Low-cost hydro resources in the north prompted the development of a high-tension grid, primarily constructed by Vattenfall, which covered the country by 1938, although formal responsibility for planning and operating the grid was not granted by the Swedish Parliament until 1946. Distribution was decentralized, with most towns initially owning their own utility; concentration has reduced the number from 2,000 in 1957 to 300 today. The present structure of generation is 55 percent state (Vattenfall), 25 to 35 percent private, and the balance municipally owned companies. The National Grid Committee consults on coordination, although it is effectively ruled by a club of the 10 to 15 largest producers and delegates operation to Vattenfall.

Norway, by contrast, has a large number of small producers and small distributors. About one-half of the market is serviced by 25 vertically integrated utilities, and most of the remainder have long-term relationships with wholesale power companies. All members of the Norwegian Power Pool have wheeling rights (the right to transmit power to their customers) on the state-owned grid. The Norwegian Electricity Board has responsibility to coordinate expansion, but local interests and environmental lobbies have effectively obstructed least-cost coordinated expansion. The fragmented nature of the industry and the locally negotiated long-term contract prices for many major energy users lead to high price dispersion and potentially large allocative losses [estimated at \$900 million per year (Bye cited in Hjalmarsson forthcoming)]. Sweden appears to have achieved greater allocative efficiency with its more concentrated structure.

Despite the relatively large number of generating companies, the Scandinavian countries—with the possible exception of Finland, which experienced duplication of the grid—appear to have solved the problem of coordinating large expansions in capacity required to achieve economies of scale by negotiations and swapping electricity over time. The main inefficiencies appear to arise at the distribution level (where the large number of utilities exhibit a wide range of efficiency), although the Norwegian requirement that each producer have adequate reserve margins (which can be based on long-term contracts with other suppliers) appears to have led to excess (and excessively costly) generating capacity (Moen 1994).

In Norway, Statkraft, as the largest and state-owned generator, acts to balance supply and demand and exercises an indirect regulatory constraint on the system. The large number of members co-responsible for the grid

appears to have led to a rather unsatisfactory grid tariff structure, which was reformed in 1992. The Norwegian system has been under pressure from large electricity users to keep prices low at considerable national cost by restricting exports to neighboring countries. Recent deregulatory moves in Norway have apparently degraded the quality of information on the state of reservoirs, which was previously essential for the efficient planning of water release over the season. This situation suggests that the rather decentralized and parochial Norwegian system is still facing problems in achieving the full benefits of coordination. In Sweden the clearer delegation of authority to Vattenfall to plan the expansion of the grid appears to have led to more satisfactory pricing and expansion than in Norway.

It is interesting to contrast the Swedish and Norwegian systems. Sweden had a sophisticated system of expansion planning and managed to exploit hydro resources in least-cost order. It also had a remarkably successful nuclear program that kept to cost and was economic against alternative hydro. Norway, with its more fragmented structure, was less successful in managing its expansion (Hjalmarsson, forthcoming; Rinde and Strom cited in Hjalmarsson forthcoming; Segelod cited in Hjalmarsson forthcoming). More recently both countries have exhibited the typical tendency to overinvestment when demand growth slackens and the self-financing constraint no longer bites (Hjalmarsson forthcoming).<sup>14</sup> This tendency is in part associated with access to cheap finance from state sources, substantial cash flow generated by low running costs of hydro and nuclear-based systems, and the prevalence of companies dominated by engineers more interested in investment than economic efficiency. The lack of representation of consumers (other than the major energy users) and of direct competition between geographically dispersed distribution companies supports reinvestment, as does the ethos to produce rather than trade reflected in the conservative reserve margins and the lack of export orientation.

Hydro systems appear to have the advantage of forcing coordination in the construction and operation of high-tension transmission systems. Hydro provides incentives for cooperation, long-term contracting, and self-regulation where the interests of the public are usually represented by municipal or state ownership of parts of the system. The main problems arise when transmission and expansion are not integrated at the country or even multicountry level. The other problem with self-regulation, also found in state-owned systems, is a tendency to excessive investment. To that extent rate of

return-regulated private industries, state-owned industries, and the mixed systems found in Scandinavia and Germany have much in common. They differ, however, in their success at achieving efficiency in distribution and least-cost expansion.

#### *United States*

Until recently most studies of the regulation of natural monopolies were based on U.S. experience. It is therefore not surprising to find a close match between the history of the U.S. industry and the concerns of the regulatory literature.<sup>15</sup> High gas prices gave electricity an initial advantage over gas for lighting in the United States not available in the United Kingdom, which enabled scale economies to be reaped more rapidly. Competition between private suppliers was consequently intense, often leading to financial failure followed by municipal takeover. As elsewhere tensions between private owners and local politics were only resolved once a stable regulatory framework was evolved in the period after 1906, the date of the first commission in Wisconsin to offer long-term franchises.

Rivalry between government and investor-owned utilities was strong, as in Britain and Germany, but constrained by the federal system. The federal government, which owned the hydro resources, was limited to supplying municipal utilities with cheap power and stimulating (by subsidy) rural electrification. Investor-owned utilities had secure property rights under the system of rate-of-return regulation and geographically distinct, vertically integrated franchise monopolies with an obligation to serve but no obligation to provide third-party access to transmission. The regulatory contract protected profits and hence lowered the cost of capital, as in Germany, while limiting monopoly abuse and hence sustaining political support for the system as it continued to deliver rapid growth and cheaper power. Tensions arose as a result of the unfortunate coincidence of a rapid escalation in nuclear construction costs caused by safety concerns and a fall in the growth rate of demand after the 1973 oil price rise. Together, these stopped the steady addition of large stations that the system was well set up to finance. Reserve margins rose from 20 percent in 1972 to more than 40 percent in 1982; they fell back to 23 percent in 1990 only because the industry virtually ceased to construct large power plants.

If regulation allowed for cheap finance, it provided little incentive for mergers to lower costs, because costs could be passed through to consumers. As a result, the number of investor-owner utilities only fell from 412 in

1938 to 236 in 1968, many of them failing to fully reap scale economies. With no effective political power for nationalization, achieving the benefits of coordination in generation, transmission, and distribution could not even rely on the self-interested pursuit of profit because of the system of cost-based regulation. The only competitive pressure comes from private power generation and, more recently, from the modest deregulation encouraged by the Public Utilities Regulatory Policies Act of 1978.<sup>16</sup> The resulting inefficiency takes various forms: rates deviate from marginal costs and may cause losses of 7 percent of costs (Gilbert and Henly 1991); employment may be 20 percent too high (Kahn and Gilbert forthcoming); and investment costs may be 10 to 15 percent too high (Kahn 1991). Although these last two inefficiencies are considerably lower than those in Britain under the Central Electricity Generating Board, tariffs in Britain were better related to costs, and the benefits of integrated dispatch were reaped.

Pressure for regulatory reform came in part from the comparison of generation costs of new CCGT, or coal plant, with the regulated prices, which were designed to recover the average costs of past, not necessarily least-cost, investments. As long as the utility owns and controls transmission, competition from new entrants is muted and largely confined to the sale of surplus private power or contracts to supply the utilities. The recent history of attempts to increase competition within a vertically integrated industry illustrates the overwhelming difficulty of finding a transparent, efficient, and stable system of access pricing for the right to transmit. That difficulty, together with the difficulty of arranging mergers or cooperation between suboptimally sized or located utilities, suggests the need to examine carefully any opportunity to restructure the electricity supply industry in a country, to ensure that the choice made does not preclude the future benefits of deintegrating transmission and dispatch from generation and distribution.

It remains to be seen if a shift to price-cap regulation and away from rate-of-return regulation will provide adequate incentives to integrate across companies without prejudicing their ability to finance investment cheaply. Although rate-of-return regulation is an attractive way of underwriting property rights and ensuring the ability to finance expansion cheaply, it appears to do so at the cost of inflexibility, risk aversion toward new technologies that may lower costs, and a reluctance to cross institutional boundaries to seek out lower-cost solutions, as well as the usual tendency to overcostly investment and employment.

*Latin America*

Spiller and Martorell (forthcoming) contrasted the success of Chile in attracting private investment into the electricity sector with the relatively poorer performance of Argentina, Brazil, and Uruguay. They observed that private investors will be reluctant to sink investments that will be subject to regulation in the absence of adequate safeguarding institutions, and that in such cases government ownership is the only viable solution. The safeguarding institution might take the form of a strong and independent judiciary, capable of upholding property rights and open to appeal from the utilities, or alternatively, an independent regulatory institution that is protected from political interference, perhaps by constitutional checks and balances or by distributing responsibilities between state and federal levels. In some cases the system of government may have evolved suitable continuity and stability, as in Mexico; in others minority coalition governments may find it difficult to obtain consensus to change the system of regulation, so that a history of weak coalitions may provide the safeguards.

Rapid demand growth reduces the need for such institutions but was not available in Latin America in the 1980s. It is therefore not surprising that Argentina, Brazil, and Uruguay had predominantly publicly owned utilities with minimal private-sector participation in 1992.<sup>17</sup> Political pressures on state-owned electricity utilities normally encourage underpriced electricity relative to long-run marginal cost, underinvestment in capacity (particularly in periods of macroeconomic stress such as Latin America experienced in the 1980s), and tariffs slanted toward favored groups, often the urban voters. All these tendencies could be seen in Argentina, Brazil, and Uruguay.

Chile, in contrast, undertook radical reform of the electricity sector in 1978. Nationalization of Chilectra in 1970 was followed by rapid inflation, a failure to adjust tariffs, and hence serious deficits at a time when the state had assumed responsibility for all electricity supply industry investment, totaling about \$200 million per year (Covarrubias and Maia 1994, B1.17). After the change of government in 1973, losses continued because of suspected inefficiencies and powerful unions; losses were addressed by tariff reforms and an attempt to make the utilities behave in a more commercial fashion. The 1978 Decree-Law 2.224 created the National Energy Commission and initiated a program of reform aimed at both deintegrating the industry to introduce competition into the power market and separating the state's com-

mercial and regulatory functions. After passage of the new electricity law in 1982, the two state-owned integrated companies, ENDESA and Chilectra, were divided into separate generation and local distribution companies. ENDESA was divided into five separate generating companies and eight distribution companies. The interconnected transmission system was placed under ENDESA's umbrella, giving that generating company potentially preferential access.

A new system of regulation was put in place in 1980 and formalized by law in 1982. The system of regulation, managed by the National Energy Commission, consists of government ministers under the Office of the Presidency. The system of price regulation is based on long-run marginal cost, itself determined by relatively simple computer models according to specified formulas, doubtlessly greatly aided by the fact that Chile is 60 percent hydro and has adequate storage hydro to buffer the price of energy over the course of the day and possibly longer. Transmission and retail prices are regulated, while large users negotiate directly. The energy commission lays down rules for dispatch of generation managed by the Economic Load Dispatch Center, which in turn estimates marginal generation costs used for settlement. Companies are free to invest in transmission and generation, and the energy commission plays a coordinating role. The system of regulating prices has survived financial turmoil and encouraged adequate investment in generation, transmission, and distribution. If the National Energy Commission wishes to depart from the formulaic rules, it can do so only with the approval of the minister of economics—and then subject to judicial appeal if companies can demonstrate that the new prices are below long-run marginal costs.

The restructured companies were subsequently privatized, and by 1991 there were 11 power-generating companies, 21 electricity distribution companies, and two integrated companies. Galal and others (1994) presented a social cost-benefit analysis of the privatization of Chilgener, one of several competing generators, and Enersis, a monopoly distribution company, both created from Chilectra. Chilgener increased its profit, investment, and productivity after divestiture; the improvement in Chilgener's profit was due mainly to a move to marginal-cost pricing and increased capacity utilization, both due to improved regulation rather than divestiture. Galal's most plausible estimate showed that the present value of world welfare was 4 billion Chilean pesos (Ch\$4 billion), 21 percent of the private value of Chilgener. Of this total, Ch\$2.7 billion went to foreign shareholders,

Ch\$3.8 billion went to domestic shareholders, and Ch\$0.1 billion to employees. Consumers saw no change in welfare, and the government felt a loss of Ch\$2.7 billion. On the alternative view that none of the post-privatization productivity increase was due to privatization per se, the government loses Ch\$6.6 billion and the country loses what the foreign shareholders gain—Ch\$2.7 billion, or 14 percent of the private value of the asset. Whether privatization of Chilgener was beneficial to Chile thus depends on whether one holds the view that privatization was essential to achieve the productivity gains or the view that these gains would have occurred as a result of earlier regulatory reforms.

Enersis, unlike Chilgener, is not subject to competition, and its external regulatory regime did not change with privatization. Nevertheless, privatization encouraged the company to reduce losses from theft and improve returns on nonoperating assets. The gainers were domestic shareholders (Ch\$40.7 billion), foreign shareholders (Ch\$2.2 billion), and paying consumers (Ch\$17.5 billion). The losers were nonpaying consumers (Ch\$9.8 billion), the government (Ch\$5.6 billion), and Chilean citizens (Ch\$26.3 billion), all of whom lost the opportunity to receive the gains from selling its holding in the other divested company, ENDESA. Those gains were instead captured by Enersis and passed to its own shareholders. The net benefit to Chile was Ch\$16.3 billion and to foreigners Ch\$2.2 billion, or together 31 percent of the private value of Enersis. Privatization was again costly to the government and resulted in considerable redistribution (some desirable, from nonpaying to paying customers), but more of the gains were captured domestically than in the case of Chilgener (Galal and others 1994).

Regulatory reform was clearly the major determinant of improvements in Chile, although there were additional gains from privatization (Galal and others 1994, p. 542). On the pessimistic side, there are concerns that investment has tended to be in small-scale generating projects (all except ENDESA's plant of 447 megawatts at Pangué are less than 150 megawatts), while ENDESA's dominant market position through ownership of the grid may be creating similar worries to those of the major British generators (Covarrubias and Maia 1994, B4.13–28).

The sequencing of reform in Chile is instructive in that the reform of the regulatory system and the restructuring of state enterprises occurred first, to ensure that the new enterprises had some experience with the regulatory regime before privatization. Privatization proceeded slowly, avoiding some of the risks of underpricing with

attendant larger transfers to shareholders, and wide share ownership created political support for the new system. By 1990 about 62 percent of the total value of the electricity supply system was in private ownership (Covarrubias and Maia 1994, B4.1).

Perez-Arriaga (1994) has described the recent reforms in Argentina. Radical reforms starting in 1992 and continuing to 1994 transformed the structure, ownership, and regulation of the electricity supply industry. As in England privatization de-integrated the industry into generation, transmission, and distribution, with the aim of introducing competition and realigning tariffs with marginal costs under a system of price regulation. As in England there is a bulk supply market in Argentina, although it differs in that the spot price is computed from generation costs that are audited rather than from submitted bids. Distribution is regulated as a natural monopoly, and generators are restricted to control less than 15 percent of total generation. By the end of 1993, 70 firms were trading in the bulk supply market, and the largest generator had less than 8 percent of total capacity.

The operation of the bulk electricity market is managed by CAMMESA, representing generators, transmission owners, distribution companies, large customers, and the secretary of state for energy. The regulator, ENRE, awards licenses, determines tariffs, and resolves disputes, subject to supervision by the secretary of state for energy but with considerable autonomy. As in Chile there is considerable emphasis on using computer models to compute prices, and distribution uses differentiated nodal prices to reflect regional cost differences. The obligation to supply is placed on distribution companies, which must ensure adequate contracts to meet their obligations and can initiate investments in transmission capacity.

The Argentine electricity reforms suggest that the lessons of radical restructuring in England can be applied to state-owned systems in developing countries providing sufficient care is taken to design the structure, the markets and their operation, and a system of regulation. The Argentine solution is more sophisticated than the English solution in the management of the bulk electricity market, the contestability of transmission, and the attempt to base prices on costs rather than bids, thus potentially reducing the market power of generators. The Argentine regulatory system, like that in England, is buttressed by licenses that can be protected through the courts, and the impact of its reforms has been similarly dramatic in increasing the efficiency of generation. It remains to be seen whether the rather complex system of regulation and

price setting can achieve the benefits that their designers anticipated, and it is too soon to judge the robustness of the regulatory system against political intervention or economic crisis. That robustness will determine whether the industry is capable of financing investment. If Argentina's program of privatization is successful, and electricity demand grows rapidly, private investors' fears are likely to be assuaged.

Both Chile (with 5 gigawatt capacity) and Argentina (with 18 gigawatt capacity) have a sufficiently well-diversified generation portfolio to contemplate separating generation from transmission to achieve a competitive, unregulated private-generation sector (although Chile may be somewhat on the small side). Costa Rica, with only 1 gigawatt of capacity, provides a more typical example of a smaller country in which deintegration and full privatization may be less attractive. As in many other countries, initial development in Costa Rica was by private concession, but rapid post-war growth and power shortages prompted public involvement, initially through the ICE (Instituto Costarricense de Electricidad), set up in 1949 to plan, implement, and coordinate electricity supply (Covarrubias and Maia 1994). Investment initially was undertaken by a foreign private firm, CNFL, which was acquired by the ICE in a friendly takeover in 1968. Costa Rica has a relatively independent regulatory body and, until the debt crisis of the early 1980s, had cost-recovering tariffs (with the usual cross-subsidies from commerce to domestic customers). Rapid inflation, attempts to buffer domestic tariffs, and foreign-financed investment created a fiscal crisis (electricity debt was 15 percent of total foreign debt) and resulted in increases in real tariffs as well as a 1990 law allowing private generators to contract to sell power to the ICE. This arrangement has apparently been successful on a modest scale, with 8 megawatts (less than 1 percent of capacity) available for dispatch in 1992. It is anticipated that up to 100 megawatts of 1,000 megawatts of new capacity in the 1990s might be private. Enthusiasm for selling state-owned generation to the private sector seems low, so the benefits are likely to derive mainly from optimizing auto-production, and possibly by exposing ICE generation to some contestability, as is happening with U.S. utilities.

Given Costa Rica's small size, its heavy dependence on site-specific hydro power (whose rents are often deemed to be national assets most readily exploited in the public sector), and its past success with the ICE and its form of regulation, such modest changes appear quite defensible. Deregulating generation may just facilitate the exercise of market power. Competition in building new capacity is

likely to yield the larger part of available efficiency gains, and the recent reforms facilitate that step.

#### *Pricing problems of different regulatory systems*

The electricity supply industry is capital intensive and durable, but it is subject to fluctuating demands. Short-run marginal costs therefore may vary widely and rapidly, and will bear a tenuous relation to average costs. Each regulatory system must balance the need for efficient pricing against the need to finance investment. The question that must be answered is, who will pay the difference between variable costs and total costs? Gilbert, Kahn, and Newbery (forthcoming) observed that most countries appear to cross-subsidize residential consumers from commercial customers, an approach that is certainly consistent with their respective political power. In most former soviet-type economies, residential consumer prices were substantially below their economic level—perhaps because of the perceived political difficulty in raising such prices in line with inflation—and a complex web of cross-subsidies was required to achieve this pricing arrangement. In some countries underpriced imported electricity provided the necessary revenue, whereas in others industrial prices were kept adequately high to generate revenue. In most, however, the overall effect was that the average price level failed to produce an adequate return on total assets. In effect the state subsidized investment. This was also true in Britain, although a large part of the subsidy was effectively for coal rather than electricity. Countries with large hydro resources usually failed to price them at scarcity value, whereas investor-financed systems typically underrewarded bondholders in the inflationary post-war period. The more decentralized systems often were able to charge domestic consumers higher prices relative to industry, as electricity costs are a smaller fraction of domestic budgets than some industrial budgets, while decentralization may defuse the salience of pricing decisions.

Observations such as these suggest first, that different regulatory systems may face very different pressures in balancing the claims of financing investment and efficient pricing, and second, that any regulatory reform therefore should ask how this balance is likely to be achieved. The next section considers the choices available and sets out the criteria for a good system of regulation.

#### **Structural choices and the design of regulation**

Reforming the electricity supply industry will raise quite different problems in countries with nationalized industries (owned and controlled by the central rather than

local government) than in those with either private (investor-owned) industries or mixed systems. Radical restructuring is far easier under public ownership; it may require clarifying the state's control over industries in countries with unclear or overlapping property rights (between workers, local municipalities, and ministries in some Central and Eastern European countries and in most developing countries). Regulatory reform without direct control over assets will be constrained by the rights of existing owners and is likely to be largely dictated by the special features of the case in question. The discussion here is restricted to the reform and possible restructuring of a nationalized or state-owned electricity supply industry.

The body advising on reform of a nationalized or state-owned electricity supply industry must answer three questions: How should the industry be structured? Which parts should be public and which private? Which parts should be regulated and how? This section addresses each question in turn.

#### *Structural questions and the need for coordination*

If generation is potentially competitive and transmission is a natural monopoly, the structure of the industry becomes an important determinant of performance and the form of regulation. At one extreme the industry could be vertically integrated from generation down to the final consumer under a single owner (either public or private, within some region), as in the United States. At the other extreme, as in England, the industry could be vertically deintegrated, with numerous unregulated, privately owned, competitive generators bidding for dispatch; with the grid under single, private-regulated ownership, delivering to a number of regulated private distribution companies; but with supply (buying, selling, and billing electricity) open to competition. Intermediate solutions might involve public or club ownership of the grid. If the country has limited capacity, or transmission constraints that limit the number of competing generators (including those in neighboring countries), generation cannot be deregulated safely. Generation then will have to remain either vertically integrated or under contract to a regulated or state-owned transmission company. Such a system will have much in common with a vertically integrated industry.

There are advantages and drawbacks in either extreme choice. If the industry is retained as a vertically integrated structure, regulation can be confined to price regulation of the basket of final products; the industry is free to choose the most efficient organization and coordination

of production and distribution. If the industry is vertically deintegrated, the network must be subject to separate regulation and the terms of access and use must be regulated. The problem here is that the grid offers a variety of different services—transmission to customers, access to insurance against power failure, freedom to schedule maintenance—that vary over space and time and whose costs are jointly determined. Efficiently pricing these services, and at the same time giving the right price signals for investment decisions (for grid expansion, plant location, and customer location), is inherently difficult. Devising a satisfactory set of prices (or judging whether a set of price proposals is satisfactory) is a challenging task for the regulator, given the natural monopoly nature of the grid.

The evidence from the English National Grid Company, which is jointly owned by the privatized regional electricity (distribution) companies, is that it is peculiarly difficult to get the relevant prices right, particularly as marginal cost will typically be below average cost, so that marginal-cost pricing will not cover total costs. In such cases nonlinear pricing is typically preferable to uniform pricing but difficult to regulate adequately. Compared to a vertically integrated industry, the outcome when the industry is vertically deintegrated is likely to be one in which intermediate transactions are less efficient. In the case of the National Grid Company, there were two main inefficiencies: the locational decisions by new generators were guided by inappropriate price signals, and there were no incentives to coordinate investments in generation and transmission to minimize total system cost.<sup>18</sup> Some of these problems are being addressed in a series of regulatory reviews.

Vertical deintegration, however, allows competitive pressures at stages where entry is feasible and may result in overall improvements in efficiency sufficient to offset the inefficiencies of transactions through the network. Vertical deintegration hinders cross-subsidization and makes pricing more transparent, which in turn may lead to close scrutiny of the value of such services as reliability, security, and national self-sufficiency.

What are the implications for a government considering a potentially radical restructuring of the electricity supply industry? Such an opportunity is rare because it requires some kind of crisis or large-scale political change. If there is doubt about which system is best in a particular country, one of the main considerations should be whether a particular structure forecloses options. If deintegration is possible, there is a good case for choosing this option—or keeping the option open by continued public ownership

of the transmission system. Continued (central) public ownership keeps open most options, whereas municipal ownership appears to create considerable obstacles to further reform—at least in some political systems. Private ownership of vertically integrated generation appears to be the most difficult system to reform because it requires overriding private property rights. If generation is to be transferred to private ownership, transmission should certainly be kept separate, initially in public ownership or as a separate company with restrictions on control by generators or individual large users or distributors. Such control not only would reintroduce vertical integration but also might allow foreclosure or other predatory actions if only some generators or distributors controlled the grid.

A key issue is whether a bulk electricity market should be created, and if so, how prices should be set and who should be allowed to trade at these prices. One extreme, represented by the English pool, is that the pool prices all electricity and the spot price clears the market (every half-hour), although agents can hedge this price through contracts. Since contracts cannot deviate far from the average spot price, the spot price is the relevant price for all generators and customers. Any license holder is free to trade at these prices, and all generators above a very small size are paid this price.

The other extreme is to restrict the market to distributors that either are vertically integrated into generation or have contracts for supply, and that are required to have adequate capacity or contracts to back their demand on average. Here the market is used for short- or medium-run balancing—essentially the U.S. and continental solution. In this case the exact mechanism for determining the price is less important so long as it ensures that payments will balance out over time with no advantage to any party. This arrangement encourages a club-like approach to operation, dispatch, and planning and is consistent with self-regulation. The requirement to match demand and supply by company potentially forgoes the benefits of trade and often leads to excess capacity.

The first solution is intended to cope with a deintegrated industry that avoids balancing trades. The price therefore is of critical concern: high prices favor generators at the expense of consumers, and low prices conversely. One key issue is whether these prices should be based on bids (as in England), possibly unrelated to the short-run avoidable cost needed for efficient dispatch, or audited costs (as in Argentina), with some other device to cover fixed costs. The advantage of a spot market is the competitive pressure placed on generators, but at the cost of either inefficient prices or unstable investment finance.

An intermediate solution may act as a cautious step toward the benefits of deintegration (and preserve that option) while leaving the problems of pricing for later resolution: the expertise previously in the integrated generation and transmission company could be transferred to a new body with control over the grid and responsibility for contracting with now-separate generators, organizing dispatch, and selling to the distribution companies and large consumers. The contracts with the generators would specify fixed and variable costs, and the second could be used to determine the merit order and the spot price—although access to this spot price (and consumers) would be restricted (perhaps to noncontracting generators) by the monopoly ownership of the grid. Consumer prices would be determined much as before to recover system costs. The advantage is that there is greater competition between generators in bidding for contracts, and the dispatch price is delinked from the means of covering costs, although at the expense of limiting direct access by generators to consumers. The drawback is that there is little pressure on the grid and the distribution companies to improve efficiency, although the system lends itself to subsequent liberalization. There are some similarities with the U.S. system of encouraging utilities to seek bids from independent generators. Short of regulatory pressures, however, there may be little to force utilities to subject their own generators to such competition.

#### *Public or private ownership?*

Vickers and Yarrow (1988) surveyed the substantial literature comparing the performance of investor-owned (that is, privately owned) electric utilities with that of publicly owned utilities (state or municipally owned). They concluded that there is little difference between public and private ownership in terms of technical or cost efficiency and cautioned against assuming that public ownership leads to greater allocative efficiency. They argued that allocative efficiency is more dependent on the form of regulation.

Pollitt (1993, 1994) provided the most recent and thorough empirical investigation of electric utilities. He subjected two data sets to exhaustive comparisons of efficiency. The first set was an international sample of 95 utilities operating in nine countries in 1986. Depending on the approach used, Pollitt found evidence for no significant difference in technical efficiency between the two ownership types but some evidence for the superior cost efficiency of private utilities.<sup>19</sup> The second data set was an international sample of 768 power plants in 14 countries in 1989, which together produced about 40 percent of world thermal electricity. This plant-level analysis (using four dif-

ferent methodologies for measuring efficiency) found that private firms are statistically significantly more technically efficient than public firms, once the efficiency scores are pooled (Pollitt 1994). The failure to find significant differences in technical efficiency in the first (and other, earlier) studies reflects the inadequacy of the sample size for detecting rather small differences in measured technical efficiency, reducing cost by between 1 and 3 percent.

To measure cost efficiency, Pollitt then considered 164 of the 213 base-load plants in the data set for which input price data could be found. He rejected the hypothesis that public utilities are as efficient as private ones, finding private utilities to be about 5 percent more efficient both in minimizing costs and overall (the difference varied with the methodology used). The evidence shows that well-run public utilities can at least equal the performance of average private utilities.

Pollitt's careful study is consistent with the view that the more important determinant of efficiency is the degree of competitive pressure put on the utility, which in turn depends on the extent to which a utility must compete for its market, and the quality of regulation, although private ownership appears to provide some additional improvement. Private owners typically perform better in competitive markets, particularly where innovation is important, or least-cost solutions require careful and informed choices, and where costs must be closely monitored. Generation is therefore a natural choice for private ownership, particularly if it is associated with open access to transmission. This combination would allow private enterprises with auto-generation plants to sell surplus power and improve the competitiveness of the bulk electricity market.

In Britain privatizing the generators and forcing them to compete in the bulk electricity market resulted in dramatic improvements in labor productivity by halving the work force within three years. Privatization also resulted in much closer control over investment costs. It is noteworthy that Nuclear Electric and British Coal, publicly owned companies that were both forced to sell into markets facing competition from private firms or imports, also improved productivity quite dramatically. In Argentina generation availability dramatically improved within a short period after the reforms. Central Costanera improved availability from 20 percent to 50 percent with a doubling of output (Perez-Arriaga 1994).

Norway introduced competition into the bulk electricity market and created Statnett Marked (as a subsidiary of Statnett, the state-owned owner of the transmission system) to operate the power pool in 1993, without altering the ownership structure of the industry. The effect has

been to induce substantial trade across former franchise boundaries with a decrease in the dispersion of prices (Moen 1994). In a hydro system like Norway, changes in patterns of supply have negligible effects on short-run costs. It is too soon to tell whether creating an integrated and competitive market will eliminate inefficient local investment in generation and induce moves toward more efficient-sized distribution companies, a large part of the goal of the reforms. In due course the Norwegian example should provide an important test of the relative importance of creating contestable power markets by restructuring the industry, with privatization. Note, however, that the Norwegian system allows private generation to compete with state and municipally owned systems.

The English distribution companies remain natural monopolies, and their performance does not appear to have changed markedly since privatization—although neither has it deteriorated. The same seems to be true in Argentina and Chile (judging from the case study of Enersis reported in Galal and others 1994). Because the distribution companies in Britain do not have large investment requirements, their considerable ability to earn profits in a protected market has not been required to finance investment (as it has for the privatized water companies with their large backlog of replacement and upgrading investment). The evidence from elsewhere is that distribution companies should be large enough to reap economies of scale and should ideally be subject to an element of benchmark regulation. Their role and ownership may also be influenced by the way in which transmission is organized, and the form of the obligation to supply, which in a deintegrated system will have to be devolved to the distribution companies.

High-tension transmission, dispatch, and other ancillary services required for the operation of the whole system present the most challenging problems. Perhaps the simplest solution is to retain this set of activities in public ownership until the rest of the industry has reached equilibrium. If transmission is separated from generation, then either the grid will have to contract for the right to dispatch power or a bulk supply market will have to determine the merit order and wholesale price (described below for England). The English solution was to transfer ownership of the National Grid Company to the distribution companies, which in turn were privatized, so that each distribution company is a part-owner of the grid. The counterpart to this arrangement would be one in which the generators jointly own the grid, but this has the obvious drawback of providing a mechanism for collusion in bidding and dispatch.

A market solution, the natural setting for a competitive generation system, will require not only the development of a spot market but also various contracts and forward or futures markets to hedge the price volatility and provide longer-term coverage. Satisfactory working of these markets will likely take some time to evolve. There is therefore a strong case for ensuring that reforms to these markets and to the structure of charges and allocation of responsibilities can be adjusted after the initial reform. Such a guarantee should be written into any license agreements if the industry is privatized. The evidence in Britain suggests that a whole series of major and minor reforms have been required for the National Grid Company and its operations. Where these have involved only the license conditions, the regulator has been able to make these changes without undue difficulty. It has proved far more difficult, however, to make changes to the pool, a contractual agreement between numerous parties with conflicting interests over which the regulator has no right to intervene except by mutual consent. The lesson here is that if reforms are likely to be necessary, it is important to ensure that the regulator has the power to make them.

The last alternative is a separately owned transmission company (or companies), the solution that Argentina adopted (Perez-Arriaga 1994). Argentina has six transmission companies: one national company (500 kilovolt) and five regional companies (220/230 kilovolt). Argentina prevents transmission companies from controlling a majority of the shares of generation or distribution companies, or major users. The license holder offers 51 percent of the shares to the highest bidder at public auction at the end of each 10-year management period. The original license holder may retain the license, or there may be a change of control, with the original license holder receiving the share receipts.

The other structural characteristic to be determined is whether the transmission company has a monopoly of transmission or generators can construct their own lines directly to customers. A related question is whether all licensed generators above a certain size (10 megawatts in Britain) must submit to central dispatch or can contract for transmission services and sell directly to customers. Such questions bear on regulation, which is addressed next.

#### *Criteria for regulation*

A good system of regulation is one that (a) enables the utility to raise finance for investment at an acceptable cost; (b) provides incentives for efficiency in operation,

pricing (and hence use), investment (in the choice of type, location, size, and costs), and innovation. These requirements may conflict to varying extents. Rate-of-return regulation, which guarantees an adequate return on capital, underwrites the ability of the utility to finance investment and, by reducing risk, ensures low-cost finance. It provides little incentive to reduce the costs of investment unless it is combined with prudential reviews, in which case the investor's security is reduced and the ability to finance may be prejudiced. Prudential reviews also may reduce willingness to innovate. Moreover, cost-based regulation frequently leads to inefficient tariff structures, which in turn may prompt bypass and duplicative investment by other suppliers.

Price-cap regulation provides good incentives for cost reduction, but cost reduction in turn creates pressures for reviews (in England, typically every five years). The prospect of future tightening of price regulation reduces investor confidence, increases regulatory risk, raises the cost of finance, and may prejudice the ability to finance investment if the regulator has no primary duty to ensure the utility's ability to finance investment.<sup>20</sup>

The difference between a cost-pass through system of regulation, such as rate-of-return regulation, and price-cap regulation of the RPI-X form widely used in British utility regulation at first appears extreme. Cost-based regulation appears to tax efficiency gains at 100 percent, whereas price-cap regulation taxes gains at zero percent. Not surprisingly, theorists have often argued for an intermediate tax rate (Laffont and Tirole 1993). In practice the contrast is not nearly so marked, and Armstrong, Cowan, and Vickers (1994) provide the most recent and comprehensive survey.

The main question concerns the determination of the efficiency factor X in the RPI-X formula (where RPI is the retail price index). Most British utility regulation legislation requires the regulator to take account of the ability of the utility to finance its investment. This immediately relates X to a required rate of return on prudent investment and suggests that price-cap regulation is like rate-of-return regulation with a predetermined regulatory lag. Although there is an important element of truth in this view, it is open to the regulator to consider other information in addition to the costs of the regulated utility in determining X at the next review. In making their decisions, regulators in Britain have certainly looked at the experience of similar utilities in other countries.

There are important differences in other dimensions—specifically, that price-cap regulation leaves the choice of relative prices fairly unconstrained, whereas

cost-based regulation typically has more to say about relative prices and is more likely to lead to cross-subsidization. Profits are likely to be more volatile under price-cap regulation, possibly increasing the cost of finance, and increasing the temptation for regulatory expropriation (Gilbert and Newbery 1988, 1994).

If regulation is to be subject to periodic review, as is almost inevitable, the review process must be carefully designed to ensure investor confidence and continued political support to sustain the system of regulation. The British solution is to grant the utilities licenses that clearly specify their rights and obligations, have legal status, and can be defended or enforced in the courts. Regulatory reviews typically require changes in license conditions, which must be agreed between the regulator and the utility or which would require legislative action to override. Such legislative intervention is sufficiently costly that its use is likely only in extreme circumstances.

The normal review is a public process in which interested parties submit evidence for regulatory reviews. The regulator then makes proposals, typically involving changes in license conditions, which the utility can either accept or have referred to an independent Monopolies and Mergers Commission. This commission can make its own recommendations to the regulator. The most difficult cases involve agreements between parties in the industry rather than, as with licenses, agreements with the regulator. Thus reforming the rules of operating the electricity pool—which is an agreement between generators, distributors, and the grid—has run into repeated difficulties. This experience again suggests that the radical step of privatization creates opportunities that, if not thoughtfully evaluated, can foreclose future modifications to the system. Irreversible decisions must be considered carefully. Retaining public ownership guards options for future change.

#### *Regulation coverage and type*

There is little dispute that the natural monopoly elements of transmission and distribution require regulation. There is considerable debate, however, about the design of such regulation. Distribution companies typically offer two types of service. In England the distribution companies distinguish between the franchise (regulated) market (below 100 kilowatts) and the nonfranchise market. For the franchise market the English companies deliver electricity at a regulated set of tariffs, which under price-cap regulation are restricted to the costs of purchased electricity plus a charge to cover the costs of transmission and supply. For the nonfranchise market the distribution com-

panies provide service at a specified cost, and any supplier (including the company itself) can then negotiate contracts with the customer sufficient to cover use of the distribution system, purchases of wholesale power from the pool, and the cost of transmission. In England there is unregulated competition in this market, although customers may choose to purchase at the published tariffs.

Distribution companies lend themselves to yardstick regulation, in which the allowed unit cost is based partly on achieved unit costs in comparable companies and partly on the costs to the individual company. The relative weights of the two components depend on the degree of confidence that the companies being compared are sufficiently similar to provide good information about achievable cost reductions. The main problem to date has been that many inputs are imperfectly measured, especially the capital base. Problems also arise in assessing the nature of terrain, property access, and other critical factors. Different econometric specifications can yield quite different estimates of the contributions of the different inputs, making problematic the derivation of a standard for the particular set of inputs of the regulated company. As data improve and time series lengthen, it may be possible to increase the weight assigned to the companies used for comparison in yardstick regulation. The results of the first regulatory review of the British distribution companies should be instructive in this regard.

In mature systems such as those in most advanced industrial countries, the need for further expansion and investment in distribution is probably small, making the issue of investor reassurance less problematic. The main regulatory issue is likely to be whether potentially competitive utilities such as gas and electricity are permitted to have joint ownership. The evidence from Germany suggests that joint ownership leads to higher prices for both fuels, leading to lower gas penetration rates than might be justified. New developments in metering may make the supply business (metering and billing) benefit from joint metering of gas, electricity, and water. Metering and billing may be best accomplished by telecommunications or cable television companies. Maintaining the contestability of the supply business to allow future mergers is consistent with competition between different fuels, each under different ownership, and may be desirable.

Politically, the main potential problem with regulation in the franchise market is the balance between the fixed charge and the energy charge. The wholesale price of electricity may be less than half the retail price, although electricity losses amount to less than one-fifth. Many of the remaining services (transmission and distribution) are

fixed independent of power flows and in the long run are related to peak demand. The sunk investment of the distribution system has many public good qualities, with the attendant problem of allocating costs. Large fixed charges and low energy charges are regressive and resisted, but privatization provides additional incentives for distribution companies to press for higher fixed charges to make revenue less susceptible to variations in demand. The practical solution is normally to allow discounted prices for bulk electricity use (for heating) and otherwise to rely on the relatively low price elasticity of demand for non-heating electricity to collect the fixed element essentially by markups on the energy element. The benefits from refining franchise tariffs to domestic customers appear low, and most of the emphasis is therefore on ensuring sensible tariff structures for commercial and industrial users. Here there is little difficulty in rebalancing tariffs among connection charges, capacity charges, and energy charges. The main problem is to devise a satisfactory system of access charges.

This problem is particularly acute for high-tension transmission. Transmission in Britain accounts for 10 percent of total costs and a larger fraction of costs for industrial users. In Argentina transmission costs are 20 percent of the total costs. Because transmission is capital intensive and durable, it is important to achieve efficiency in investment. The National Grid Company, which is required to invest to ensure that the system meets specified security standards, receives revenue proportional to average cold spell demand (a measure of peak demand for which the system must be adjusted). The company was vested with a set of use-of-system charges that differentiated between different regions, to encourage new generation to locate in areas of deficit supply. These charges proved unsatisfactory, however, and were revised in 1992 (after about 10 gigawatts of new capacity had already entered).

The main problem in regulating the grid is to provide incentives to its owners to adopt the least-cost solution to generating and transmitting electricity to the correct level of security. Transmission and generation often can substitute for each other and hence require coordination, which means the grid must construct a set of charges that induces the right investment in generation or transmission as appropriate.<sup>21</sup> The British experience is that continual adjustment by the regulator may be required, because many of the changes in tariffs and reallocation of responsibilities involved considerable transfers of revenue between generators and customers, for the whole system is close to a zero-sum game.

Because transmission investments are expensive, the potential for reducing costs may be considerable. Many earlier regulatory systems encouraged overinvestment, usually by providing excessive security and reliability standards. A shift toward more footloose, gas-fired generation at smaller scales should reduce the need for grid investment.

The remaining regulatory question is, should generation be regulated or left subject to competitive market pressures? In the English system the two unregulated, privatized fossil generators dominate the bulk electricity supply market, or "pool," setting the price 90 percent of the time. (The balance is set by pumped storage, which arbitrages by buying when the price is low and selling when the price is high. So even this component ultimately depends on prices set by the duopolists.) The pool was set up as a spot market (more accurately, a "day ahead" market) for the dispatch and pricing of electricity. This spot market, or "pool," is the most radical part of the 1990 reforms. Every morning, generators must declare which of their generating sets will be available the next day and announce prices for each set.<sup>22</sup> The grid dispatcher then computes the least financial cost of meeting the predicted demand and pays the same system marginal price (the bid price of the most expensive set required to operate) to all the generating sets actually dispatched.<sup>23</sup>

Because the pool price varies widely over the day and year, the resulting price risks must be hedged by contracts. At vesting the generators and suppliers were provided with contracts of up to three years duration. Most were contracts for differences, under which a generator receives, in addition to the normal pool price for any sales, a sum equal to the difference between the specified strike price and the pool price, multiplied by the specified number of units contracted. There also is a market for electricity forward agreements, which allow the main components of electricity price uncertainty (such as the pool price between certain weekday hours or the capacity charge) to be hedged up to a year ahead, similar to a futures market.

There has been considerable dissatisfaction with the results because the duopolists' were perceived as having substantial market power in the pool. When the vesting contracts expired, the distribution companies had to renegotiate new contracts. The duopolists' initial offers were unattractive compared with offers by potential entrants, independent power producers building combined cycle gas turbine (CCGT) stations. These independent producers were able rapidly to put together 15-year contracts to buy gas, backed by 15-year contracts

to sell base-load power to the distribution companies—enabling the stations to be financed largely by bank loans with a small equity, typically partly held by the purchasing distribution company. The distribution companies were able to convince the regulator that they had met their obligation to purchase economically, because they could demonstrate that the contracts signed with the independent producers both were more favorable than any then on offer from the majors and hedged the risks of future price rises caused by sulfur limits. The majors failed to respond to this threat by counteroffers of comparable long-term contracts. It is unclear whether they feared that increasing their contract cover would reduce their market power, they were anxious not to forestall entry for fear of a charge of predatory pricing, or they merely miscalculated (all motives probably played a role). The outcome was a remarkably rapid flurry of entry, leading to contracts that could not be overruled in a privatized electricity market.

The lack of competition in generation led to high prices, which induced excess entry. Unfortunately, these entrants supply base-load power, whereas the pool price is set by mid-merit and peaking stations whose ownership remains concentrated in the duopolists (Newbery 1994b, p. 34). The only pressure on prices is that the time-weighted price now must be kept below the price at which further entry would be profitable. The regulator has ruled that pool prices should be capped for two years, until the duopolists divest 6 gigawatts (out of about 40 gigawatts) into a new company (or companies) to increase competition in the price-setting part of the market.

The Argentine solution in many ways is at the other extreme. Because generation was split into 22 companies, the largest of which has less than 8 percent of total capacity, there is the potential for considerable competition. Nevertheless, generators are required to declare their generating costs (which are subject to auditing), and these are used to determine the merit order and dispatch. In England there is no requirement that bids must be equal to costs. The director general of electricity supply, in his 1993 Pool Price Statement (OFFER 1993), concluded that average pool revenues were above medium-run avoidable costs (even including company-level as well as plant-level costs) after April 1993, when vesting contracts had lapsed. Short-run avoidable costs relevant for short-run scheduling decisions and the merit order are below the medium-run avoidable costs on which decisions about plant availability are made.

The problem that must be resolved is twofold: how to arrange for efficient dispatch [based on short-run (24-

hour) avoidable costs], and how to remunerate capital costs (or, more generally, how to establish the difference between long- and short-run avoidable costs). To some extent the problem can be solved by contracts between generators and suppliers. But there is a tension between the requirement that generators sign long-term contracts to recover total costs, freeing them to bid into the pool at short-run avoidable costs to ensure efficient dispatch, and the requirement that consumers be willing to accept long-term contracts at prices considerably above the pool prices used to determine the merit order.

The Argentine and English systems both attempt to resolve this tension by paying for available capacity, given by

$$LOLP \times [VOLL - \max(SMP)],$$

where *LOLP* is the loss of load probability, the risk that demand will exceed capacity; *VOLL* is the value of lost load, which is set administratively to reflect the cost of demand exceeding supply (currently about \$4 per kilowatt hour); and *SMP* is the generating set's bid price. Argentina pays an additional capacity charge for available capacity of \$10 per megawatt hour (during off-valley weekdays), regardless of type of plant. Although on the face of it, cost-based regulation of generation might appear rather unattractive, there are incentives for efficiency in that dispatch is determined by cost, and payment is equal to the system marginal price, equal to the cost of the marginal unit.

Even if the generating industry can be sufficiently fragmented to encourage competition (and Green and Newbery 1992 found that five noncolluding generators would be sufficient in the United Kingdom, assuming that they sold into a unified market), there remains the problem that transmission constraints may make effective market areas rather small, with few competing generators. The Argentine solution of cost-based bidding avoids this problem, whereas the English system allows plants that must run to meet system constraints to be paid their bid, not the system marginal price (which is determined by ignoring constraints). It is open to the National Grid Company to contract with such stations for the services they supply, and the exercise of such market power can be reduced by requiring stations to submit bids that remained in force for some predetermined period (such as six months, possibly with fuel indexing), rather than allowing the exercise of short-run market power. Once again, however, reforming the pool rules (which determine how generators are to bid and be paid) requires

agreement by the members, which include the generators, and such agreement is unlikely.

The other question that must be addressed is whether entry should be essentially free (that is, licenses to connect are issued automatically to those paying the fee and meeting the technical requirements) or should be controlled to avoid excessive entry, an unbalanced fuel mix (due to inadequate attention to security of supply), or other market failures. If access prices to the grid are well designed (a big "if"), and if the contract market is contestable and properly managed (see below), then the case favoring free entry is to underwrite competition and avoid a system that could rapidly revert to the flaws of central planning or cartel cozi-ness. Again there is too little experience to judge how best to manage this aspect, and many reforms under current discussion are reluctant to make the full step to unregulated entry. The issue is of enormous importance because gas generation looks set to command most new expansion in Europe in the near future, which is almost certainly best undertaken by new entrants.

Another longer-run issue not yet addressed is whether generation will ever be allowed to (or can) develop as a truly competitive sector. Given the large fixed costs (and the irreversibility of investment) and the homogeneity of the product, if generation were competitive, prices would fluctuate widely, as in metals markets. Aluminum is an interesting example of a similarly irreversible investment (although the capital cost is a smaller fraction of the total than for electricity). Aluminum prices follow a random walk between buffers, determined by the price at which further investment looks attractive and that at which exit is rational, but between which excess capacity and fierce competition are the normal order. Although the wholesale price of electricity might be very volatile, transmission and distribution margins are stable (and a large fraction of domestic prices), whereas regulation encourages distribution companies to adjust franchise prices only annually. The political problems of volatile spot prices thus may be avoidable (just as the retail price of coffee, tea, or sugar exhibits little of the extreme volatility shown in the spot markets). Whether the shareholders would tolerate the volatility or whether the generators would exert strong pressures to merge or diversify remains an open question.

### **Managing the interface between the regulated and competitive sectors**

Separating transmission from generation is a key reform that enables competitive pressure to be put on generation by allowing entry by other types of generators. The evidence from Argentina and Chile is that this reform is a

natural precursor to privatization and allows mixed solutions or an evolutionary approach to later liberalization. It forces attention to be paid to the need for an appropriate regulatory framework, which is often the critical requirement to place finances and pricing on a rational basis. It encourages, but does not force, transparency in pricing, which undermines cross-subsidization and inefficient averaging (over time or over distance). And it preserves options for future reform and thus is crucial if privatization is planned.

If transmission is separated from generation, many of the problems of managing the interface between the regulated and competitive sectors become relatively minor, compared with the difficulties of regulating access to a transmission system owned by incumbent monopoly generators (possibly vertically integrated through to distribution), as in Germany and the United States. Regulating access charges in a vertically integrated system (such as many telephone systems) creates endless argument between entrants and the network owner and requires forceful regulation if it is to succeed.

The major potential problematic interfaces in the English system concern the distribution companies' ability to hold equity stakes in independent power producers with which they also have long-term contracts. In the United States such an arrangement would be judged illegal as a "sweetheart deal." In England the sanction is that the regulator can refuse to allow the costs of such contracts to be passed through to the captive franchise market if he decides that they do not meet the economic purchasing condition in the distribution company's license. Threats of benchmark regulation would have a similar effect in making the English distribution companies less willing to sign long-term contracts. The balance to be struck here is that without long-term contracts, the market for generation is unlikely to be contestable and the advantages of competition will be diluted. Excessive entry may be avoided, but most of the pressure on bid prices, if that is the form chosen for the bulk-supply market, will be removed. The English regulatory system here appears to be the least unsatisfactory solution.

### **Directions for future research**

One of the most promising forms of policy research is to undertake cost-benefit analyses of regulatory reforms and privatizations. The World Bank-sponsored study by Galal and others (1994), an excellent example, provides two case studies of electricity privatizations (Chilgener and Enersis in Chile). The techniques set out there (and also in Jones, Tandon, and Vogelsang 1990) can be

applied if the post-reform or post-privatization period is sufficiently long. Argentina is an obvious choice for such an analysis in due course. Most radical reforms in the electricity supply industry are quite recent, but it should soon be possible to provide a fairly comprehensive assessment of the British reforms.

The World Bank itself should be well placed to conduct comparative studies of the efficiency of the electricity supply in different developing countries. Such studies, like a World Bank (1993) study for Latin America, should relate differences in performance to differences in regulatory structure, controlling for differences in the availability of fuel, extent of economies of scale, level of development of the country, competence of the bureaucracy, and other relevant facts. Perhaps the most exiting direction for further research is to relate the potential options for regulatory improvement to the political endowments of the country in question, to explore what regulatory reforms might feasibly be introduced and which would mesh with the political, institutional, and legal structures within the country. A more ambitious task would be to ask how these institutional and legal frameworks might be improved to increase the likelihood of successful regulatory reform in particular industries such as electricity.

One practical question for which information and evidence are accumulating is the choice of the appropriate financial structure and sequencing of sales for privatized electricity industries. Several issues are involved. For the natural monopoly components of the industry, provided the regulatory framework is reasonably predictable, the utilities should have moderately low risk and be able to borrow at reasonably favorable terms. The issue will then be to determine the appropriate debt–equity structure on privatization to provide sufficient incentives to the management to keep costs low, and not to diversify outside the expertise of the management. In Britain many privatized utilities that were cash-rich and sold with low or negligible debt have found their post-privatization profits embarrassingly large. Some utilities have repurchased their shares and hence redistributed their income; others have paid high dividends; but some have diversified, often with disappointing results.

The second question that has attracted considerable interest is whether the industry should be sold in tranches, so that once the industry has established a track record for profits and regulatory uncertainties have been resolved, the government obtains some of the benefit of improvements in efficiency that were heavily discounted by the buyers at privatization.

Some of the most pressing and unresolved issues were mentioned earlier in this chapter, and further research as well as accumulating experience may go some way to reduce ignorance. Although the vexed question of reconciling efficiency in dispatch with cost-reflective tariffs and the problem of recovering capital costs in a competitive bulk electricity supply market have been subject to repeated review in Britain (OFFER 1994), they remain largely unresolved. As other countries face the same problem, different alternatives may be proposed and tried, and they then may be compared.

Several more fundamental questions remain: Is a bulk supply electricity market open to all a feasible long-run solution for managing third-party access? Or is the continental-preferred alternative for single-purchaser systems—in which the prices paid to generators are not necessarily closely related to the prices paid by buyers—more likely to be sustainable, and would it yield adequate benefits from the more limited form of competition?

The pressure created by radical reforms has prompted considerable interest in a host of technical questions: the desirability of nodal pricing for transmission systems; the problems of remunerating generating sets that are either prevented from supplying from grid capacity constraints or required to supply out of merit order to relieve capacity constraints; and the more general question of access pricing, connection charges, and providing suitable incentives for the management of the transmission system. Again, progress is likely to involve both theoretical investigation and comparisons between alternative solutions.

A range of issues concern environmental regulation and its interface with energy regulation, as well as problems of interaction between different regulatory regimes, such as that for gas and electricity in Britain. These bear on the whole design of competition policy within the country and its appropriate institutional representation. Is it better to have a hierarchical structure or, as in the United Kingdom, parallel institutions specialized by industry? Again, comparative institutional research that places these institutions within the wider legal and political system is likely to shed light on these design questions.

Finally, there are questions about the appropriate system for small countries where the installed capacity may be small compared to the minimum economic scale of plant. Some of these countries might benefit from regional integration, although grid constraints are likely to limit competition. The real question is whether the

option of openly contestable generation is worth retaining or single-purchaser bidding procedures should be used.

### Conclusions

Most industrial countries were able to solve the problem of financing capital-intensive electricity, either by providing sufficiently strong guarantees of fair rates of return to regulated private investors, by allowing self-regulation of cartelized, often-mixed public and private systems, or by state ownership with access to the budget. During the period of rapid demand growth, this was the main problem to solve, and the incentives to do so were strong.<sup>24</sup> As long as the system was able to meet demands at gradually falling real prices, there seemed little reason to make fundamental structural changes. Vertically integrated electricity supply industries appeared well placed to reap the benefits of coordinating investment and cross-subsidizing to maximize support for their continued existence. The number of events and circumstances—the fall in demand growth rates after 1975, the problems of nuclear power, tighter environmental regulations, the exhaustion of scale economies and technical advances in conventional generation, and the possibility of competitive entry by the new combined cycle gas turbine technology at small scale and lower capital cost—collectively conspired to cast doubts on the desirability of the old central planning solutions to managing the electricity supply industry. Introducing competition and market-based solutions began to look attractive.

The argument advanced here is that effective competition requires the deintegration of the grid from generation and the privatization of generation (and possibly also distribution) in order to create a market for bulk electricity. This in turn has far-reaching effects on the structure of relative prices, reducing the ability to cross-subsidize and putting competitive pressures on fuel supply industries, making subsidies harder to justify. Paradoxically, costs may fall as efficiency is increased, labor shed, and costly fuels such as coal and nuclear replaced by gas, whereas prices may rise as subsidies (to capital and fuel) are removed. If demand growth resumes and new investment in transmission and generation is required, prices will have to be adequate to reward private investors and may have to rise further in some countries, although the avoided cost to the public treasury may be the major benefit. The major challenge to the design of regulation for such a deintegrated industry will be to ensure that the bulk electricity market is adequately competitive without so encumbering it with regulation that

neither investor confidence nor market efficiency is undermined.

In developing countries the main problem is to improve the financial and economic performance of the electricity supply industry. To do so will require rebalancing tariffs, eliminating costly interruptions, and reducing construction and operating costs and construction delays. Allowing private investors entry into generation and possibly transmission and distribution is attractive on all scores, providing that entry both is competitive and takes place in a regulatory environment that reduces risks and hence costs sufficiently. While the lessons from industrial countries, especially Britain, appear relevant for and are beginning to be applied in developing countries, experience in developing countries is rapidly accumulating, reducing the uncertainties in designing reforms. The evidence in Chile in particular shows the importance of creating a sound and independent system of regulation, commercialization, and competition, even for state-owned utilities, and the relative unimportance of rushing into privatization.

In Eastern Europe the objective of privatizing utilities in order to reduce public debt is hampered by low tariffs and an unsatisfactory mechanism for setting and reviewing tariffs. If these problems could be remedied by the creation of a satisfactory regulatory regime, much of the financial urgency for privatization would evaporate (Newbery 1994a).

### Notes

1. The power sector would be able to finance all investment at an unchanged gearing ratio if the financial rate of return exceeded the rate of growth of capacity. The average annual rate of growth of power was about 7 percent for middle-income countries in 1960–90, compared with an average economic (but not financial) rate of return on World Bank projects of 11 percent (World Bank 1994, figure 3 and table 1.2). Had the financial rate of return increased to the economic rate of return, financing should not have been a problem.
2. British Telecom was the first utility to be privatized, but it had been preceded by various manufacturing, extraction, and service firms, and it was followed by many more, including British Airways, British Steel, and British Airports. See Vickers and Yarrow 1988 for a partial list.
3. A good example is Costa Rica, where in the 1980s the foreign debt of the electricity supply industry was 15 percent of total foreign debt (Covarrubias and Maia 1994, D2.1).
4. With inflation, the nominal interest rate on debt could easily exceed a sensible real rate of return. In that case, unless the nominal debt is increased, the real value of the debt will be rapidly

eroded. The assumption here is that the nominal debt can be increased in line with prices.

5. *Financial Times* (May 13, 1994) reported rumors that returns will be limited to 12 percent, considered unattractive because they are denominated in Yuan and because on January 1, 1994, the official exchange rate was abolished and is to be replaced with an electronic swap market. The article noted that the target rate of return may be part of a regulatory reform aimed primarily at the underpricing of power. To date, most foreign investment has required full government guarantees. "There is a great deal of difficulty in putting in place a structure that is sufficiently robust to achieve foreign project financing. . . ." Hence, foreign investment may fail to achieve the incentive effects of shifting equity risk to investors.

6. The English industry in this context includes Wales but not Scotland, which was privatized later and with a differently regulated vertically integrated structure.

7. Spiller (1993) observed that in 1962 the Jamaican government informed Jamaica Telephone Co. that it wished to renegotiate the terms of its license on expiration in 1966, reducing its durability substantially. JTC stopped all investment and eventually sold out to another company. In Bolivia the municipality of La Paz in 1984 started negotiations over the renewal of the license of the electricity company, due to expire in 1990. The company suspended all investment activity after 1984, and the license was still not satisfactorily renewed by 1991. Suspending investment is a completely rational act for a utility suddenly faced with uncertainty over its future regulatory regime.

8. The qualification, unfortunately important in many developing countries, requires that regulation be insulated by means of legal contracts and licenses from party political change in countries such as the United Kingdom with sovereign parliamentary freedom to pass new laws. Thus Spiller (1993) argued that Argentina failed to attract much foreign interest when it privatized its telephone companies precisely because of obscurities in the proposed regulation and tariff-setting rules. He also noted that some countries have had to accept low sales prices and high rates of return earned by the foreign private operators, who are anxious to recoup their investment before they are caught in a regulatory trap. That it might seem irrational for a government to put itself in such a position is no guarantee that governments in some countries will not do so.

9. This discussion draws on Newbery and Green forthcoming, which provides the supporting evidence for the abbreviated claims advanced here. See also Hannah (1979, 1982) for a fuller historical account.

10. See the review of the evidence in Vickers and Yarrow 1988, Chapter 2 and especially §2.5, 39–43, and the discussion in the section in this chapter on public vs. private ownership. Vickers and Yarrow cite evidence from the Edison Electric Institute (1985) that ownership has little effect on internal efficiency in the U.S. electricity industry, where both public and private firms are vertically

integrated, regulated monopolies. Studies on refuse collection suggest that private firms are more efficient than public firms, but this is explained by competition. When public and private firms compete via competitive tendering, cost differences between them vanish (Savas 1977). Caves and Christensen (1980) offered a similar observation in their comparison of public and private railroads in Canada.

11. Supply involves contracting for the delivery of electricity to the customer, metering, and billing.

12. This discussion is based on Müller and Stahl forthcoming.

13 This discussion is based on Hjalmarsson forthcoming.

14. Reserve margins in Sweden (capacity peak/capacity) were 58 percent in 1980 and 45 percent in 1990 (Gilbert and Kahn forthcoming).

15. This discussion is based on Kahn and Gilbert forthcoming.

16. The act's scope was subsequently broadened considerably by the Energy Act of 1992, which gave the Federal Energy Regulatory Commission the power to mandate transmission access.

17. This account, of course, grossly simplifies. Argentina initially developed via private, vertically integrated, concession-holding utilities, which, according to Covarrubias and Maia (1994), after 1940 were unable to finance the rapid growth in required capacity and were increasingly supplemented by federally owned utilities. It is not clear whether this failure of private finance was because of (a) restrained tariffs, (b) the step change in efficient scale required in an integrated and increasingly hydro system, (c) the absence of post-war international capital, or (d) the ready availability of World Bank finance. The state sector nonetheless continued to expand by acquiring private companies until no investor-owned utilities remained by 1982. Chile also had a private system until the 1930s, when politically constrained tariffs and the low growth of the Depression years broke the regulatory compact. In 1944 the government created the state-owned ENDESA to extend and integrate the system, and build large hydro plants, although the private Chilectra continued its thermal expansion program. Integration of transmission was associated with nationalization of most of generation and transmission, culminating in 1970 with the nationalization of Chilectra. Only 20 percent of distribution remained in private hands (Covarrubias and Maia 1994).

18. For example, in some cases it may be cheaper to keep peaking turbines available to meet power demands within a region to which transmission capacity is limited, rather than invest in additional transmission capacity.

19. Following Farrell (1957), technical efficiency is measured as the extent to which the utility reaches the technical production frontier, variously estimated; cost efficiency is the extent to which the utility minimizes costs at prevailing input prices. A utility can be technically efficient but not minimize its costs.

20. Price-cap regulation was first introduced for British Telecommunications following publication of a report by Professor

Stephen Littlechild (1983), the present director general of electricity supply. See Vickers and Yarrow 1988 for a history and discussion of the introduction of price-cap regulation for British Telecommunications and Beesley and Littlechild 1989 for the standard analysis of the desirability of price-cap regulation.

21. It may be cheaper to build another generator to serve a region with limited transmission capacity than to expand that capacity. But conversely, it may be cheaper to locate generation near the mine mouth or port and connect to distant customers. The fewer the transmission constraints, the less spare generating capacity is required.

22. A set is an individually controlled generating turbine and switch gear.

23. There are further payments if generating sets are required to operate out of merit order to meet transmission constraints and if declared capacity is tight relative to demand (capacity payments).

24. Compare the experience of Costa Rica with that of other countries, such as Argentina and Chile, which all faced the difficulty of financing rapid growth in capacity after the war. Costa Rica set up a state company to coordinate investment, which was undertaken by foreign capital, whereas the other countries publicly financed investment in public utilities. It is interesting to speculate whether the poor financial performance in many developing countries was effectively underwritten by international lending agencies, whose loans reduced the urgency of devising or evolving a system capable of delivering the power demanded.

## References

- Armstrong, Mark, Simon Cowan, and John Vickers. 1994. *Regulatory Reform—Economic Analysis and British Experience*. Cambridge, Mass.: MIT Press.
- Asian Development Bank. 1988. *Financing Public Sector Development Expenditure in Selected Countries: An Overview*. Manila: Asian Development Bank.
- . 1991. *Key Indicators of Developing Asian and Pacific Countries 22* (July). Manila: Asian Development Bank.
- Beesley, Michael E., and Steven C. Littlechild. 1989. "The Regulation of Privatized Monopolies in the United Kingdom." *Rand Journal of Economics* 20(3): 454–72.
- Berg, S. V., and J. Tschirhart. 1988. *Natural Monopoly Regulation*. Cambridge: Cambridge University Press.
- Besant-Jones, J. E., ed. 1993. "Reforming the Policies for Electric Power in Developing Countries." World Bank, Industry and Energy Department, Washington, D.C.
- Caves, D. W., and L. R. Christensen. 1980. "The Relative Efficiency of Public and Private Firms in a Competitive Environment: The Case of Canadian Railroads." *Journal of Political Economy* 88: 958–76.
- Covarrubias, A. J., and S. B. Maia. 1994. "Reforms and Private Participation in the Power Sector of Selected Latin American and Caribbean and Industrialized Countries." Vol. 2. Report 33. World Bank, Latin American and Caribbean Technical Department, Washington, D.C.
- Edison Electric Institute. 1985. *Analysis of the Differences among Alternative Forms of Utility Ownership in the U.S.A.* Washington, D.C.
- Farrell, Michael J. 1957. "The Measurement of Productive Efficiency." *Journal of the Royal Statistical Society* 120: 253–81.
- Farrer, T. H. 1902. *The State in Its Relation to Trade*. London: Macmillan.
- Foster, Christopher D. 1993. *Privatization, Public Ownership, and the Regulation of Natural Monopoly*. London: Basil Blackwell.
- Galal, Ahmad, Leroy Jones, Pankaj Tandon, and Ingo Vogelsang. 1994. *The Welfare Consequences of Selling Public Enterprises*. New York: Oxford University Press.
- Gilbert, Richard J., and J. Henly. 1991. "The Value of Rate Reform in a Competitive Electric Power Market." In Richard Gilbert, ed., *Regulatory Choices: A Perspective on Developments in Energy Policy*. Berkeley: University of California Press.
- Gilbert, Richard J., and Edward Kahn, eds. Forthcoming. *International Comparisons of Electricity Regulation*. New York: Cambridge University Press.
- Gilbert, Richard J., and David M. Newbery. 1988. "Regulation Games." CEPR Discussion Paper 267. Centre for Economic Policy Research, London.
- . 1994. "The Dynamic Efficiency of Regulatory Constitutions." *The Rand Journal* 25(4): 538–54.
- Gilbert, Richard J., Edward Kahn, and David Newbery. Forthcoming. "Introduction." In Richard J. Gilbert and Edward Kahn, eds., *International Comparisons of Electricity Regulation*. New York: Cambridge University Press.
- Green, R. J., and D. M. Newbery. 1992. "Competition in the British Electricity Spot Market." *Journal of Political Economy* 100(5): 929–53.
- Hannah, Lesley. 1979. *Electricity before Nationalization*. London: Macmillan.
- . 1982. *Engineers, Managers, and Politicians: The First Fifteen Years of Nationalized Electricity Supply in Britain*. London: Macmillan.
- Henney, Alex. 1992. *The Electricity Supply Industries of Eleven West European Countries*. London: Energy Economic Engineering, Ltd. and European Energy Economics.
- Hjalmarsson, L. Forthcoming. "The Market for Electricity in Scandinavia: Regulation and Performance." In Richard J. Gilbert and Edward Kahn, eds., *International Comparisons of Electricity Regulation*. New York: Cambridge University Press.
- Jones, Leroy, Pankaj Tandon, and Ingo Vogelsang. 1990. *Selling Public Enterprises: A Cost-Benefit Methodology*. Cambridge, Mass.: MIT Press.

- Kahn, Edward. 1991. "Risks in Independent Power Contracts: An Empirical Survey." *The Electricity Journal* 4(9): 30–45.
- Kahn, Edward, and Richard J. Gilbert. Forthcoming. "Competition and Industrial Change in United States Electric Power Regulation." In Richard J. Gilbert and Edward Kahn, eds., *International Comparisons of Electricity Regulation*. New York: Cambridge University Press.
- Kohli, K. N. 1987. "Financing Public Sector Development Expenditure: The Asian Experience." *Asian Development Review* 5(2).
- Laffont, Jean-Jacques, and Jean Tirole. 1993. *A Theory of Incentives in Procurement and Regulation*. Cambridge, Mass.: MIT Press.
- Littlechild, Stephen. 1983. *Regulation of British Telecommunications Profitability*. London: Her Majesty's Stationary Office.
- Moen, Jan. 1994. "Electricity Utility Regulation, Structure, and Competition. Experiences from the Norwegian Electric Supply Industry." Norwegian Water Resources and Energy Administration, Oslo.
- Müller, J., and K. Stahl. Forthcoming. "Regulation of the Market for Electricity in the Federal Republic of Germany." In Richard J. Gilbert and Edward Kahn, eds., *International Comparisons of Electricity Regulation*. New York: Cambridge University Press.
- Newbery, David M. 1992. "The Role of Public Enterprises in the National Economy." *Asian Economic Review* 10(2): 1–34.
- . 1994a. "Restructuring and Privatising Electric Utilities in Eastern Europe." *The Economics of Transition* 2(3): 291–316.
- . 1994b. "The Impact of Sulfur Limits on Fuel Demand and Electricity Prices in Britain." *Energy Journal* 15(3): 19–41.
- . Forthcoming. "Removing Coal Subsidies: Implications for European Electricity Markets." *Energy Policy*.
- Newbery, David M., and Richard Green. Forthcoming. "Regulation, Public Ownership, and Privatization of the English Electricity Industry." In Richard J. Gilbert and Edward Kahn, eds., *International Comparisons of Electricity Regulation*. New York: Cambridge University Press.
- OFFER (Office of Electricity Regulation). 1993. "Pool Price Statement." Birmingham, England, July.
- . 1994. "Trading Outside the Pool." Birmingham, England, July.
- Perez-Arriaga, Ignacio. 1994. *The Organization and Operation of the Electricity Supply Industry in Argentina*. London: Energy Economic Engineering Ltd.
- Pollitt, Michael G. 1993. "The Relative Performance of Publicly Owned and Privately Owned Electric Utilities: International Evidence." Ph.D. diss., Oxford University.
- . 1994. "Technical Efficiency in Electric Power Plants." Cambridge University, Faculty of Economics, Cambridge, England.
- Savas, E. S. 1977. "Policy Analysis for Local Government: Public Versus Private Refuse Collection." *Policy Analysis* 3: 49–74.
- Short, R. P. 1984. "The Role of Public Enterprises: An International Statistical Comparison." In R. Floyd, C. Gary, and R. Short, eds., *Public Enterprises in Mixed Economies: Some Macroeconomic Aspects*. Washington, D.C.: International Monetary Fund.
- Spiller, Pablo. 1993. "Institutions and Regulatory Commitment in Utilities' Privatization." *Industrial and Corporate Change* 2 (3): 317–80.
- Spiller, Pablo T., and L. V. Martorell. Forthcoming. "How Should It Be Done? Electricity Regulation in Argentina, Brazil, Uruguay and Chile." In Richard J. Gilbert and Edward Kahn, eds., *International Comparisons of Electricity Regulation*. New York: Cambridge University Press.
- Vickers, John, and George Yarrow. 1988. *Privatization: An Economic Analysis*. Cambridge, Mass.: MIT Press.
- World Bank. 1993. *The World Bank's Role in the Electric Power Sector: Policies for Effective Institutional, Regulatory, and Financial Reform*. A World Bank Policy Paper. Washington, D.C.
- . 1994. *World Development Report 1994*. New York: Oxford University Press.
- Yergin, Daniel. 1992. *The Prize: The Epic Quest for Oil, Money, and Power*. New York: Simon and Schuster.

# Regulating the power sector

Anthony Churchill

Regulation of the power sector is primarily a political issue. Although there are complex and difficult technical questions to be addressed, these are less important than the predominant need for societies to establish acceptable conflict resolution mechanisms. All societies develop formal and informal rules governing the exchange of property rights; without these rules, transactions costs can rise significantly. The institutions that societies establish both to make and to enforce the rules include parliaments and other legislative bodies, courts and law enforcement systems, and quasi-public boards.

If electric power were just another commodity, there would be no need for special rules or institutions to make and interpret these rules. The normal laws governing commercial transactions would apply. Electric power, however, is produced and distributed under conditions of monopoly. The existence of real or potential monopoly rents will inevitably bring forth competing claims for those rents.

While industrial countries have effectively managed the redistributive claims of society without excessively compromising overall efficiency, their experience provides little in the way of guidance. In fact, much of the focus of the literature is on techniques and sophisticated incentive systems, either market-based (price caps or rate of return minus adjustments for efficiency changes) or in the form of command and control (integrated resource planning). Because most developing countries lack a basic institutional infrastructure, few can use these techniques with any degree of confidence. Improving the regulatory framework will thus have to focus first on developing more effective political means of resolving conflict. Sophisticated pricing formulas are meaningless if there is no basic agreement on how gains and losses will be shared.

A simple framework can provide some useful insights into the regulatory process. Three types of institutional mechanisms describe the nature of social interactions (figure 15.1). The first, labeled politics, depends on voice

or deliberation; the second emphasizes law and is concerned with equity or representation; and the third, concerned with business or negotiations, is focused on market transactions.<sup>1</sup> Few institutions are found at the extremes. The interesting areas are those where two or more circles overlap.

The regulatory process and its institutions in most developing countries today fall in the right corner of figure 15.1. The incipient regulatory bodies are simply a department of government, run by civil servants or technicians, responsible to a minister. The major advantage of this type of structure is that it simplifies political control. The disadvantage is the predominance of short-run political imperatives that can result in excessive economic costs.

In the United States the institutional structure lies at the top of the triangle, where the judicial process dominates. There are numerous quasi-judicial bodies, each with one or more commissioners who are usually political appointees and who sit in judgment over the interested parties. The independence of the commissioners can vary considerably as can the openness of the discussion. The effectiveness of these commissioners depends on the existence of well-established judicial procedures and a tradition of public participation. Their costs, however, are not insignificant.

Developing countries must find an institutional structure that utilizes the existing strengths of the deliberative or cooperative traditions of many societies, yet recognizes the weakness of the judicial process and places greater reliance on more impersonal market forces. In the industrial economies there appears to be greater confidence in the use of market mechanisms in directing the public interest and an increasing willingness to treat the power sector within the normal commercial framework.<sup>2</sup> The developing world, perhaps justifiably, does not have this same sense of confidence in its commercial environment; although governments still seek to express the public interest, they are looking for a new social compact that

will be less destructive of efficiency and financial viability than those presently in force.

**Why regulate?**

Governments play a role in almost all matters of business and commercial activity, even if only to provide the framework for the establishment of property rights and enforcement of contracts. When governments go beyond these “rules of the game” and specify a special set of rules that apply uniquely to one activity, they are regulating the activity.<sup>3</sup>

The importance of the power industry to economic growth, the strategic nature of energy choices in defining the national well-being, and the potential redistribution of income that can be achieved by determining access to services are some of the reasons given for states exercising a greater than normal degree of control over this activity. In many cases, these aspects have been considered of sufficient importance to warrant direct control or ownership by the state. In many industrial countries they have justified a substantial level of state intervention. In France and the United Kingdom, for example, investments in nuclear power were viewed in terms of strategic or defense-related decisions.

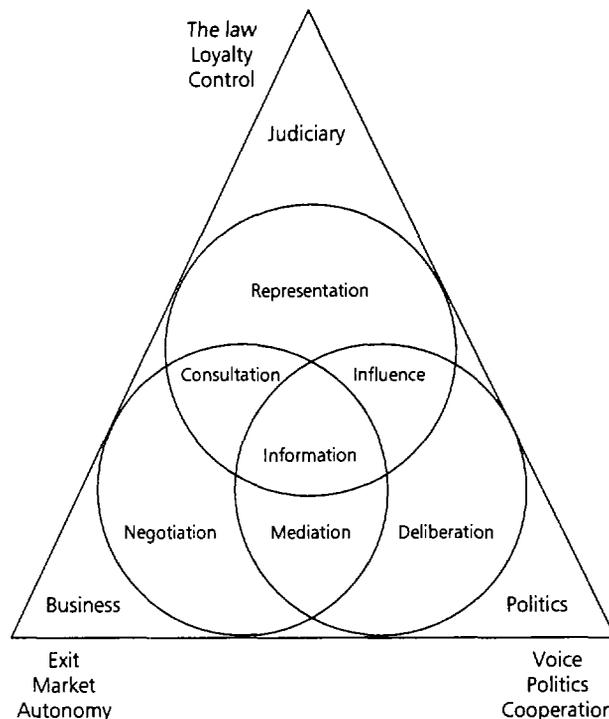
These considerations have invited broader political controls and thus subjected the industry to a negotiated process in which varying interests and groups compete for leverage. Commercial concerns for profitability and financial viability are only one of the competing elements in this process that have resulted in a social compact, or understanding, under which the industry is managed in each society.

In an increasing number of countries, the strategic or public good aspects of the power industry are being questioned. In New Zealand the power industry is regarded as just another commercial activity, subject only to the rules governing such activity. In Chile the government has withdrawn from most of the specialized regulation of the sector and settled for simple pricing rules to control monopoly profits. In the United States, the concept of obligation to serve in exchange for monopoly rights is being eroded by competition from independent power producers. The existing and increasingly unstable social compacts are being rewritten, the role of the industry in society redefined, and the boundaries between public and private interests redrawn.

*The regulator’s tasks*

The regulator must be concerned with resolving social conflict, improving accountability, and ensuring trans-

**Figure 15.1 Fundamental institutions and their interactions**



Source: Hirschman 1970.

parency, in addition to setting the rules of the game for ownership, investment, and operations.

*Resolving social conflict.* Most countries have tended to rely on specialized political processes for dealing with the governance of the energy sector in general and of utilities in particular. The best known of these is the U.S. system, in which separate regulatory bodies with quasi-judicial authority provide and enforce the rules of the game so that conflicts of interest are resolved at the least cost to the basic economic concerns of profitability and efficiency. Regulatory commissions at the federal and state levels, usually politically appointed, are empowered to hold public hearings, call witnesses, and take other steps to seek all relevant information. Various interests are encouraged to come forward with their views, specialized information services such as newsletters and bulletin boards reach the affected parties, and controversial issues are reported in the press. Most other industrial countries have similar mechanisms for focusing the debate and decisionmaking process.

The main features of these regulatory systems are their openness and inclusiveness. The staffs of these bodies provide the necessary technical support but are seldom in the

position to make actual decisions. Most of the issues are resolved at the political level, among interested parties, within a framework established by the broader body politic. Without this level of disaggregation of the political process, the growing complexities of governance in this sector can rapidly overwhelm the major political institutions.

In most developing countries regulatory institutions are still thought of as largely technical bodies deciding on technical matters (for example, the price of electricity) rather than as institutions designed to resolve social conflict. Public hearings are a novelty. Most governments are still unwilling to delegate real responsibility to specialized political institutions. The public at large has not been drawn into the process and often feels excluded. Developing these regulatory institutions is a difficult task that generally receives inadequate attention from officials in developing countries.

*Improving accountability.* In the past the state regulated or controlled almost all aspects of the power industry. Most state enterprises or public monopolies made up their own rules with little outside review and no external mechanisms of enforcement. The appalling environmental and health hazards that characterize the energy industries in Eastern Europe are extreme examples of the problems that can be created where accountability is nonexistent.

Experience has shown that it is almost impossible for one part of the government to enforce rules on another part. Developing-country governments will have to establish an arm's length relationship with the industries they regulate. The pervasiveness of the state in all aspects of business operations in the power sector, including ownership, has blurred the boundary line between the state's dual roles as an advocate of the industry and the guardian of the public interest. The weakness of many judicial systems only compounds the problem. The resolution of these inherent conflicts of interest will remain difficult if the state retains either a role in the ownership of the industry or responsibility for raising capital.

The enforcement of rules and regulations with respect to the environment, health, and safety cannot be left only to regulatory institutions. Enforcement also will require strengthening a wide range of public institutions. Although a regulatory body can set standards for effluent quality, for example, it seldom has the means for either collecting or testing samples. This responsibility is usually assigned to another branch of government, often at the local level. Neither does a regulatory body have an enforcement capability, which must rest with the courts and the law enforcement system.

Up to this point in the power industry, responsibility for the rules and their enforcement have been left to the public enterprises. Accountability inevitably has been weak. But the public will be unaccepting of weak accountability where private interests play a key role. As the private sector becomes more involved in the power industry, the state will face new, more rigorous standards of accountability and will have to strengthen a whole set of underdeveloped institutions, from the courts to municipal government.

*Ensuring transparency.* If private capital is to fund the sector, it will have to make most of the decisions. Private investors are unlikely to put resources into industries in which the state is guiding most of the decisions. They also are unlikely to put their resources into a sector that lacks stable and predictable rules. Regulation thus must focus on a clear, simple, and consistent set of rules.

Regulation has traditionally focused on prices or rates of return. A far more productive focus is entry and exit, in other words, competition. In the case of nontraded commodities and services, particularly electric power and to some extent gas, public regulation must protect consumers from abuse of monopoly power. But to minimize the need for this type of regulation, it should not be automatically assumed that many of these services can be delivered only under conditions of "natural" monopoly. The granting of legal monopolies should be regarded with suspicion. At a minimum, other parties should be permitted to contest any service provided under monopoly conditions.

Experience suggests that more competitive services can be developed as alternatives to those provided under monopoly conditions. Third-party access to transmission and distribution systems, for example, makes it possible for energy suppliers and consumers to reach normal commercial agreements on the quality and pricing of energy services, particularly for large commercial and industrial consumers, which in most cases are responsible for three-quarters of total demand.

In the provision of networked or common-carrier services, a constructive role for the state will be to facilitate the bargaining arrangements between interested parties. In electric power transmission or gas pipelines, for example, the state could facilitate closure on prices and other matters by inviting all parties to share in the ownership of the common facility. The role of the state should be to ensure that all relevant parties, including those acting in the general public interest, are at the bargaining table and that negotiations are carried out in a transparent manner. By facilitating an open and honest exchange, the state will

help ensure that the ultimate solution reflects a fair distribution of costs.

### **Who should regulate?**

Regulation is generally considered to be a public good and therefore the function of government. But how should the government exercise this function? In particular, who should regulate? Should it be a technical or a political group? Who should appoint the regulators? How independent should the regulatory bodies be, and from whom? What are the checks on abuse of regulatory power? Who should participate in the decisionmaking process? Can the regulatory authority make decisions or is it merely an advisory group? Should there be more than one regulatory authority? At what level of government? Should there be a special authority for each sector?

The answers to these questions will depend on the process of reform under way and, equally important, on the strength and capacity of market-based institutions, the independence and effectiveness of the judiciary, and the openness of the political system.

In developing countries there has been a tendency to appoint administrative boards that act in an advisory capacity to the minister. With few exceptions, these boards have focused narrowly on tariff and pricing issues; investment decisions and environmental issues are handled by other bodies. Yet even within their narrow mandate, they have not been notably successful. As long as the real decisions are made at a ministerial level, the temptation is to bypass the regulator. The makeup of these boards reflects their status, and they often become patronage jobs. Few boards have the technical capacity to critically review the information provided by the dominant monopoly.

Attempts to increase the autonomy of the regulators through either legislation or the "independence" of the appointees have failed because of gaps in the underlying institutional structure. In the United States, for example, where such regulatory bodies are common, there is a strong judiciary system for arbitrating disputes, a technical capacity for managing information, and a greater willingness on the part of the political system to delegate control. The openness of the system encourages debate and compromise.

Until the necessary institutional structure is in place, the regulatory systems of most developing countries will have to rely on a combination of more effective political or voice systems and use of market information. The issue is not that regulatory bodies lack independence, but that they lack the political representation required to be effective.

The trouble with the present, predominantly political approach to regulation in developing countries is not the deliberative process itself but the narrowness of participation. There is little public debate and no structured forum where interested parties can exchange information or engage in a process of negotiation and mediation. When a tariff increase is required, the general manager consults with the minister, who sees his job on the line. An impasse is quickly reached. There is no structure that allows the power company to explain or defend the need for a rate increase or helps consumers to understand the link between tariffs and the quality of services.

Opening up the process is essential. One way to do so is to recognize the regulatory board as a political body rather than confining its mandate to technical reviews, as is now the practice. By explicitly making the regulatory board a political body, a broader representation of interests can be invited. In other words, it is a move toward the center of the triangle (see figure 15.1), where deliberations are influenced by better representation. The objective is to provide a structure within which difficult and often contentious issues can receive an open airing—where users as well as producers have a voice.

A regulatory body that has no means of hearing or receiving complaints of consumers and other parties can create serious political difficulties. Yet the feedback mechanism must go beyond the complaints of consumers and provide the government and the general public with information on what is working and what is not. This means that the regulatory body must be in a position to gather information on costs, service quality, prices, how complaints have been handled, and other aspects of operations. There are several recent examples of privatizations in which the enterprises have had no legal obligation to provide the regulators with appropriate information. The lack of publicly available information can undermine the credibility of the reform process.

How much independence should be given to regulatory bodies? It would be unrealistic to assume that they could be completely isolated from the normal political process. Who is appointed and their tenure, authority, and scope are all political decisions. Ultimately, government leaders must answer to the body politic for all aspects of governance, including regulation. Different societies will draw different lines of responsibility around their institutions; over time, many of these boundaries will be redrawn.

The effectiveness of these institutions will be governed less by their independence than by the degree to which the process is open, accountabilities are clear, and ade-

quate information made available. It is possible, for example, for a regulatory board to be a purely advisory committee to the ministry and to have little or no independence. Its effectiveness will depend on how well it represents the interested parties, the openness of its decisionmaking process, and the reliability and comprehensiveness of the information it receives. The timely provision of adequate information will be a critical part of the process of negotiation and compromise behind any regulatory decisions. An unrepresentative body with secret processes and limited access to information will be ineffective. Each country will have to evolve a system that best fits its own institutions and political processes.

How can regulatory capture by the industry be avoided? The experience of New Zealand is instructive in this regard. The regulation of all utilities falls under the Commerce Commission, which is responsible for commercial activities. In other words, utilities are treated as any other industrial or commercial activity, which lessens the probability of capture. Given its broad mandate, the commission focuses on entry and exit, the competitive conditions and monopoly practices in specific markets, the enforcement of commercial codes and contracts, and other similar issues. In the United Kingdom, separate regulatory offices for each of the utilities focus on maintaining the competitive framework that has been established and on preventing the abuse of monopoly power. In the United States some states, for example, Ohio, have regulatory commissions with authority over multiple sectors.

This last option may not be particularly effective in most developing countries, however. The single authority covering all utilities requires a well-functioning commercial code and its supporting institutions. Few developing countries, for example, have policies or legal structures that effectively address anticompetitive practices.

In this regard, the role of the dominant supplier will be of particular concern. In almost all cases, the development of competitive alternatives will take time, and the new suppliers will be small relative to the existing monopoly. The national monopoly through its power purchase arrangements and day-to-day operations is usually in a position to influence the commercial success of any potential competitor. The monopoly is likely to have mixed views on the participation of independent producers in the generation of electric power. On the one hand, it will welcome the additions to capacity; on the other, it may well see the project as a competitive threat and use its controlling position either to undermine the project or collude with the private investors to share the gains from noncompetitive practices.

It will be essential for the government to establish, through the regulatory process, a credible, arm's length relationship with the dominant supplier. Widening participation in the deliberative process will help, but with its superior access to information and substantial resources, the present monopoly will inevitably be in a strong position.

Consider, for example, the issue of the price to be charged to alternative suppliers for use of or access to existing transmission and distribution systems. Because these enterprises are not usually operated on a commercial basis, this information rarely exists. When it is available, the costs are likely to be an arbitrary allocation of inaccurate historical costs. Most governments have relied exclusively on data generated by their own utilities and have no alternative sources of information. Thus, when the monopoly is asked to produce the required figures, it is in a position to set prices on the basis of cost estimates that can be easily manipulated for its own self-interest.

In practice this means if regulation is to be effective in encouraging new entrants into the business, all participants must be treated equally. This will only happen if steps are taken to place the existing monopoly on a commercial basis. Thus removal of subsidies and harmful noncommercial objectives imposed by government, usually followed by corporatization of existing state enterprises, is a critical first step in regulatory reform. In the short run, the most effective means of protecting against the dominance of a few special interests is to widen the base of participation in the deliberative process.

### **What should be regulated?**

Along with safety, standards, protection of consumers and similar issues of public concern, the usual areas of regulatory action are prices, investment decisions, and entry and exit.

#### *Prices*

Since there are no market signals to clearly guide the structure of prices in an industry in which costs vary considerably, the task of setting tariff categories falls on the regulator. Few issues are as difficult as the pricing of the services provided by natural monopolies. Given the complexity of relating prices to costs, regulatory systems have tended to favor gross simplifications of the rules for pricing. In most cases an "average" price for a variety of services and costs is used. Under rate-of-return systems of regulation such as that in the United States, prices are set to ensure that revenues are sufficient to cover costs, including capital costs. The distribution of these costs

over the customer base or load curve is usually accomplished by a set of rather arbitrary accounting rules and legal precedents. Only in recent years have regulatory systems permitted time-of-day-based pricing even though a major element of costs is related to the time of consumption. The World Bank and other international lenders have introduced the concept of long-run marginal costs as a complement to rate-of-return rules. When long-run marginal costs are used, prices are set so that, on average, they cover projected system expansion costs.

In some ways, focusing on one simple variable—either the rate of return or long-run marginal costs—simplifies the regulatory requirements. In others, it complicates it.

Under the pressure of political and other demands, the tariff structures of most regulated monopolies have evolved into a complicated pattern of cross-subsidies that usually bears little relationship to real costs. In both industrial and developing countries, these cross-subsidies have grown to the point where it is difficult to achieve even the “average” price objectives. Often, industrial or commercial users subsidize the more numerous residential consumers. Peak users are usually subsidized by off-peak users, rural users by urban users, and a few special industries or consumer groups by everyone else. In recent years, particularly in the United States, a new class of cross-subsidies has been introduced to achieve environmental and other social objectives. The end result is that most systems wind up with tariff structures that have little to do with costs and with few incentives to minimize costs.

The information systems are similarly distorted. An integrated monopoly has little incentive to collect information on the various parts of the system, particularly if pricing decisions do not take into account the costs of the different parts of the service. If there are no time-of-day charges, for example, there is no incentive to collect usage information by time of day. If plants are not dispatched on the basis of short-run marginal costs, this information will not be available. In the United States, for example, the information collected has more to do with accounting rules, regulatory requirements, and the tax structure than with the economic costs of different services.

These types of pricing systems put an enormous burden on regulation, and the overall objective of achieving both efficient production and efficient consumption of the services provided is lost. Attempts have been made to design pricing rules that focus on efficiency rather than the redistribution of benefits. One of the most promising of these is one in which the regulator starts with existing prices and returns. In exchange for the freedom to adjust prices within tariff categories, the monopoly is expected

to lower average prices by some annual percentage. This pricing system has been particularly popular in the United Kingdom and the United States in telecommunications, a sector in which the rate of technical progress has produced dramatic declines in cost over time. The principle is simple: The monopoly is encouraged to improve efficiency by being allowed to keep, for some fixed period of time, the gains that arise from these improvements.

This pricing system is not without its problems, however. The determination of the annual adjustment factor and the length of time for which it will apply is not a simple matter in practice. The application of this pricing system in the power sector has been discussed but has not been implemented in any significant system. The pricing formulas used in Chile come close. Efficiency is promoted by basing prices on “best practices.” If a firm can exceed the best practices for a period of time, it is permitted to keep the gains. The reason these cost-plus-or-minus systems have been slow to take hold in the power sector is probably that the likely direction of changes in costs is less obvious than in the telecommunications sector. Then too, the existing pattern of cross-subsidies is so well entrenched that any attempts to introduce a new pricing system are likely to generate substantial opposition.

#### *Investment decisions*

Providing an adequate framework for investment decisions is probably the greatest challenge for regulators. Large, complex, and often “lumpy” investments invite the application of administrative discretion over simple rules.

External lenders have made popular the use of least-cost planning techniques from which to select investment alternatives. But because such techniques call for a great many discretionary assumptions, there is the risk that investment decisions will be based on social and political objectives rather than purely economic considerations. In addition these techniques were designed for public sector investments, and their implementation generally requires the existence of a monopoly.

The private sector works with a different set of assumptions, particularly about risks, and is more concerned about minimizing risk than costs. Private investors would be unwilling, for example, to undertake an investment with a long construction period or substantial construction risks, a high probability with hydro projects (see Churchill 1994). Some countries have attempted to address these differences by having the public sector make the investment decisions and then inviting the private sector to compete for the project. But because of the

risks involved, private financing has not come forward with much enthusiasm in the absence of substantial public guarantees.

Although least-cost planning techniques may be useful as an indicative planning tool, they are incompatible with competitive markets. In the United States these techniques have been used by private regulated monopolies in well-established market structures. In developing countries, however, it is unlikely that much private capital will be forthcoming unless the investors, rather than the government or its utility, make the fundamental decisions on what to build. If the private sector does not make the decisions, it will not assume the risks.

One of the arguments against allowing the private sector to make these fundamental decisions is that it will be "excessively" concerned with short-term considerations. In actual practice the results are somewhat mixed. In the United States a number of private utilities made long-run decisions regarding their investments in nuclear power. Although many of these decisions turned out to be costly, the implicit understanding between regulators and investors that all costs would be passed on to consumers permitted the investors to make "public" decisions. In developing countries, by contrast, the experience with least-cost planning has prevented neither excessively costly decisions nor the undertaking of high-risk projects (see World Bank 1990). In other words, there is little evidence to support the view that decisions made by the public sector will necessarily be "right."

#### *Entry and exit*

In the past decade some important developments have supported the view that competition is the most effective and efficient framework for the regulatory process. The theory of contestable markets, in particular, has had a major impact on the intellectual foundations on how governments should regulate (see Baumol and others 1988). The regulatory reform process in New Zealand offers a dramatic example of the power of these ideas. The reform process in that country has focused on entry and exit conditions, or contestability. The regulator's job is to ensure fair entry into any aspect of a regulated business. Investors willing to put their own resources into a business have a legal right to do so, provided they bear the investment risk. Most regulation today, however, prohibits any form of competition. In Costa Rica, for example, even modest levels of autogeneration were prohibited until recently. For the most part, governments have reserved electric power for a state monopoly and have actively discouraged competition.

Finding a regulatory path between protecting monopoly privileges and introducing competition has proved to be extraordinarily difficult. Once any degree of competition is introduced, the system is pushed in the direction of either returning to the former restrictive practices or allowing further competition. In the United States, for example, once independent power producers were permitted entry, pressures for third-party access were quick to grow; today the system is being pushed in the direction of a complete unbundling of services in a competitive market framework.

A few developing countries have allowed limited entry into power generation; the next step is to introduce competition by allowing generators access to customers. In the most limited form, the generator is permitted to sell part of the output from his cogeneration plant to third parties rather than being limited to the dominant monopoly. Demands for third-party access to transmission and distribution systems immediately result.

One way of handling the conflicts between control and competition, and providing a more explicit transition process, is the approach taken by Portugal. Some 75 percent of the estimated future demand has been reserved for planned public additions to capacity, usually through a competitive bidding process. Decisions on the remaining 25 percent are to be left to the marketplace.

The Portuguese experiment will be an interesting development to watch. What if demand growth differs from the planned estimates? Where will the additions or cutbacks occur? The inevitably slow and cumbersome nature of the public procurement process suggests that decisions once made will be hard to change and that the market-led sector will be forced to adjust. But because private investment decisions are characterized by greater speed and flexibility, this capacity may well be installed first, and it may be the less responsive, publicly induced capacity that will have to adjust. Regardless of how the adjustment process plays out, a healthy tension between market and planned decisions will be created. With careful management, this approach may result in a satisfactory transition process.

#### **Institutional options**

One of the major difficulties in considering a new regulatory structure is that the sector itself is undergoing substantial change. The development of a regulatory framework must be part of the process of structural reform of the sector. This section examines the regulatory implications of the structural reforms in the power sector that are under way in many countries.

*Integrated public monopoly*

The most common institutional structure for the power sector in almost all developing countries is the integrated public monopoly. In larger countries (for example, Brazil and India) there may be several regional monopolies together with some administrative divisions, particularly at the wholesale level, of generation, transmission, and distribution. Public ownership and, more important, public responsibility for raising the necessary capital are the key features of this arrangement.

Many countries are contemplating transforming these public service institutions to commercially oriented enterprises, which would subject them to normal corporate pressures for profitability and financial viability. Malaysia and the Republic of Korea are going one step further by offering shares to the general public. The changing value of the shares will reflect the performance of the enterprise.

Commercialization and corporatization require a well-defined regulatory framework. If corporatization is to achieve better results than the previous direct controls, rules on how prices and returns are to be calculated must be established and enforced. Australia and New Zealand took advantage of the shift in enterprise structure to make significant changes in management and labor relations. But such changes are likely to be unsustainable one-time gains unless the issues of pricing and competition are addressed within a reasonably short period of time. Some of the francophone countries have attempted to develop more efficient forms of governance between the utilities and government through specific agreements or "contract plans." However, these efforts have not been successful because it has proved difficult to limit the interference of the political system.

Whether the utility operates as a public enterprise or as a public corporation, under the integrated public monopoly arrangement entry into any part of the sector is restricted, and pricing and investment decisions are made by government. External lenders have typically attempted to instill some discipline into the system using targeted variables such as the rate of return on assets or the requirement that prices, on average, must cover long-run marginal costs. Regulatory boards or price commissions sometimes are established to assist in developing tariff structures. Market signals are weak and have little or no influence on pricing.

Even where there are "independent" boards, these have little power and usually function in an advisory role. Since the minister (or the cabinet) makes the ultimate decision, it is the political process that determines the

results. There is little public debate. Few countries have a tradition of public hearings, and there is no structured forum where interested parties can exchange information or engage in a process of negotiation and mediation.

Yet a more open decisionmaking process is a necessary but far from sufficient condition for an effective regulatory regime. In developing appropriate regulatory structures, the system must be made more sensitive to market signals. Even in the United States, where discussion and representation are structured and reasonably open, the system has an ultimate check in the market. Public commissions and interest groups cannot make or change the rules without keeping a careful eye on capital markets. In recent years, for example, regulatory bodies have introduced the concept of "due prudence" in determining which investments can be included in the rate base. As a result, the risk to investors and the cost of capital have increased. All parties—investors, utilities, and consumers—have had to modify their behavior in response to market signals.

*Privately financed, regulated monopoly*

The present system in most industrial countries is the privately financed, regulated monopoly. Whether ownership is public or private, the key feature is the use of private capital markets to raise the necessary finance. To satisfy the requirements of the capital market, a rate of return on capital is targeted. The market acts favorably or unfavorably, depending on whether this target is achieved. In the United States where private ownership is permitted, greater accountability exists for investment decisions. Where private ownership is absent, politically motivated and usually expensive investment decisions are possible (as in the case of nuclear power in Canada and the United Kingdom and lignite plants in Australia), and the government must ultimately bail out the system. Although the capital market provides some information on the efficiency of investment decisions, explicit and implicit government debt guarantees generally mute this feedback.

In order to attract private capital, a clear set of rules and a means of enforcing the rules on both the government and the enterprise is required. The relationship between the government (politics) and the enterprise must be predictable and subject to an open and well-understood judicial process. An independent judiciary is thus an essential requirement. For the most part, the signals from the capital market are a commentary on the ability of the regulatory system to balance the competing demands of the political system.

Although the use of market-determined targets injects an important element of discipline into the overall operation of the system, the determination of the actual pricing structure is subject to a substantial element of political control. In the United States complex legal and accounting structures have been developed to support the tariff-setting process, freeing regulatory bodies to focus on other objectives. In recent years the regulatory system has been used to reinforce environmental goals, using the ability to discriminate among consumer classes.

A number of developing countries are considering moving closer to the U.S. model by opening the enterprise's capital structure to private shareholders and organizing the enterprise so as to improve its access to capital markets. In Argentina and Malaysia corporatization and privatization took place at the same time. Privatization adds another dimension to the structural reform process and increases the urgency of establishing an appropriate regulatory frame. In the case of Argentina, the uncertainty regarding the rules under which the newly privatized entities (essentially regional monopolies) would operate, undoubtedly affected both the number of bidders and the price paid.

Failure to provide an appropriate regulatory framework creates risks not only for the private investor but for the government as well. In the absence of a clear set of rules, governments may be tempted to negotiate ad hoc deals either with investors or in the distribution of shares. Investors will naturally try to hedge all risks by seeking the collusion or active cooperation of those responsible for the sector. Some of the deals negotiated in the last few years involving build-own-operate-transfer (BOOT) or build-own-operate (BOO) plants have unduly favored investors. In other situations, particularly in the telecommunications field, financially strapped governments (specifically, the governments of Jamaica and Mexico) have received attractive prices for these assets by guaranteeing exclusive rights. It is not always clear that such deals are in the country's longer-term interests.

As a practical matter, however, it will not always be possible to specify the rules with much certainty. Few countries have much experience with the type of regulations that will be required in a privatized power environment or, more important, with their enforcement. Even with the best set of rules, investors are likely to be cautious until some experience is gained and to require both substantial guarantees and high rates of return to compensate for potential adverse actions by the government or its agents.

This lack of experience led Australia and New Zealand to postpone privatization for some time following the

commercialization and corporatization of their power sectors. Their rationale was that the government and the enterprise should take the time to work out the rules and acquire experience with their enforcement before privatization. There are some risks in this strategy, however. Once the enterprise is corporatized—and particularly if significant improvements in management and labor relations are made in the process—the pressure to write the rules and follow through with the privatization may be relieved, as seems to be the situation in New Zealand.

It is not clear, however, that the results from postponing privatization will be much different than other similar attempts to rewrite the rules of the game between governments and state-owned utilities (such as the contract plan). It is difficult to maintain an arm's length relationship between what are essentially different parts of the same government. The involvement of private parties with a financial stake offers a clearer separation of responsibilities and greater opportunities for the enforcement of contracts. In the end, ownership counts.

#### *Monopoly with competitive procurement of generation*

A number of countries have introduced competition on the generating side of the business, whether public or private, into what has been a vertically integrated electric power monopoly. India, Malaysia, and the Philippines are struggling to write the rules governing the competitive procurement of generation.

Competition has been so limited, however, that what has been left is another form of public procurement, although a more efficient one. The private firm has become a contractor to the existing monopoly for a set of specialized services that now includes finance. Although in some cases there may be competition among private firms for the provision of these services, more often than not it is a "negotiated" deal. The public utility or the government decides what plants are to be built, what technologies are to be used, where plants are to be located, what is to be produced, and at what price. The Hub River project in Pakistan and the proposed private sector plants in Jamaica are of this type. Since the public sector makes most of the decisions, it winds up having to cover most of the risks. The private party is simply selling management, technical, and, increasingly, financial services to the government or its agents.

In these circumstances, the main regulatory concern is over the nature of the contracting process between the private producers and the utility, which has become a monopsonist in the market for capacity and energy. To protect his investment, the private investor will focus on

obtaining a satisfactory power purchase contract from the utility and will look to the government to underwrite the risks with respect to its own behavior or that of the utility.

In the United States, with a predictable regulatory framework and a strong judicial system for contract enforcement, it has been possible for independent producers to work out satisfactory power purchase contracts. Moreover, because a greater degree of competition is being introduced, the rule-making process is greatly simplified. In Virginia, for example, the regional monopoly, VEPCO, has requested construction bids for power plants mainly on the basis of the price it is prepared to pay for power. VEPCO has indicated the amount of power it needs and approximately when and where, stated the price it is prepared to pay, and then asked for bids. The developer then takes the risk on the number of plants to build and their location. Given the final price (and certain technical qualifications), the profitability of the enterprise will depend on the developer's decisions. The utility and the public sector need not be involved in those decisions. In the case of VEPCO and other utilities that have developed this process for additions of future capacity, bidders have come forward with proposals that more than satisfy the capacity requirements.

Thus by fixing a price or a set of rules on how prices will be established, the monopoly avoids the difficulties being experienced by countries as diverse as Honduras and India as they try to contract for new plant. The limitation on this approach in developing countries, however, is the lack of credible rules or operating experience with pricing regimes. Private suppliers will be unwilling to enter into such contracts with the dominant public monopoly unless significant government guarantees are forthcoming. In the Philippines the procurement of private generation capacity has been possible only with the government assuming all risks with respect to prices and quantities. The independent producer receives a physical quantity of fuel from the dominant utility and then converts it to kilowatt hours for a processing fee, taking no risks with respect to either input or output prices.

Chile is the only developing country in which investors have been willing to undertake capacity expansions on the basis of relatively firm expectations about prices and market structure. Perhaps investor confidence can be increased over time to the point where contracts based on the price of power alone will be sufficient to induce investments in new capacity.

Expanding generation capacity through public procurement may increase efficiency in plant construction and operation. Yet it may be impossible to develop capac-

ity on the required scale, given the institutional and regulatory limitations that push governments to take all of the risks. The various build-operate-transfer (BOT) schemes have all required a long and complex process of negotiation, and many, particularly the larger ones such as that in Turkey, have failed to get off the ground.

#### *Unbundling services*

The limitations for efficiency improvements of simply moving to more competitive procurement of generation are obvious if they take place within the traditional monopoly structure. The alternative focuses on unbundling the services offered by the vertically integrated monopoly, splitting off those that can be provided under competitive circumstances from those in which natural monopoly elements prevail.

*Generation.* Must generation be provided under conditions of monopoly? A few decades ago the answer would have been yes. Small systems were dominated by one or two large, central power plants characterized by significant economies of scale. As systems have grown and generation technologies changed, economies of scale have become less relevant.

Perhaps an even more important reason the competitive provision of generation services is now possible is the existence of communication and information systems that permit the dispatching of electric power within a more market-oriented framework. In small systems usually operating within a well-defined region, it was necessary to maintain close control over generation and its dispatch, ensuring, on the basis of technical criteria, adequate reserve requirements and service quality. As systems have grown and become interconnected, new protocols have been established, and a number of power pools have developed in both North America and Europe to trade power and reserve requirements among large regional monopolies. These power pools reflect various degrees of system integration, from exchanging power at the margin to more centralized control of the combined systems.

Since members of the pools had different sets of owners—different governments in the case of Europe and different combinations of private and public owners in the case of North America—governments or the pool members have had to develop rules to govern their interaction. Initially, most of these arrangements were fairly simple and reflected only marginal transactions. But as greater integration has been achieved, and in particular (where independent generators have become part of the system), rules have of necessity become more complex. This com-

plexity inevitably forced the rule systems to simulate what might happen in a competitive wholesale market.

In the United Kingdom the power sector has developed into a wholesale market for generation services, with generators actually bidding at half-hour intervals for a place on the load curve, for the provision of spinning reserves, and for other technical services. The order and amount dispatched are determined by the prices offered rather than, for example, the traditional engineering-determined merit order used in most systems. In other words market prices have substituted for technical parameters. The profitability of each plant on the system is thus dependent on the ability of the plant owner to compete with respect to prices and costs. With market-determined prices, other market-based mechanisms have developed to arbitrage risk, namely a small but growing futures market for kilowatt hours.

The existence of market-determined wholesale prices has greatly simplified the regulator's job. The main task now is not to determine prices (or the structure of tariffs) but to ensure entry into and exit from the system—that is, that the existing group of generators dominated by two large companies is unable to exclude new entrants or grant favorable treatment to existing producers. The regulator's objective is thus to encourage competition and prevent the abuse of monopoly power.

One of the major advantages of the present U.K. system is the small amount of resources required for its regulation. The Office of the Regulator is staffed by only a few people. In contrast, the U.S. system requires a virtual army of lawyers and other technical experts. Regulatory costs are kept to a minimum in New Zealand because electric power has no separate regulatory body and is regulated under the Commerce Commission, which focuses mainly on entry and exit conditions.

An important feature of the U.K. system and one used on a limited basis in the United States is to permit the supplier direct access to the customer. In the United States independent power producers are building plants on the basis of heat and power contracts signed with one or two consumers, usually bypassing the regional utilities transmission system. In the state of New York the utilities are requesting permission of the regulators to directly negotiate contracts with their major customers, since competition from unregulated suppliers has placed these utilities, with their numerous cross-subsidies and pricing rules, at a disadvantage.

*Transmission.* Transmission services appear to have the strongest element of natural monopoly. In most cases transmission either is part of a vertically integrated monopoly or

is run as a separate and regulated public monopoly, as in India. In the United Kingdom the transmission system is owned by the private regional distribution companies and is subject to public regulation regarding who may sell and buy from the system. In the United States legislation enacted in late 1992 started the process of opening up the transmission systems owned by the regional monopolies to permit third-party access.<sup>4</sup> Sweden and a few other countries have permitted third-party access, but France and Germany, in the interest of protecting their state utilities from competition, have rejected it and stalled its introduction in the European Union.

Third-party access permits generators to reach consumers directly without the mediation of the regional utility. The large, 2,000-megawatt Tysdale plant built by Enron in the United Kingdom was financed on the basis of purchase contracts negotiated between the company and power consumers. The contracts would not have been signed if producers and consumers had not been confident that adequate transportation arrangements existed. Because Enron had been granted third-party access rights, transmission was ensured.

Permitting some degree of competition through third-party access to high-voltage transmission lines for larger customers—which in most developing countries could account for more than 75 percent of the load—could be an attractive way to compel greater efficiency from the system by creating pressures to price energy to reflect costs. If given a choice, customers will shop around for prices and qualities of services that best meet their needs. Customers in many developing countries now have few choices: either poor-quality service at subsidized prices from the public monopoly or expensive autogeneration. To obtain higher-quality service customers in Indonesia and Nigeria pay a multiple of the public price through autogeneration.

Competition also will limit the ability of suppliers to cross-subsidize one class of consumers at the expense of others. If utilities are able to negotiate directly with major customers, subsidies to certain classes of residential customers, for example, will no longer be possible. Regulators thus will be confronted with the true costs of providing various subsidized services. Changing the distributional impact of power tariffs will of course necessitate some difficult political decisions.

Transmission services can be further unbundled by separating out the system management. In an integrated monopoly decisions regarding which plants are on line and which in reserve status, which plants provide spinning reserves, and which links are used to transmit power, are all centralized functions. Whether third-party access is

permitted or the transmission lines are privately owned, there will still be the need for centralized system management. This need is usually advanced as an argument against third-party access and for continuation of an integrated monopoly. Recent experience suggests, however, that it is possible to separate out the system management functions without affecting performance. In the case of the U.S. power pools, for example, many of the system management functions have been delegated to jointly owned and managed control centers. In Europe a number of the national utilities have conceded considerable authority to centralized management centers. In the United Kingdom the National Grid Company is both the owner of the transmission network and the manager of the system. In Australia consideration is being given to establishing a national grid or system management function separate from the ownership of the transmission system.

The system management functions are clearly a natural monopoly and present a number of issues for public regulation in addition to those discussed earlier with regard to ownership and access to the transmission system. System managers are in a position to determine the operations and profitability of all parts of the system. Unless the rules are clearly specified and understood by all parties, there is the possibility of considerable discord and political fallout, particularly where the legal structure for contract dispute resolution is underdeveloped.

One alternative is to manage the system as a public or state enterprise, as electric power is managed in New Zealand.<sup>5</sup> However, state ownership and control of system management alongside private ownership of generation, transmission, and distribution facilities will result in difficulties that cast doubt on the viability of the approach, particularly in developing countries. Public management will require an extraordinarily strong judicial system if it is not quickly to become another instrument of political control. The signals of the marketplace inevitably will be weak.

Efficient system management requires the ability to make almost instantaneous decisions. Failure of a generating plant, for example, requires an immediate search for alternatives. Many but not all of these decisions can be anticipated and automated. Managers will require a large amount of discretionary authority, an area in which the public sector usually does not perform well.

Apart from the day-to-day and hour-to-hour decisions is the issue of writing the system rules. The profitability of a power plant will depend on the number of hours and times of the day it is in operation. If, for example, there is an outmoded coal-fired plant in one section of the country that is expensive compared with alternatives available to the sys-

tem, system operators will be tempted to limit generation from this plant. But what if the plant is in the prime minister's district? Or what if it means that a large number of coal miners (voters) will be laid off? These examples show how easy it is for noneconomic criteria to dominate system decisions. Even in the case of the United Kingdom, where most plants are dispatched on the basis of prices bid, a political decision has been made always to dispatch the nuclear plant.

The role of the public sector should be limited to controlling the abuse of monopoly power and facilitating bargaining among the parties involved. The joint ownership of the system management by generators and distributors provides the mechanism for the various parties to work out the rules of the game. It would be difficult for one group, for example, the generators (or an individual generator), to extract monopoly profits from system operations without the knowledge of all concerned. The tendency is for the parties to arrive at a set of economic prices and costs that reflects a fair distribution of the burden. The role of government (or public regulation) is to ensure the openness of the process and that all parties with a legitimate right to join the "club" are allowed to do so.

The North American power pools are a limited version of this type of joint-ownership arrangement. Members exchange information on costs and formulate how profits are shared. In the United Kingdom the system management function is under the ownership of the Regional Electricity Corporations, which are the regional distributors and presumably have a strong interest in maintaining efficient, equitable operation.

*Distribution.* Distribution systems remain one of the more difficult areas for reform. In New Zealand contestability was introduced at the generation stage, and transmission was structured as a national monopoly; the municipal distribution systems were left untouched. The gains in efficiency in the form of lower wholesale prices were not passed on to the bulk of consumers but instead absorbed by the regional distribution monopolies. Recent legislation prevents the regional monopolies from absorbing future efficiency gains. In addition the regional monopolies no longer have exclusive rights to supply their territory.

In the United Kingdom customers with a maximum demand of 1 megawatt and greater were able to choose their suppliers from April 1992 to March 1994. In 1993-94, 37 percent of the sites in this market, accounting for 57 percent of the demand, chose to take second-tier supply, that is, from a supplier other than the local distribution company. In April 1994, when the franchise limit was lowered to 0.1 megawatts, customers followed a similar pattern of choosing other suppliers.

Australia and a few other countries are considering similar arrangements, permitting competition for at least larger customers. In countries with reasonably well-functioning billing and payment systems, there is no reason that competition cannot be extended to the level of the residential household. The United Kingdom has in fact publicly announced its intention to do so.

These changes in the distribution system constrain the degree of monopoly of the service. Customers deal directly with the supplier, and the wire service, or common carrier, transports the energy at established rates, thereby limiting the public regulatory burden to ensuring access and setting access prices to the distribution and transmission networks. Long-distance telephone services have reached this point in a number of industrial countries.

Although this type of competition in distribution limits the need for regulatory intervention to an area covering less than 20 percent of total costs, its implementation for any but larger customers would be difficult in most developing countries. Metering systems, and cost-accounting and billing practices, are likely to be inadequate to meet the information needs of such a system.

One alternative that several countries have considered is benchmark competition for regional distribution monopolies. In this case the function of the regulator is to improve accountability by comparing the performance of the distributors. Since most are likely to face the same set of wholesale prices, the range of profits and prices should reflect their relative efficiency. If the regulatory authorities have control over consumer prices, tariffs can be set to reflect best practices and owners of the distribution monopoly forced to absorb the inefficiencies. Widespread knowledge of the relative performances of the various distributors will apply public pressure on poor performers.

The province of Ontario, in Canada, adopts such a system. Electricity is wholesaled by Ontario Hydro, which has a virtual monopoly on generation and transmission, but power is distributed by municipal electric associations. Ontario Hydro "suggests" a set of retail rates to the associations that presumably reflects a sufficient margin to cover distribution costs. These rates are widely publicized, and it is difficult for any association to put in place a different tariff structure. The profitability of the association, or the deficit to be paid by the municipality, provides an incentive for efficiency.

Benchmark competition is not without its problems. Key parameters necessary to make it work are the openness of the information system and the willingness of government or the regulators to accept the profits and losses of the distributors. Moreover, the issue of what is a rea-

sonable set of rates at both the wholesale and the retail levels is still a matter of considerable public dispute. In the case of Ontario, the utility and the municipalities frequently disagree about the structure of tariffs. The municipal electric associations, which must deal with the final customer, have recently complained that the wholesale rates are excessive and that they are bearing the burden for inefficient investment decisions at the wholesale level.

### **Directions for future policy research**

There is a significant gap between rules and enforcement capacity, with a temptation to keep the rules and try to strengthen their administration. Although this may be a useful long-term goal, in the medium term it may generate a counterproductive disrespect for the rules. Enforcement capacity is highly dependent on overall institutional progress (honest government, for example) and is unlikely to change much in the short run. Research is needed to identify ways to shore up weak administrative and institutional capacities. Use of private firms, possibly of international origin, to monitor compliance with rules, for example, is an alternative for overcoming weaknesses in public administration. In the case of weakness in judicial bodies, can more effective use be made of the press to shame firms into compliance or bring public pressure to bear? Can competition be structured so as to make compliance with rules a competitive advantage?

Most discussions of reform in the electric power sector focus on introducing competition in generation. If this part of the sector were competitive, one-third of the cost of producing electric power would still be produced under conditions of natural monopoly. How to regulate the wire services, transmission and distribution, and most important of all, the dispatch or grid management functions around a competitive generation sector, is a problem that is only beginning to receive attention.

If competition in generation is to be effective, it is critical to have clear rules of the game on who will be dispatched, in what order, and when; unclear rules can result in manipulation of the system. There is little experience to draw on. In the United Kingdom, for example, the regulator has intervened in situations in which the physical location of plants on the transmission network, for technical reasons, has permitted the generators to extract monopoly profits. In the United States the opening up of the transmission system under the Energy Policy Act of 1992 has introduced a whole new set of regulatory problems for the Federal Energy Regulatory Commission. Resolving these problems promises to be a long and difficult process.

Developing countries will have the opportunity to regulate these services on the basis of more rational criteria than are likely to be applied in the United States. Argentina and Chile, two of the countries that have started the process, are encountering implementation problems that are revealing some of the weakness of the initial framework. The development of a better theoretical framework and a deeper understanding of the practical problems of implementation should be high on the research agenda.

### Conclusions

The reform of the power sector involving a shift from command and control to public ownership of integrated monopolies, to greater reliance on market signals and, finally, to private ownership, is on the agenda in most countries. These changes require governments to develop regulatory mechanisms that incorporate the public interest into the sector's decisionmaking process. With few exceptions developing-country governments have had little experience in the effective use of public regulation in a market-driven setting. The models from industrial countries are of limited use. The public monopoly has been the dominant form of organization; where it is not (as in the United States), the institutional and legal structures are so well developed that they do not serve as a useful paradigm to be emulated.

Developing countries thus have no choice but to experiment with home-grown solutions. Inevitably, the regulatory framework will lag behind sector reform. Care must therefore be taken to avoid overly rigid initial solutions. Flexibility will be the key to successful regulatory reform.

Above all, recognition must be given to the essentially political nature of the process. Regulation is a means of taking into account society's concerns. In the power sector redistributive concerns have been at the forefront, often at the expense of efficiency. Redressing this imbalance will produce both gains and losses among the different constituencies. New, more efficient institutional mechanisms will be needed to resolve the inevitable conflicts. Appeals to "independent" or "nonpolitical" regulatory institutions will not work.

The task of developing the appropriate conflict resolution mechanisms or regulatory institutions will be easier if openness, accountability, and transparency are stressed. In particular an emphasis on more inclusive or participatory mechanisms is desirable. At the same time greater reliance on competition and market signals will make the regulatory task easier.

### Notes

1. The use of triads of this type is common in academic and popular management literature. For further discussion see R. W. Keidel, 1985, 1987.
2. Regulation is not dead in the industrial world, however. Concerns for the environment are being substituted for distributional and other economic objectives by those seeking to redefine the existing social compact. Integrated resource planning and some forms of demand-side management now being pursued by regulators are objectives being imposed in the name of the general public interest.
3. The term "regulation" as used here generally refers to economic regulation. Most countries have a variety of licensing and other requirements with respect to the physical siting of facilities, health and safety requirements, and other issues. Although regulations may in some cases be specific to a sector, they are usually part of an overall package of rules governing most industrial and commercial activities.
4. Third-party access to transmission lines (or pipelines, in the case of oil and gas) occurs when a seller of the service other than the owner of the transmission line has access to the physical facilities for the transport of power on its own account. Essentially, the owner of the transmission facility acts as a common carrier. Third-party access generally is strongly opposed by the existing carrier because it gives competitors direct access to customers in its previously monopolized territory.
5. New Zealand has split up the three functions of the system, but they remain under public ownership. Although competition is permitted in generation, the existence of excess capacity in the dominant monopoly has effectively discouraged entry. Whether it will be possible to maintain the transmission and system management functions as a public monopoly remains to be seen; private ownership of parts of the generation and distribution facilities is likely to make continued public control difficult.

### References

- Baumol, W. J., and others. 1988. *Contestable Markets and the Theory of Industry Structure*. New York: Harcourt Brace Jovanovich.
- Churchill, Anthony. 1994. "Meeting Hydroelectric's Financing and Development Challenges." *Hydro Review World Wide* 2: 22-32.
- Hirschman, A. O. 1970. *Exit, Voice, and Loyalty*. Cambridge, Mass.: Harvard University Press.
- Keidel, R. W. 1985. *Game Plans: Sports Strategies for Business*. New York: Dutton.
- . 1987. "Team Sports Models as a Generic Organizational Framework." *Human Relations* 40: 591-612.
- World Bank. 1990. "Understanding the Costs and Schedules of World Bank-Supported Hydro-electric Projects." Industry and Energy Paper 31. Industry and Energy Department, Washington, D.C.







