

Timing and Duration of Exposure in Evaluations of Social Programs

Elizabeth M. King
Jere R. Behrman

The World Bank
Development Research Group
Human Development and Public Services Team
August 2008



Abstract

Impact evaluations aim to measure the outcomes that can be attributed to a specific policy or intervention. Although there have been excellent reviews of the different methods that an evaluator can choose in order to estimate impact, there has not been sufficient attention given to questions related to timing: How long after a program has begun should one wait before evaluating it? How long should treatment groups be exposed to a program before they can be expected to benefit from it? Are there important time patterns in a program's impact? Many impact evaluations assume that interventions

occur at specified launch dates and produce equal and constant changes in conditions among eligible beneficiary groups; but there are many reasons why this generally is not the case. This paper examines the evaluation issues related to timing and discusses the sources of variation in the duration of exposure within programs and their implications for impact estimates. It reviews the evidence from careful evaluations of programs (with a focus on developing countries) on the ways that duration affects impacts.

This paper—a product of the Human Development and Public Services Team, Development Research Group—is part of a larger effort in the department to improve the technical quality of impact evaluations in the World Bank. Policy Research Working Papers are also posted on the Web at <http://econ.worldbank.org>. The author may be contacted at eking@worldbank.org.

The Policy Research Working Paper Series disseminates the findings of work in progress to encourage the exchange of ideas about development issues. An objective of the series is to get the findings out quickly, even if the presentations are less than fully polished. The papers carry the names of the authors and should be cited accordingly. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

Timing and Duration of Exposure in Evaluations of Social Programs

Elizabeth M. King and Jere R. Behrman

Forthcoming, *World Bank Research Observer*

Keywords: Impact evaluation, duration of program exposure, social programs in developing countries, implementation delays

Corresponding author: Elizabeth M. King, The World Bank, 1818 H Street, NW, Washington,

DC 20433; eking@worldbank.org

* We are grateful to Laura Chioda and to three anonymous referees for helpful comments on a previous draft. All remaining errors are ours. The opinions and conclusions expressed in this paper do not necessarily reflect those of the Executive Directors and officials of the World Bank or of its member governments.

Introduction

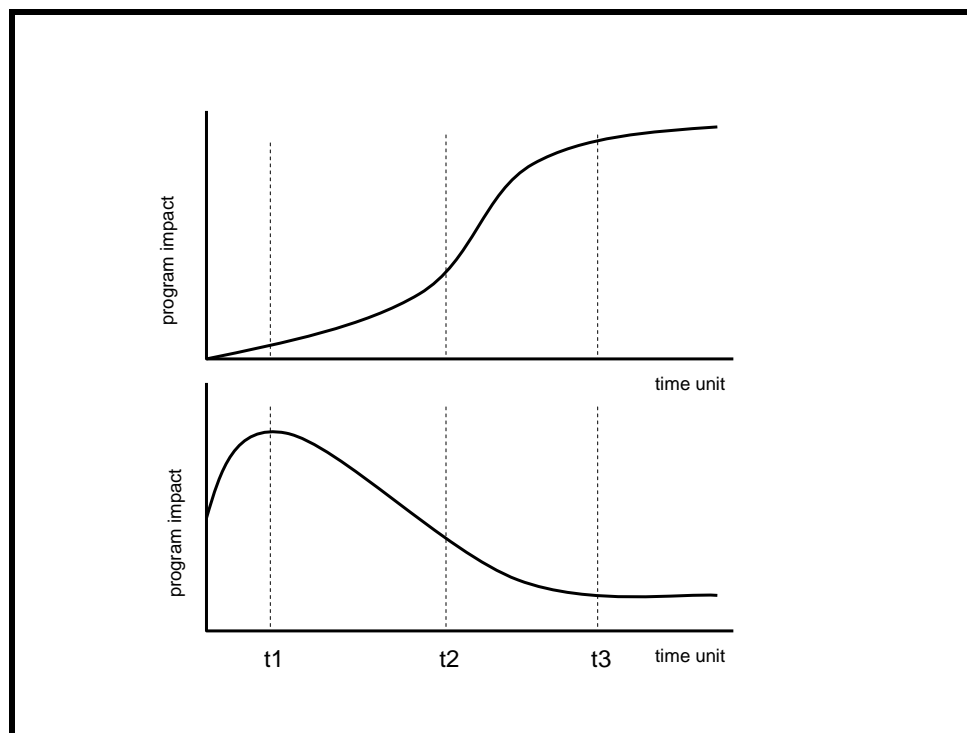
A critical risk that faces all development aid is that it will not pay off as expected—or that it will not be perceived as effective—in reaching development targets. Despite billions of dollars spent to improve health, nutrition, learning and household welfare, surprisingly little is known about the impact of specific social programs in developing countries. One reason for this is that governments and the development community tend to scale up programs fast even in the absence of credible evidence, reflecting an overwhelming impatience with waiting to pilot and assess new programs adequately before expanding them. This impatience is understandable given the urgency of the problems being addressed, but such impatience can result in costly but avoidable mistakes and failures; the same impatience can also result in really promising new programs being terminated too soon after a rapid assessment shows negative or no impact.

Recent promises for substantially more aid from rich countries as well as large private foundations, however, have intensified interest in assessing aid effectiveness. This interest is reflected in a call for more evaluations of the impact of donor-funded programs in order to learn which interventions work, which do not and why.¹ Researchers are responding enthusiastically to this call. There have been important developments in evaluation methods as they apply to social programs, especially on the question of how best to identify a group with which to compare intended program beneficiaries, that is, a group of people who would have had the same outcomes as the program group without the program.²

Arguably as important but relatively understudied, however, is the timing question in evaluations. This question has many dimensions: One pertains to how long after a program has been launched one should wait before evaluating it, how long treatment groups should be exposed to a program before they can be expected to benefit from it, either partially or fully, and

how to take account of the heterogeneity in impact that is related to the duration of exposure. This timing issue is relevant for all evaluations but particularly so for the evaluation of social programs that require changes in the behaviors of both service providers and service users in order to bring about measurable outcomes. If one evaluates too early, one risks finding only partial or no impact; if one waits too long, one risks losing donor and public support for the program or a scaling up of a badly-designed program. Figure 1 illustrates this point by indicating that the true impact of a program may not be immediate or constant over time, for reasons that we discuss in this paper. Comparing two hypothetical programs for which the time path of impact differs, we see that an evaluation undertaken at time $t1$ would indicate that the case in the bottom panel has a higher impact than the case in the top panel, while an evaluation at time $t3$ would suggest the opposite result.

Figure 1. The timing of evaluations can affect impact estimates



This paper discusses key issues related to the timing of programs and the time path of their impact, and how these have been addressed in evaluations.³ Many evaluations treat interventions as if they were instantaneous, predictable changes in conditions and equal across treatment groups. Many evaluations also implicitly assume that the impact on individuals is dichotomous (that is, that individuals are either exposed or not) as might be the case in a one-shot vaccination program that provides permanent immunization; there is no consideration of the possibility that the effects are different depending on variations in program exposure.⁴ Whether the treatment involves immunization or a more process-oriented program such as community organization, the unstated assumptions are often that the treatment occurs at a specified inception date and that it is implemented completely and in precisely the same way across treatment groups.

There are several reasons why implementation is neither immediate nor perfect, why the duration of exposure to a treatment differs not only across program areas but also across ultimate beneficiaries, and why varying lengths of exposure might lead to different estimates of program impact. This paper discusses three broad sources of variation in duration of exposure, and reviews the literature related to those sources (see Appendix Table 1 for a list of the studies reviewed). One source pertains to organizational factors that affect the leads and lags in program implementation, as well as to timing issues related to program design and the objectives of an evaluation. A second source refers to spillover effects, including variation that arises from the learning and adoption by beneficiaries and possible contamination of the control groups. Spillover effects are external (to the program) sources of variation in the treatment; while these may pertain more to compliance than timing, they can appear and intensify with time and so affect estimates of program impact. A third source pertains to heterogeneous responses to

treatment. Although there can be different sources of heterogeneity in impact, the focus here is on those associated with age or cohort, especially as these cohort effects interact with how long a program has been in effect.

Organizational Factors and Variation in Program Exposure

Program design and the timing of evaluations

How long one should wait to evaluate a program depends on the nature of the intervention itself and the purpose of the evaluation. For example, in the case of HIV/AIDS or tuberculosis treatment programs, adherence to the treatment regime over a period of time is necessary for the drugs to be effective. While drug effectiveness in treating the disease is likely to be the outcome of interest, an evaluation of the program might also consider adherence rates as an intermediate outcome of the program—and so the evaluation need not take place only at the end of the program but during the implementation itself. In the case of worker training programs, workers must first enroll for the training and then some time passes during which the training occurs. If the training program has a specific duration, the evaluation should take place after the completion of the training program. However, timing may not be so easy to pin down if the timing of the intervention itself is the product of a stochastic process, such as when a market downturn causes workers to be unemployed, triggering their eligibility for worker training, and when a market upturn causes trainees to quit the program to start a job (as Ravallion et al. 2005 observes in Argentina's *Trabajar* workfare program). In cases where the timing of entry into (or exit from) a program itself differs across potential beneficiaries, the outcomes of interest depend on an individual selection process as well as on the passage of time, and an evaluation of these programs should address selection bias. Randomized evaluations of trials with well-defined start and end dates do not address this issue.

The timing of a program may, in fact, be used for identification purposes. For example, some programs are implemented in phases, and if the phasing is applied randomly, the random variation in duration can be used for identification purposes in estimating program impact (Rosenzweig and Wolpin 1986 is a seminal article on this point). One example is Mexico's PROGRESA (Programa de Educación, Salud y Alimentación) which was targeted to the poorest rural communities when it began. The program identified the potential beneficiaries using administrative and census data on measures of poverty (Behrman and Todd 1999). Out of the 506 communities identified for the evaluation sample, about two-thirds were randomly selected to receive the program activities during the first two years of the program starting in mid-1998, while the remaining one-third received the program in the third year, starting in the fall of 2000. The group that received the intervention later has been used as a control group in evaluations of PROGRESA (see, for example, Schultz 2004 or Behrman, Sengupta and Todd 2005).

One way to regard duration effects is that, given constant dosage or intensity of a treatment, lengthening duration of exposure is akin to increasing intensity and thus the likelihood of greater impact. Two cases show that impact is likely to be underestimated if the evaluation coverage is too short. First, skill development programs are an obvious example of the importance of the duration of program exposure because beneficiaries who attend only part of a training course are less likely to benefit from the course and attain the program goals than those who complete the course. In evaluating the impact of a training course attended by students, Rouse and Krueger (2004) distinguish between students who completed the computer instruction offered through the Fast ForWord Program and those who did not. The authors define completion as a function of the amount of training attended and the actual progress of students toward the next stage of the program as reflected in the percentage of exercises at the current

level mastered at a prespecified level of proficiency.⁵ The authors find that, among students who received more comprehensive treatment as reflected by the total number of completed days of training and the level of achievement of the completion criteria, performance improved more quickly on one of the reading tests (but not all) that the authors use.

Banerjee et al. (2007) evaluate two randomly assigned programs in urban India: a remedial training program that hired young women to teach children with low literacy and numeracy skills, and a computer-assisted learning program. Illustrating the point that a longer duration of exposure intensifies treatment, the remedial program raised average test scores by 0.14 standard deviation in the first year and 0.28 standard deviation in the second year of the program, while computer-assisted learning increased math scores by 0.35 standard deviations in the first year and 0.47 standard deviations in the second year. The authors interpret the larger estimate in the second year as indicating that the first year of the program laid the foundation for the program to help the children benefit from the second year of the program.

Lags in implementation

One assumption that impact evaluations often make is that once a program starts, its implementation occurs at a specific and knowable time that is usually determined at a central program office. Program documents, such as World Bank project loan documents, typically contain official project launch dates, but these dates often differ from the date of actual implementation in a project area because when a program actually is initiated depends on supply- and demand-related realities in the field. For example, a program requiring material inputs (such as textbooks or medicines) relies on when those inputs arrive in the program areas; the timing of the procurement of the inputs by the central program office is not necessarily an accurate indicator of when those inputs arrive at their intended destinations.⁶ In a large early childhood

development program in the Philippines, administrative data indicate that the timing of the implementation differed substantially across program areas; because of lags in central procurement, not all providers in the program areas had received required training three years after project launch (Armeccin et al. 2006). Besides material inputs, snags in information flows and project finances can also delay implementation. Delays in providing the information about intended household beneficiaries in Mexico's and Ecuador's conditional cash transfer programs prevented program operators in some sites from making punctual transfers to households (Rawlings and Rubio 2005; Schady and Araujo 2008).⁷ In Argentina, the targeting performance of its Trabajar program was weakened by poor municipalities having a harder time raising cofinancing required for the sub-projects (Ravallion 2000).

The problem of implementation lags can be partly addressed if careful and complete administrative data on timing are available for the program; cross-referencing such data with information from public officials or community leaders in treatment areas could reveal the institutional reasons for variation in implementation. For example, if there is an average gap of one year between program launch and actual implementation, then it is reasonable for the evaluation to make an allowance of one year after program launch before estimating program impact.⁸ However, actual information on dates often is not readily available so studies have tended to allot an arbitrary grace period to account for lags.

Assuming a constant allowance for delays, moreover, may not be an adequate solution if there is wide variation in the timing of implementation across treatment areas as a result of the program involving a large number of geographical regions or a large number of components and actors. In programs that cover several states or provinces, region or state fixed effects might control for duration differences if the differences are homogeneous within a region or state and if

the delays are independent of unobservable characteristics in the program areas that may also influence program impact. An evaluation of Madagascar's SEECALINE program defined the start of the program in each treatment site as the date of the first child-weighting session in that site. The area-specific date takes into account the program's approach of gradual and sequential expansion, as well as the expected delays between the signing of the contract with the implementing NGO and when a treatment site is actually open and operational (Galasso and Yau 2006). This method requires detailed program monitoring data.

If a program has many components, the solution could hinge on the evaluator's understanding of the technical production function and thus on identifying the elements that must be present for the program to be effective. For example, in a school improvement program that requires additional teacher training as well as instructional materials, the materials might arrive in schools at about the same time, but the additional teacher training might be achieved only over a period of several months, perhaps due to differences in teacher availability. Whether the effective program start should be defined according to the date when the materials arrive in schools or when all (or majority?) of the teachers have completed their training is a question that an evaluator has to address in deciding on the timing of the evaluation. In the Madagascar example above, although the program has several components (e.g., growth monitoring, micronutrient supplementation, de-worming), the inception date of each site was pegged to a growth monitoring activity, that of the first weighing session (Galasso and Yau 2006).

Although the primary objective of evaluations usually is to measure the impact of programs, often they also monitor progress during the course of implementation and thus help to identify problems that need correction. An evaluation of the Bolivia Social Investment Fund illustrates this point clearly (Newman et al. 2002). One of the program components was to

improve the drinking water supply through investments in small-scale water systems. The first laboratory analysis of water quality, however, showed little improvement in program areas. Interviews with local beneficiaries explained why: contrary to plan, personnel designated to maintain water quality lacked training, inappropriate materials were used for tubes and the water tanks, and the lack of water meters made it difficult to collect fees needed to finance maintenance work. After training was provided in all the program communities, a second analysis of water supply indicated significantly less fecal contamination in the water in those areas.

How do variation in program start and lags in implementation affect estimates of impact? Variation in program exposure that is not incorporated into the estimation of impact almost surely biases downward the intent-to-treat (ITT) estimates of program impact, especially if such impact increases with the exposure of the beneficiaries who are actually treated. But the size of this underestimation, for a given average lag across communities, depends on the nature of the lags. If the program implementation delays are not random, it matters if they are inversely or directly correlated with unobserved attributes of the treated groups that may positively affect program success. If the implementation lags are *directly* correlated with unobserved local attributes (e.g., central administrators may put more effort into getting the programs started in areas that have worse unobserved determinants of the outcomes of interest), then the true ITT effects are underestimated to a larger extent. If implementation delays are instead *inversely* associated with the unobserved local attributes (e.g., the management capability of local officials to implement a program may be weaker also in areas that are worse off in terms of the unobserved determinants of desired outcomes), then the ITT effects are underestimated to a lesser extent. If instead the program delays are random, the extent of the underestimation

depends on the variance in the implementation lags (still given the same mean lag). Greater random variance in the lags, all else equal, results in greater underestimation of the ITT effects because of a larger classical random measurement error in a right-side variable that biases the estimated coefficient more towards zero.

Implementation delays per se do not necessarily affect estimates of treatment-on-the-treated (TOT) effects if the start of the treatment for individual beneficiaries has been identified correctly. In some cases, this date of entry can be relatively easy to identify as, for instance, when the dates on which beneficiaries enroll in a program can be captured through a household or facility survey or administrative records (e.g., school enrollment rosters or clinic logbooks). In other cases, however, the identification may be more difficult, such as when beneficiaries are unable to distinguish among alternative, contemporaneous programs or to recall their enrollment dates or when the facility or central program office does not monitor beneficiary program enrollments.⁹ Nonetheless, even if the variation in treatment dates within program areas is handled adequately and enrollment dates are identified fairly accurately at the beneficiary level, nonrandom implementation delays bias TOT estimates. Even a well-specified facility or household survey is still going to leave the concern of unobservables that may be related to the direction and size of the program impact, and the duration of exposure, like program take-up, has to be treated as endogenous. The problem of selection bias motivates the choice of random assignment to estimate treatment effects in social programs.

Learning by providers

A different implementation lag is associated with the fact that program operators (or providers of services) themselves face a learning curve that depends on time in training and on-the-job experience, most likely producing some variation in the quality of program

implementation that is independent on whether there has been a lag in the procurement of the training. This too is an aspect of the reality of program operation that often is not captured in impact evaluations. The evaluation of Madagascar's SEECALINE program allotted a grace period of 2-4 months for the training of service providers but it is more likely that much of the learning by providers happened on-the-job after the formal training.

While the learning process of program operators may delay full program effectiveness, another effect could be working in the opposite direction. The "pioneering effect" means that implementers exhibit extra dedication, enthusiasm, and effort during the first stages because the program represents an innovative endeavor to attain an especially important goal. (A simplistic diagram of this effect is shown in Figure 1, bottom panel.) Jimenez and Sawada (1999) find that newer EDUCO schools in El Salvador had better outcomes than older schools, holding constant the school characteristics, a finding that they interpret as evidence of a Hawthorne effect—that is, newer schools were more motivated and willing to undertake reforms than the older schools. If such a phenomenon exists, it would exert an opposite pull on the estimated impacts and, if sufficiently strong, might offset the learning effect, at least in the early phases of a new program. Over time, however, this extra dedication, enthusiasm, and effort are likely to wane. If there are heterogeneities in this unobserved pioneering effect across program sites that are correlated with observed characteristics (e.g., schooling of program staff), the result will be biased estimates of the impact of such characteristics on initial program success.

Spillover Effects

The observable gains from a social program during its entire existence, much less after only a few years of implementation, could be an underestimate of its full potential impact for several reasons that are external to the program design. First, evaluations are typically designed

to measure outcomes at the completion of a program, and yet the program might yield additional and unintended outcomes in the longer run. Second, while the assignment of individuals or groups of individuals to a treatment can be defined, program beneficiaries may not actually take up an intervention—or may not do so until after they have learned more about the program. Third, with time, control groups or groups other than the intended beneficiaries might find a way of obtaining the treatment, or they may be affected simply by learning about the existence of the program, possibly because of expectations that the program will be expanded to their area. If non-compliance is correlated with the outcome of interest, then the difference in the average outcomes between the treatment and the control groups is a biased estimate of the average effect of the intervention. We discuss these three examples below.

Short-run and long-run outcomes

Programs that invest in cumulative processes such as a child's physiological growth and accumulation of knowledge require the passage of time, implying that longer program exposure would yield greater gains, though probably with diminishing marginal returns. In addition, such cumulative processes could lead to outcomes beyond those originally intended—and possibly beyond those of immediate interest to policymakers. Early childhood development (ECD) programs are an excellent example of short-run outcomes that could lead to long-run outcomes outside those envisioned by the original design. These programs aim to mitigate the multiple risks facing very young children and promote their physical and mental development by improving nutritional intake and/or cognitive stimulation. The literature review by Grantham-McGregor et al. (2007) identifies studies that use longitudinal data from Brazil, Guatemala, Jamaica, the Philippines and South Africa that establish causality between pre-school cognitive development and subsequent schooling outcomes. The studies suggest that one standard

deviation increase in early cognitive development predicts substantially improved school outcomes in adolescence, as measured by test scores, grades attained and dropout behavior (e.g., 0.71 additional grade by age 18 in Brazil).

Looking beyond childhood, Garces et al. (2002) find evidence from the U.S. Head Start program that links preschool attendance not only to higher educational attainment but also to higher earnings and better adult social outcomes. Using longitudinal data from the Panel Study of Income Dynamics, they conclude that, controlling for the participants' disadvantaged background, exposure to Head Start for whites is associated with significantly lower dropout rates in the short run, and in the long run 30 percent greater probability of high school completion and 28 percent higher likelihood of attending college, as well as higher earnings in their early twenties. For African-Americans, participation in Head Start is associated with a 12-percentage-points lower probability of being booked for or charged with a crime.

Another example of an unintended long-run outcome is provided by Angrist et al.'s (2004) evaluation of Colombia's school voucher program at the secondary level (PACES or Programa de Ampliación de Cobertura de la Educación Secundaria), which finds longer-run outcomes beyond the original program goal of increasing the secondary school enrollment rate of the poorest youths in urban areas. Using administrative records, the follow-up study finds that the program increased also high school graduation rates of voucher students in Bogota by 5-7 percentage points, consistent with the earlier outcome of a 10-percentage point increase in eighth-grade completion rates (Angrist et al. 2002), and correcting for the greater percentage of lottery winners taking college admissions tests, the program increased test scores by two-tenths of a standard deviation in the distribution of potential test scores.

In their evaluation of a rural roads project in Vietnam, Mu and van de Walle (2007) find

that rural road construction and rehabilitation produced larger gains as more time elapsed after project completion because of developments external to the program. The impacts of roads depend on people using them, so the increased availability of bicycles or motorized vehicles to rural populations connected by the roads is necessary for the benefits of the roads project to be apparent. But the impact of the new roads also include other developments that have arisen more slowly, such as a switch from agriculture to non-agricultural income-earning activities and an increase in secondary schooling following on a rise in primary school completion. These impacts grew at an increasing rate as more months passed, taking two years more on average to emerge.

In the long-run, however, impacts can also vanish so estimates based on short periods of exposure are not likely to be informative about issues such as the extent of diminishing marginal returns to exposure that would be an important part of the information basis of policies. In Vietnam, the impact of the rural roads project on the availability of foods and on employment opportunities for unskilled jobs emerged quite rapidly and then waned as the control areas caught up with the program areas, an effect we return to below (Mu and van de Walle, 2007). In Jamaica, a nutritional supplementation-cum-psychological stimulation program for children under two yielded mixed effects on cognition and education years later (Walker et al. 2005). While the interventions benefited child development and even at age 11 stunted children who received stimulation continued to show cognition benefits, small benefits from supplementation noted at age 7 were no longer present at age 11. In fact, impact could vanish much sooner after a treatment ends. In the example of two randomized trials in India, although impact rose in the second year of the program, one year *after* the programs had ended, impact dropped: for the remedial program, the gain fell to 0.1 of a standard deviation and was no longer statistically

significant; for the computer learning program, the gain dropped to 0.09 of a standard deviation though still significant (Banerjee et al. 2007).

Chen, Mu and Ravallion (2008) point to how longer-term effects might be invisible to evaluators in their evaluation of the long-term impact of the Southwest China Project which gave selected poor villages in three provinces funding for a range of infrastructure investments and social services. The authors find only small and statistically insignificant average income gains in the project villages four years after the disbursement period. They attribute this partly to significant displacement effects which are due to the government cutting the funding for non-project activities in the project villages and reallocating resources to the non-project villages. Because of these displacement effects, the estimated impacts of the project are likely to be underestimated. To estimate an upper bound on the size of this bias, they assume that the increase in spending in the comparison villages is equal to the displaced spending in the project villages. Under this assumption, the upper bound of the bias could be as high as 50 percent—and it could be even larger if the project actually has positive long-term benefits.

Long-term benefits, however, are often not a powerful incentive to support a program or policy. The impatience of many policymakers with a pilot-evaluate-learn approach to policymaking and action is usually coupled with a high discount rate that results in little appetite to invest in programs for which benefits are mostly enjoyed in the future. Even aid agencies exhibit this impatience and yet programs that have long-run benefits would be just the sort of intervention that development aid agencies should support because local politicians are likely to dismiss them.

Learning and adoption by beneficiaries

Programs do not necessarily attain full steady-state effectiveness after implementation

commences because learning by providers and beneficiaries takes time, or because a transformation of accountability relationships that may be necessary does not happen immediately, or because the behavioral responses of providers and consumers may be slow in becoming apparent.

The success of a new child immunization or nutrition program depends on parents learning about the program and bringing their children to the providers and the providers giving treatment. In Mexico's PROGRESA the interventions were randomly assigned at the community level so a simple comparison between eligible children in the control and treatment localities would have been sufficient to estimate the program TOT effect if program uptake were perfect (Behrman and Hoddinott 2005). However, not all potential beneficiaries sought services: only 61–64 percent of the eligible children aged 4 to 24 months and only half of those aged 2 to 4 years actually received the program's nutritional supplements. The evaluation found no significant ITT effects, but did find that the TOT effects were significant despite individual and household controls.

In Colombia's secondary education voucher program, information played a role at both the local government level and the student level (King, Orazem and Wohlgemuth 1999). Since the program was co-funded by the central and municipal governments, information to municipal governments was critical to securing their collaboration. At the beginning of the program, the central government met with the heads of the departmental governments to announce the program and solicit their participation, and in turn, the departmental governors invited municipal governments to participate. In addition, dissemination of information to families was particularly important because participation was voluntary and the program was targeted only to certain students (specifically those living in neighborhoods classified among the two lowest

socioeconomic strata in the country) based on specific eligibility criteria. Some local governments used newspapers to disseminate information about the program.

In decentralization reforms, the learning and adoption processes are arguably more complex because the decision to participate and the success of implementation depend on many more actors. Even the simplest form of this type of change in governance (say, the transfer of the supervision and funding of public hospitals from the national government to a subnational government) entails a shift in the accountability relationships between levels of government and between governments and providers. In Nicaragua's autonomous schools program in the 1990s, for example, the date a school signed the contract with the government was considered to be the date the school officially became autonomous. In fact, the signing of the contract was merely the first step toward school autonomy, and it would have been followed by training activities, the election of the school management council, the development of a school improvement plan, and so on. Hence, the full impact of the reform on outcomes would have been felt only after a period of time, and the size of this impact might have increased gradually as the elements of the reform were put in place. It is not easy, however, to determine the length of the learning period; the evaluation finds a lack of agreement among teachers, school directors and parents in the so-called autonomous schools on whether their schools had become autonomous and the extent to which this had been achieved (King and Özler 1998). An in-depth qualitative analysis in a dozen randomly selected schools confirms that school personnel had different interpretations of what had been achieved (Rivarola and Fuller 1999).

Studies of the diffusion of the Green Revolution in Asia in the mid-1960s highlight the role of social learning among beneficiaries. If an individual learns about a new technology from the experiences of his neighbors (their previous decisions and outcomes) before adopting the

technology; this wait-and-see process accounts for some of the observed lags in the adoption of high-yielding seed varieties in India at the time (Foster and Rosenzweig 1995; Munshi 2004). In rice villages, the proportion of farmers who adopted the new seed varieties rose from 26 percent in the first year following the introduction of the technology to 31 percent in the third year; in wheat villages, the proportion of adopters increased from 29 percent to 49 percent. Farmers who did not have neighbors with comparable attributes (such as farm size or characteristics unobserved in available data such as soil quality) may have had to carry out more of their own experimentation, which was probably a more costly form of learning because the farmers bore all the risk of the choices they made (Munshi 2004).

The learning process at work during the Green Revolution is similar to that described by Miguel and Kremer (2003, 2004) about the importance of social networks in the adoption of new health technology, in this case deworming drugs. Survey data on individual social networks of the treatment group in rural Kenya reveal that social links provide non-treatment groups better information about the deworming drugs and thus lead to higher program take-up. Two years after the start of the deworming program, school absenteeism among the treatment group had fallen by about one-quarter on average, and there were significant gains in several measures of health status, including reductions in worm infection, child growth stunting and anemia, and gains in self-reported health. But children whose parents had more social links to early treatment schools were significantly less likely to take deworming drugs. The authors speculate that this disappointing finding could be due to overly optimistic expectations about the impact of the drugs or to the fact that the health gains from deworming take time to be realized while the side effects of the drugs are immediately felt.

Providing information about a program, however, is no guarantee of higher program

uptake. One striking example of this is captured by an evaluation of a program in Uttar Pradesh, India which aimed to strengthen community participation in public schools by providing information to village members (Banerjee et al. 2008). More information apparently did not lead to higher participation by the Village Education Committee (VEC), by parents, or by teachers. The evaluators attribute this poor result to more deep-seated information blockages: the fact that village members were unaware of the roles and responsibilities of the VEC despite the fact that these committees had been in existence since 2001 and that a large proportion of the VEC members were not even aware of their membership.

To the extent that the nutritional component in PROGRESA was undersubscribed because parents lacked information about the program and its benefits or that the community mobilization in Uttar Pradesh failed because basic information about the roles and powers of village organizations is difficult to convey, impact evaluations that do not take information diffusion and learning by beneficiaries into account obtain downward-biased ITT and TOT impact estimates. The learning process might be implicit, such as when program information diffuses to potential beneficiaries during the course of implementation and perhaps primarily by word-of-mouth, or it could be explicit as when a program includes an information campaign to potential beneficiaries during a well-defined time period.

Two points are worth noting about the role of learning in impact evaluation. One is the simple point discussed above—that learning takes time. Related to the process of expanding effective demand for a program, there is a steady-state level of effective demand among potential beneficiaries (effective in the sense that the beneficiaries actually act to enroll in or use program services).¹⁰ This implies that ITT estimates of program impact are biased downward if the estimates are based on data obtained prior to the attainment of this steady-state effective demand.

The extent of the bias depends on whether learning (or the expansion of effective demand) is correlated with unobserved program attributes; specifically, there is less downward bias if this correlation is positive. There may be heterogeneity in this learning process, that is, those programs that have better unobserved management capabilities promote more rapid learning, while those that have worse management capabilities are faced with slower learning. Heterogeneity in learning would affect the extent to which the ITT and TOT impacts that are estimated before a program has approached effectiveness are downward-biased, but less so if the heterogeneity in learning is random.

The second point is that the learning process itself may be an outcome of interest in an impact evaluation. How beneficiaries learn and decide to participate is often external to a program since the typical assumption is that beneficiaries will take up a program if the program exists. In fact, the exposure of beneficiaries to specific communication interventions about a program may be necessary to encourage program uptake. There is a large literature, for example, that shows a strong association between exposure to information campaigns through mass media and the use of contraceptive methods and family planning services. Such campaigns have been aimed at informing potential beneficiaries of the availability of services as well as at breaking down sociocultural resistance to the services (e.g., Westoff and Rodriguez 1995, Adongo et al. 1997). To understand how learning takes place is to begin to understand the “black box” that lies between program design and outcomes—and if this learning were promoted in a random fashion, it could serve as an exogenous instrument for the estimation of program impact.

Peer effects

The longer a program has been in operation, the more likely it is that specific interventions will spill over to populations beyond the treatment group and thus affect impact

estimates. Peer effects increase impact as in the case of the Head Start example already mentioned. Garces et al. (2002) find strong within-family spillover effects—and higher birth-order children (that is, younger siblings) seem to benefit more than their older siblings, especially among African-Americans, because older siblings are able to teach younger ones. Hence, expanding the definition of impact to include peer effects adds to impact estimates.

Peer effects also arise when specific program messages (either directly from communications interventions or from observing treatment groups) diffuse to control groups and alter their behavior in the same direction as in the treatment group. While this contagion is probably desirable from the point of view of policymakers, it likely depresses impact estimates since differences between the control and treatment groups are diminished. Another form of leakage that grows with time may not be so harmless from the point of view of program objectives. For programs that are targeted only to specific populations, time allows political pressure to build for the program to be more inclusive and even for non-targeted groups to find ways of obtaining treatment (such as through migration into program sites). Because of the demand-driven nature of the Bolivia Social Investment Fund, for instance, not all communities selected for active promotion applied for and received a SIF-funded education project, but some communities not selected for active promotion nevertheless applied for promotion and obtained an education project (Newman et al. 2002).

Heterogeneity of Impact

How program impact varies by the observable characteristics of the beneficiaries holds important lessons for policy and program design. Our focus here is when duration or timing differences interact with the sources of heterogeneity in impact. One important source of heterogeneity in some programs is cohort membership.

Cohort effects

The age of beneficiaries may be one reason why duration of exposure to a program matters, and the estimates of ITT and TOT impacts can be affected substantially by whether the timing with regard to the ages of the beneficiaries is targeted toward critical age ranges. Take the case of early childhood development (ECD) programs, such as infant feeding and preschool education, which target children for just a few years after birth. This age targeting is based on the evidence that a significant portion of a child's physical and cognitive development occurs at a very young age, and that the returns to improvements in the living or learning conditions of the child are highest at those ages. The epidemiological and nutritional literatures emphasize that children under three years of age are especially vulnerable to malnutrition and neglect (see Engle et al. 2007 for a review). Finding that a nutritional supplementation program in Jamaica did not produce long-term benefits for children, Walker et al. (2005) suggest that prolonging the supplementation or supplementing at an earlier age, during pregnancy and soon after birth, may have benefited later cognition and may have been more effective than the attempt to reverse the effects of undernutrition through supplementation at an older age. Applying evaluation methods to drought shocks, Hoddinott and Kinsey (2001) also conclude that in rural Zimbabwe children in the age range of 12 to 24 months are the most vulnerable to drought shocks; these children lose 1.5–2 centimeters of physical growth, while older children 2 to 5 years of age do not seem to experience a slowdown in growth.¹¹ In a follow-up study, Alderman, Hoddinott and Kinsey (2006) conclude that the longer the exposure of young children to civil war and drought, the larger the negative effect of these shocks on child height; moreover, older children suffer less than younger children in terms of growth.¹²

Interaction of cohort effects and duration of exposure

As discussed above, the impacts of some programs crucially depend on whether or not an intended beneficiary is exposed to an intervention at a particularly critical age range, such as during the first few years of life. Other studies illustrate that the duration of exposure during the critical age range also matters. Frankenberg, Suriastini, and Thomas's (2005) evaluation of Indonesia's Midwife in the Village Program show just this. The program was supposed to expand the availability of health services to mothers and thus improve children's health outcomes. By exploiting the timing of the (nonrandom) introduction of a midwife to a community, the authors distinguish between the children who were exposed to a midwife at a very young age and the older children who were living in the same community but who had not been exposed to a midwife when they were very young. The authors group the sample of children into three birth cohorts; for each group, the extent of exposure to a village midwife during the vulnerable period of early childhood varied as a function of whether the village had a midwife and, if so, when she had arrived. In communities that had a midwife from 1993 onward, children in the younger cohort had been fully exposed to the program when data were collected, whereas children in the middle cohort had been only partially exposed. The authors conclude that partial exposure to the village midwife program conferred no benefits in improved child nutrition, while full exposure from birth yielded an increase in the height-for-age Z-score of 0.35 to 0.44 standard deviations among children aged 1 to 4.

Three other studies test the extent to which ECD program impacts are sensitive to the duration of program exposure and the ages of the children during the program. Behrman, Cheng, and Todd (2004) evaluated the impact of a preschool program in Bolivia, the Proyecto Integral de Desarrollo Infantil. Their analysis explicitly takes into account the dates of program enrollment of individual children. In their comparison of treated and untreated children, they find

evidence of positive program impacts on motor skills, psychosocial skills, and language acquisition that are concentrated among children 37 months of age and older at the time of the evaluation. When they disaggregated their results by the duration of program exposure, the effects were most clearly observed among children who had been involved in the program for more than a year.

Like the Bolivia evaluation, the evaluation of the early childhood development program in the Philippines mentioned above finds that the program impacts vary according to the duration of exposure of children, although this variation is not as dramatic as the variation associated with children's ages (Armezin et al. 2006). Administrative delays and the different ages of children at the start of the program resulted in the length of exposure of eligible children varying from 0 to 30 months, with a mean duration of 14 months and a substantial standard deviation of six months. Duration of exposure varied widely even controlling for a child's age. The study finds that children 2- and 3-year-olds exposed to the program had Z-scores 0.5 to 1.8 standard deviations higher for motor and language development than children in the control areas, and that these gains were much lower among older children.

Gertler (2004) also estimates how duration of exposure to health interventions in Mexico's PROGRESA affects the probability of child illness using two models—one assumes that program impact is independent of duration, and the other allows impact to vary according to the length of exposure. The interventions required that children under two years be immunized, visit nutrition monitoring clinics, and obtain nutritional supplements, and that their parents receive training on nutrition, health, and hygiene; children between two and five years of age were expected to have been immunized already, but were to obtain the other services. Gertler finds no program impact after a mere six months of program exposure for children under three

years of age, but with 24 months of program exposure the illness rate among the treatment group was about 40 percent lower than the rate among the control group, a difference that is significant at the 1 percent level.

The interaction of age effects and the duration of exposure has been examined also by Pitt, Gibbons and Rosenzweig (1993) and Duflo (2001) in Indonesia and by Chin (2005) in India in their evaluations of schooling programs. These studies use information on the region and year of birth of children, combined with administrative data on the year and placement of programs, to measure duration of program exposure. Duflo (2001), for example, examines the impact of a massive school construction program on the subsequent schooling attainment and on the wages of the birth cohorts affected by the program in Indonesia. From 1973 to 1978, more than 61,000 primary schools were built throughout the country, and the enrollment rate among children aged 7–12 rose from 69 percent to 83 percent. By linking district-level data on the number of new schools by year and matching these data with intercensal survey data on men born between 1950 and 1972, Duflo defines the duration of exposure of an individual to the program. The impact estimates indicate that each new school per 1,000 children increased years of education by 0.12–0.19 percent among the first cohort *fully* exposed to the program.

Chin (2005) uses a similar approach in estimating the impact of India's Operation Blackboard. Taking grades 1–5 as the primary school grades, ages 6–10 as the corresponding primary school ages, and 1988 as the first year that schools would have received program resources, Chin supposes that only students born in 1978 or later would have been of primary school age for at least one year in the program regime and therefore potentially exposed to the program for most of their schooling. The evaluation compares two birth cohorts: a younger cohort born between 1978 and 1983 and was therefore potentially exposed to the program and an

older cohort. The impact estimates suggest that accounting for duration somewhat lowers the impact as measured, but it remains statistically significant, though only for girls.

Conclusions

This paper has focused on the dimensions of timing and duration of exposure that relate to program or policy implementation. Impact evaluations of social programs or policies typically ignore these dimensions; they assume that interventions occur at a specified date and produce intended or predictable changes in conditions among the beneficiary groups. This is perhaps a reasonable assumption when the intervention itself occurs within a very short time period and has an immediate effect, such as some immunization programs, or is completely under the direction and control of the evaluator, as in small pilot programs. In the examples we have cited (India's Green Revolution, Mexico's PROGRESA conditional cash transfer program, Madagascar's child nutrition SEECALINE program, and an early childhood development program in the Philippines, among others), this is far from true. Indeed, initial operational fits and starts in most programs and a learning process for program operators and beneficiaries can delay full program effectiveness; also, there are many reasons why these delays are not likely to be the same across program sites.

We have catalogued sources of the variation in the duration of program exposure across treatment areas and beneficiaries, including program design features that have built-in waiting periods, lags in implementation due to administrative or bureaucratic procedures, spillover effects, and the interaction between sources of heterogeneity in impact and duration of exposure. The findings of some evaluations demonstrate that accounting for these variations in length of program exposure alters impact estimates significantly, so ignoring these variations can generate misleading conclusions about an intervention. Appendix Table 1 indicates that a number of

impact evaluation studies do incorporate one or more of these duration effects. The most commonly addressed source of duration effects is cohort affiliation. This is not surprising since many interventions such as education and nutrition programs are allocated on the basis of age, in terms of both timing of entry into and exit from the program. On the other hand, implementation lags are recognized but often not explicitly addressed.

What can be done to capture timing and the variation in length of program exposure? First, *improve the quality of program data*. Such data could come from administrative records on the design and implementation details of a program, combined with survey data on program take-up by beneficiaries. Program data on the timing of implementation are likely to be available from program management units, but these data may not be available at the desired level of disaggregation, which might be the district, community, providers, or individual depending on where the variation in timing is thought to be the greatest. Compiling such data on large programs that decentralize to numerous local offices could be costly. There is obviously a difference in the primary concern of the high-level program manager and of the evaluator. The program manager's concern is the disbursement of project funds and the procurement of major expenditure items, whereas the evaluator's concern would be to ascertain when the funds and inputs reach treatment areas or beneficiaries.

Second, *choose the timing of the evaluation keeping in mind the time path of program impacts*. Figure 1 illustrates that program impact, however measured, can change over time for various reasons discussed in the paper, so there are risks of not finding significant impact when a program is evaluated too early or too late. The learning process by program operators or by beneficiaries could produce a curve showing increasing impact over time, while a pioneering effect could show a very early steep rise in program impact that is not sustainable. Figure 1 thus

suggests that early rapid assessments to judge the success of a program could be misleading, and also that repeated observations may be necessary to estimate true impact. Several studies that we reviewed measured their outcomes of interest more than once after the start of the treatment, and some compared short-run and long-run effects to examine whether the short-run impact had persisted. Possible changes in impact over time imply that evaluations should not be a once-off activity for any long-lasting program or policy. In fact, as discussed above, examining long-term impacts could point to valuable lessons about the diffusion of good practices over time (Foster and Rosenzweig 1995) or, sadly, about how governments can reduce impact by implementing other policies that (perhaps unintentionally) disadvantage the program areas (Chen, Mu and Ravallion 2008).

Third, *apply an appropriate evaluation method that takes into account the source of variation in duration of program exposure*. As discussed above, impact estimates are affected by the length of program exposure depending on whether or not the source of variation in duration is common within a treatment area and whether or not this source is a random phenomenon. Some pointers are: If the length of implementation lags is about equal across treatment sites, then a simple comparison between the beneficiaries in the treatment and control areas would be sufficient to estimate the average impact of the program or the ITT effects under many conditions—although not if there are significant learning or pioneering effects that differ across them. If the delays vary across treatment areas but not within those areas and if the variation is random or independent of unobservable characteristics in the program areas that may also affect program effectiveness, then it is also possible to estimate the ITT effects with appropriate controls for the area or with fixed effects for different exposure categories. In cases where the intervention and its evaluation are designed together, such as in pilot programs, it is possible and

desirable to explore the time path of program impact by allocating treatment groups to different lengths of exposure in a randomized way. This treatment allocation on the basis of duration differences can yield useful operational lessons about program design, so it deserves more experimentation in the future.

Notes

Research Manager, Development Research Group, World Bank; and William R. Kenan, Jr. Professor of Economics, Department of Economics, University of Pennsylvania.

¹ For instance, the International Initiative for Impact Evaluation (3IE) is being set up by governments of several countries, donor agencies and private foundations to address the desire of the development community to build up systematically more evidence about effective interventions.

² There have been excellent reviews of the choice of methods as applied to social programs. See, for example, Grossman (1994), Heckman and Smith (1995), Ravallion (2001), Cobb-Clark and Crossley (2003), and Duflo (2004).

³ To keep the discussion focused on the timing issue and the duration of exposure, we avoid discussing the specific evaluation method (or methods) that is used by the empirical studies that we cite. However, we restrict our selection of studies to review to those that have a sound evaluation design, whether experimental or using econometric techniques. We also do not discuss estimation issues such as sample attrition bias which is one of the ways in which a duration issue has been taken into account in the evaluation literature.

⁴ See Heckman, Lalonde and Smith (1999) for a review.

⁵ Because Rouse and Krueger (2004) define the treatment group more stringently, however, the counterfactual treatment received by the control students becomes more mixed, and a share of these students is contaminated by partial participation in the program.

⁶ In their assessment of the returns to World Bank investment projects, Pohl and Mihaljek (1992) cite construction delays among the risks that account for a wedge between ex ante (appraisal) estimates and ex post estimates of rates of returns. They estimate that, on average, projects take considerably more time to implement than expected at appraisal: six years rather than four years.

⁷ In Mexico's well-known PROGRESA program, payment records from an evaluation sample showed that 27 percent of the eligible population had not received benefits after almost two years of program operation, possibly a result of delays in setting up the program's management information system (Rawlings and Rubio 2005). In Ecuador's *Bono de Desarrollo Humano*, the lists of the beneficiaries who had been allocated the transfer through a lottery did not reach program operators and so about 30 percent of them did not take up the program (Schady and Araujo 2008).

⁸ Chin (2005) makes a one-year adjustment in her evaluation of Operation Blackboard in India. Although the Indian government allocated and disbursed funds for the program for the first time in fiscal year 1987, not all schools received program resources until the following school year. In addition to the delay in implementation, Chin also finds that only one-quarter and one-half of the project teachers were sent to one-teacher schools while the remaining project teachers were used in ways the central government had not intended. Apparently, the state and local governments had exercised their discretion in the use of the OB teachers.

⁹ In two programs that we know, administrative records at the individual level were maintained at local program offices, not at a central program office, and local recordkeeping varied in quality and form (for example, some records were computerized and some were not), so that a major effort was required to collect and check records during the evaluations.

¹⁰ Information campaigns for programs that attempt to improve primary school quality or to enhance child nutrition through primary school feeding programs in a context in which virtually all primary-school-age children are already enrolled would seem less relevant than such campaigns as part of a new program to improve preschool child development where there had previously been no preschool programs.

¹¹ The authors estimate the impact of atypically low rainfall levels by including a year's delay because the food shortages would be apparent only one year after the drought, but before the next harvest was ready.

¹² To estimate these longer-run impacts, Alderman, Hoddinott and Kinsey (2006) combine data on children's ages with information on the duration of the civil war and the episodes of drought used in their analysis. They undertook a new household survey to trace children measured in earlier surveys.

References

- Alderman, Harold, John Hoddinott, and William Kinsey. 2006. "Long Term Consequences of Early Childhood Malnutrition." *Oxford Economic Papers* 58 (3): 450–74.
- Angrist, Joshua D., Eric Bettinger, Erik Bloom, Elizabeth M. King, and Michael Kremer. 2002. "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment." *American Economic Review* 92 (5): 1535–58.
- Angrist, Joshua, Eric Bettinger, and Michael Kremer. 2004. "Long-term consequences of secondary school vouchers: Evidence from administrative records in Colombia." National Bureau of Economic Research Working Paper No. 10713, August.
- Angrist, Joshua D., and Victor Chaim Lavy. 2001 "Does Teacher Training Affect Pupil Learning?: Evidence from Matched Comparisons in Jerusalem Public Schools." *Journal of Labor Economics* 19 (2): 343–69.
- Armeecin, Graeme, Jere R. Behrman, Paulita Duazo, Sharon Ghuman, Socorro Gultiano, Elizabeth M. King, and Nannette Lee. 2006. "Early Childhood Development through an Integrated Program: Evidence from the Philippines." Policy Research Working Paper 3922, The World Bank, Washington, DC.
- Banerjee, Abhijit, Shawn Cole, Esther Duflo, and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics*. 122 (3): 1235-1264.
- Banerjee, Abhijit V., Rukmini Banerji, Esther Duflo, Rachel Glennerster, and Stuti Khemani. 2008. "Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India." Policy Research Working Paper 4584, The World Bank, Washington,

DC.

- Behrman, Jere R., and John Hoddinott. 2005. "Programme Evaluation with Unobserved Heterogeneity and Selective Implementation: The Mexican 'Progresa' Impact on Child Nutrition." *Oxford Bulletin of Economics and Statistics* 67 (4): 547–69.
- Behrman, Jere R., Yingmei Cheng, and Petra E. Todd. 2004. "Evaluating Preschool Programs When Length of Exposure to the Program Varies: A Nonparametric Approach." *Review of Economics and Statistics* 86 (1): 108–32.
- Behrman, Jere R., Piyali Sengupta, and Petra Todd. 2005. "Progressing through PROGRESA: An Impact Assessment of Mexico's School Subsidy Experiment." *Economic Development and Cultural Change* 54 (1): 237-275.
- Chen, Shaohua, Ren Mu and Martin Ravallion. 2008. "Are there lasting impacts of aid to poor areas?" Policy Research Working Paper 4084, The World Bank, Washington, DC.
- Chin, Aimee. 2005. "Can Redistributing Teachers across Schools Raise Educational Attainment?: Evidence from Operation Blackboard in India." *Journal of Development Economics* 78 (2): 384–405.
- Cobb-Clark, Deborah A., and Thomas Crossley. 2003. "Econometrics for Evaluations: An Introduction to Recent Developments." *Economic Record* 79 (247): 491–511.
- Duflo, Esther. 2001. "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment." *American Economic Review* 91 (4): 795–813.
- Duflo, Esther. 2004. "Scaling Up and Evaluation," in F. Bourguignon and B. Pleskovic (eds.), *Annual World Bank Conference on Development Economics: Accelerating Development*. Washington, DC: The World Bank.

- Engle, Patrice L., Maureen M. Black, Jere R. Behrman, Meena Cabral de Mello, Paul J. Gertler, Lydia Kapiriri, Reynaldo Martorell, Mary Eming Young, and the International Child Development Steering Group. 2007. "Strategies to Avoid the Loss of Developmental Potential in More Than 200 Million Children in the Developing World." *Lancet* 369 (January): 229–242.
- Foster, Andrew D., and Mark R. Rosenzweig. 1995. "Learning by Doing and Learning from Others: Human Capital and Technical Change in Agriculture." *Journal of Political Economy* 103 (6): 1176–1209.
- Frankenberg, Elizabeth, Wayan Suriastini, and Duncan Thomas. 2005. "Can Expanding Access to Basic Health Care Improve Children's Health Status?: Lessons from Indonesia's 'Midwife in the Village' Programme." *Population Studies* 59 (1): 5–19.
- Galasso, Emanuela, and Jeffrey Yau. 2006. "Learning through Monitoring: Lessons from a Large-Scale Nutrition Program in Madagascar," Policy Research Working Paper 4058, The World Bank, Washington, DC.
- Garces, Eliana, Duncan Thomas and Janet Currie. 2002. "Longer-Term Effects of Head Start." *American Economic Review* 92 (4): 999-1012.
- Gertler, Paul J. 2004. "Do Conditional Cash Transfers Improve Child Health? Evidence from Progresa's Control Randomized Experiment." *American Economic Review* 94 (2): 336–41.
- Grantham-McGregor, Sally, Yin Bun Cheung, Santiago Cueto, Paul Glewwe, Linda Richter, Barbara Strupp, and the International Child Development Steering Group. 2007. "Developmental potential in the first 5 years for children in developing countries," *Lancet* 369 (9555, January): 60-70
- Grossman, Jean Baldwin. 1994. "Evaluating Social Policies: Principles and U.S. Experience."

- World Bank Research Observer* 9 (2): 159-180.
- Heckman, James J., and Jeffrey A. Smith. 1995. "Assessing the Case for Social Experiments." *Journal of Economic Perspectives* 9 (2, Spring): 85-110.
- Heckman, James J., R. J. Lalonde, and Jeffrey A. Smith. 1999. "The economics and econometrics of active labor market programs," in Orley Ashenfelter and David Card (eds.), *Handbook of Labor Economics, Volume III*. Amsterdam: North-Holland.
- Hoddinott, John, and William Kinsey. 2001. "Child Growth in the Time of Drought." *Oxford Bulletin of Economics and Statistics* 63 (4): 409–36.
- Jimenez, Emmanuel and Yasuyuki Sawada. 1999. "Do Community-Managed Schools Work? An Evaluation of El Salvador's EDUCO Program." *World Bank Economic Review* 13 (3): 415-441.
- King, Elizabeth M., Peter F. Orazem, and Darin Wohlgemuth. 1999. "Central Mandates and Local Incentives: Colombia's Targeted Voucher Program," *World Bank Economic Review* 13 (3): 467-491
- King, Elizabeth M., and Berk Özler. 1998. "What's Decentralization Got to Do with Learning?: The Case of Nicaragua's School Autonomy Reform." Working Paper on Impact Evaluation of Education Reforms 9 (June), Development Research Group, World Bank, Washington, DC.
- Kohler, Hans-Peter, Jere Behrman, and Susan C. Watkins. 2001. "The Density of Social Networks and Fertility Decisions: Evidence from South Nyanza District, Kenya." *Demography* 38(1): 43-58.
- Manski, Charles F. 1996. "Learning about Treatment Effects from Experiments with Random Assignment of Treatments." *Journal of Human Resources* 31 (4): 709–33.

- Miguel, Edward, and Michael Kremer. 2003. "Social Networks and Learning about Health in Kenya," National Bureau of Economic Research Working Paper and Center for Global Development, manuscript, July 2003.
- Miguel, Edward, and Michael Kremer. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica* 72(1): 159-217.
- Mu, Ren, and Dominique van de Walle. 2007. "Rural roads and poor area development in Vietnam," Policy Research Working Paper Series, The World Bank.
- Munshi, Kaivan. 2004. "Social Learning in a Heterogeneous Population: Technology Diffusion in the Indian Green Revolution." *Journal of Development Economics* 73 (1): 185–213.
- Newman, John L., Menno Pradhan, Laura B. Rawlings, Geert Ridder, Ramiro Coa, and Jose Luis Evia. 2002. "An impact evaluation of education, health, and water supply investments by the Bolivian Social Investment Fund," *World Bank Economic Review* 16(2): 241-274.
- Pitt, Mark M., Mark R. Rosenzweig, and Donna M. Gibbons, 1993, "The Determinants and Consequences of the Placement of Government Programs in Indonesia," *The World Bank Economic Review* 7:3 (September), 319-348.
- Pohl, Gerhard, and Dubravko Mihaljek. 1992. "Project Evaluation and Uncertainty in Practice: A Statistical Analysis of Rate-of-Return Divergences of 1,015 World Bank Projects." *World Bank Economic Review* 6 (2): 255–77.
- Ravallion, Martin. 2001. "The Mystery of the Vanishing Benefits: An Introduction to Impact Evaluation." *World Bank Economic Review* 15(1): 115-140.
- Ravallion, Martin. 2002. "Are the Poor Protected from Budget Cuts? Evidence for Argentina," *Journal of Applied Economics* 5(1): 95-121.
- Ravallion, Martin, Emanuela Galasso, Teodoro Lazo, and Ernesto Philipp. 2005. "What Can Ex-

- Participants Reveal about a Program's Impact?" *Journal of Human Resources* 40(1): 208-230.
- Rawlings, Laura B. and Gloria M. Rubio. 2005. "Evaluating the Impact of Conditional Cash Transfer Programs," *World Bank Research Observer* 20(1): 29-55.
- Rivarola, Magdalena, and Bruce Fuller. 1999. "Nicaragua's Experiment to Decentralize Schools: Contrasting Views of Parents, Teachers, and Directors." *Comparative Education Review* 43 (4): 489-521.
- Rouse, Cecilia Elena, and Alan B. Krueger. 2004. "Putting Computerized Instruction to the Test: A Randomized Evaluation of a 'Scientifically Based' Reading Program." *Economics of Education Review* 23 (4): 323-38.
- Schady, Norbert, and Maria Caridad Araujo. 2008. "Cash transfers, conditions, and school enrollment in Ecuador," *Economía* 7(2), forthcoming.
- Schultz, T. Paul. 2004. "School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program." *Journal of Development Economics* 74:2 (June), 199-250.
- Todd, Petra and Kenneth I. Wolpin, 2007, "Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility: Assessing the Impact of a School Subsidy Program in Mexico." *American Economic Review* 96(5): 1384-1417.
- Walker, Susan P., Susan M. Chang, Christine A. Powell, and Sally M. Grantham-McGregor. 2005. "Effects of Early Childhood Psychosocial Stimulation and Nutritional Supplementation on Cognition and Education in Growth-Stunted Jamaican Children: Prospective Cohort Study." *Lancet* 366 (9499): 1804-07.
- Westoff, Charles F. and German Rodriguez. 1995. "The Mass Media and Family Planning in Kenya," *International Family Planning Perspectives*. 21(1): 26-31.
- World Bank. 2005. "Improving the World Bank's Development Effectiveness: What Does

Evaluation Show?" Evaluation report, Operations Evaluation Department, World Bank, Washington, DC.

Appendix Table 1 Examples of Evaluations That Consider Timing Issues and Duration of Program Exposure in Estimating Program Impact

Studies	Country	Intervention	Sources of variation in timing and duration of program exposure					
			Implementation lags	Short-run and long-run outcomes	Learning by beneficiaries	Learning & use by beneficiaries	Cohort effects	Cohort interacted with duration of exposure
Angrist et al. (2002); Angrist et al. (2004)	Colombia	School voucher program for secondary level		x		x		
Armecin et al. (2007)	Philippines	Comprehensive early childhood development program (ECD)	x				x	x
Banerjee et al. (2007)	India	Balsakhi school remedial program & computer-assisted learning program		x				x
Behrman & Hoddinott (2005)	Mexico	PROGRESA nutrition intervention				x		
Behrman et al. (2004)	Bolivia	PIDI preschool program					x	x
Behrman et al. (2005); Schultz (2004)	Mexico	PROGRESA education intervention					x	
Chin (2005)	India	Operation Blackboard: additional teachers per school	x				x	x
Duflo (2001)	Indonesia	School construction program		x			x	x
Foster & Rosenzweig (1995)	India	Green Revolution: new seed varieties				x		
Frankenberg et al. (2005)	Indonesia	Midwife in the Village program					x	x
Galasso & Yau (2006)	Madagascar	SEECALINE child nutrition program	x		x	x		x
Garces et al. (2002)	United States	Head Start program: ECD		x				
Hoddinott & Kinsey (2001); Alderman, Hoddinott & Kinsey (2006)	Zimbabwe	Drought shocks; civil war		x			x	x
Jimenez & Sawada (1999)	El Salvador	EDUCO schools: community participation			x			
Gentler (2004)	Mexico	PROGRESA health & nutrition services					x	x
King & Ozler (1998); Rivarola & Fuller (1999)	Nicaragua	School autonomy reform			x			
Miguel & Kremer (2003, 2004)	Kenya	School-based deworming program				x		
Mu & van de Walle (2007)	Vietnam	Rural roads rehabilitation project		x		x		
Munshi (2004)	India	Green Revolution: new seed varieties				x		
Rouse & Krueger (2004)	United States	Fast ForWord Program: computer assisted learning						x
Walker et al. (2005)	Jamaica	Nutrition supplementation		x			x	

Note: Review articles on early childhood development programs (e.g., Engle et al. 2007 and Grantham-McGregor et al. 2007) cover a long list of studies that we mention in the text but are not listed in this table; many of those studies examine age-specific effects and some examine short- and long-run impacts.