World Bank Reprint Series: Number 162

Graham Pyatt, Chau-Nan Chen, and John Fei

The Distribution of Income by Factor Components

Reprinted with permission from *The Quarterly Journal of Economics*, vol. 94 (1980), pp. 451–73.

THE DISTRIBUTION OF INCOME BY FACTOR COMPONENTS*

GRAHAM PYATT
CHAU-NAN CHEN
JOHN FEI

The paper provides a rigorous and exact formulation of the relationship between the Gini measure of inequality in total income across families, and corresponding measures of inequality in such components of total income as wages, transfer income, etc. It is shown that serious problems of bias arise when individual family data are not available and when data on averages for families grouped by the size of total income are used instead. These problems are illustrated with reference to data for Taiwan, 1964 to 1976.

I. INTRODUCTION

In a recent paper, "Growth and the Family Distribution of Income by Factor Components," John Fei, Gustav Ranis, and Shirley Kuo [1978] set out a methodology for the statistical decomposition of family income inequality in terms of such components of family incomes as wages, income from agricultural activities, property income, etc. Their analysis translates the statistical decomposition of income inequality into meaningful economic components of growth and inequality, to yield a mapping from the statistical magnitudes into the conceptual framework of economic theory. A methodology that allows this to be done is clearly of great interest for any attempt to understand the process of economic development. The Fei, Ranis, and Kuo (F-R-K) paper also provides a numerical illustration of the methodology with reference to Taiwan for the period 1964 to 1972.

The measure of inequality used is the Gini coefficient. Links with the Lorenz curve make this coefficient an attractive statistic for the purpose at hand, and recent contributions to the literature on the Gini coefficient have added to our understanding of its properties and characteristics. The present paper is directed toward furthering this understanding in the context of the F-R-K application, with particular emphasis on the issues that arise when data for individual families are not available and the decomposition of family income has to be

* We are indebted to Gustav Ranis and John H. Duloy for comments on an early draft and for their contributions to a dialogue, over an extended period, which have led to the present paper. We are also grateful to T. W. Tsao for his most valuable contribution in preparing the data reported in this paper. It is to be noted that the views expressed in this paper do not necessarily reflect those of the World Bank or its affiliates.

The Quarterly Journal of Economics, November 1980 0033-5533/80/0095-0151$02.30

based on grouped data of the type that are readily accessible in published form.

An aspect of the problems raised by grouped data has previously been discussed in the literature. Specifically, Gastwirth [1972] analyzed the errors involved when a Gini coefficient is calculated from grouped data as opposed to individual observations. The present analysis shows that this problem is only a minor facet of the issues that arise when grouped data are used as a basis for the decomposition of total income into additive factor components. The nature of the broader issues has been the subject of exchanges between the present authors for some time. Using results in an early round of these exchanges, Gary Fields [1979] examined the problems using some 1967–1968 data for Colombia.[1] He writes: "[The evidence] suggests that for this particular decomposition problem with this particular type of grouped data, the option of doing nothing at all rather than using what imperfect data we have deserves serious consideration." [2] This is a strong conclusion, which requires a perspective. This paper attempts to provide one, first from a conceptual point of view, and then in terms of empirical results for Taiwan, using data for individual families that have recently become available.

In discussing the problems that arise when a Gini coefficient is decomposed into additive factor components, the present analysis builds on a statistic known as the concentration ratio. Some authors, especially in the statistical literature, use the term "concentration ratio" synonymously with that of a Gini coefficient. However, in the definition of the former, set out in Section II below, the concentration ratio is a more general concept, with respect to which the Gini coefficient is a special case. So, too, is the pseudo-Gini coefficient, which plays such an important part in the F-R-K methodology. However, true Ginis and pseudo-Ginis are not the same thing, and a clarification of the difference is central to the present discussion.[3]

The next section (Section II) of this paper provides a definition of the concentration ratio for a variable z with respect to some other variable t, when both z and t are observed for each household. The Gini coefficient for z is then defined as the concentration ratio of z with respect to itself. Based on this we are able to show that, for both grouped data and for individual family data when available, the Gini coefficient of inequality for total family incomes can be decomposed

1. It is stated incorrectly in Fields [1979] that the fundamental result, which is given as equation (36) below, is derived in Fei, Ranis, and Kuo [1978]. See Fields [1979], footnote 2, p. 328.
2. Fields [1979], p. 334.
3. See Section III below.

exactly as a sum over all types of factor incomes, of the cross products of two terms, namely, (i) the share of each factor in total income, and (ii) the concentration ratio for the distribution of each type of factor income with respect to that for total income. Moreover, when individual family data are available, the concentration ratio for each type of factor income can be expressed as the product of (a) a correlation effect depending on the respective rankings of households according to total income and income of the given factor type; and (b) the Gini coefficient for the distribution of the particular type of factor income. Hence, with individual family data, we get an exact decomposition of the Gini coefficient for income inequality into factor shares, correlation effects, and the Gini coefficient for each factor. This decomposition is illustrated for Taiwan (1964 to 1976) in Section VI below.

In the absence of individual family data, it is still possible to obtain an exact decomposition of the Gini coefficient for total income (based on grouped data) into a weighted average of the concentration ratios for each type of factor income with respect to total income. In the F-R-K paper these concentration ratios were interpreted as being (approximately) equal to the factor Gini coefficients. However, using the individual family data now available, we show in the numerical results in Section V that Gini coefficients and concentration ratios can be markedly different both in level (by a factor of 2) and trends over time (including turning points). Accordingly, our results show that, in general, it is not safe to interpret the concentration ratios for factor components of income as if they were Gini coefficients, since the factor components also depend on correlation effects. We therefore conclude that the decomposition of the total income Gini adopted by F-R-K is legitimate, but their interpretation of its components is not.

Before presenting our numerical results in Sections V and VI, a conceptual basis is established in Sections II to IV. As previously noted, the concentration ratio is defined in Section II, and exact decomposition by additive factor components is explored for the case where individual data are available. Section III then presents the exact decomposition applicable to grouped data. Problems of accuracy and interpretation are then discussed in Section IV. It is shown that the accuracy problem discussed in Gastwirth [1972] is essentially irrelevant to the main issue, which is the extent to which factor concentration ratios can be interpreted as Gini coefficients. The theoretical discussion is then completed: first by showing that the concentration ratios always underestimate the corresponding factor Ginis, and then

by deriving necessary and sufficient conditions for this bias to be zero.

The results on exact decomposition that are obtained in this paper are largely anticipated by the work of Rao [1969], which, in turn, has antecedents in the literature. It seems, however, that the earlier statistical literature has not focused on the problems of empirical analysis, and especially those which arise from grouped data and the use of secondary source material. Insofar as the pioneering work of Fei, Ranis, and Kuo is being reproduced in contemporary studies,[4] the analysis and results that are presented here may caution other investigators about the limitations that are inherent in working with secondary, as opposed to primary, data sources. In addition, our results show that translation from the exact Rao decomposition of the Gini for total income into the framework of economic theory is less straightforward than Fei, Ranis, and Kuo have assumed.

II. Gini Coefficients and Concentration Ratios

Before coming to the problems that arise when only grouped data are available, and since the new results in Section VI below are based on individual family data, it is useful to start our discussion with general definitions of Gini coefficients and concentration ratios that can be applied in the various situations to be encountered subsequently.

The starting point is to assume that there are n families and that two variables, z_i and t_i, are observed for each ($i = 1,2,\ldots,n$). Families are to be ranked according to t_i: the ranking of the ith family will be denoted $r(t_i)$ with the convention that $r(t_i) = 1$ for the family for which t_i is smallest, and $r(t_i) = n$ for the family for which t_i is largest. If two or more families have the same value for t_i, they are each to be given the average of the ranks that they would get if there were an infinitesimal difference between them. With these conventions the average of all ranks $r(t_i)$ is given by

$$\bar{r}(t_i) = \frac{1}{n}\sum_{i=1}^{n} r(t_i) = \frac{(n+1)}{2} = \bar{r}. \tag{1}$$

The average rank is therefore independent of the ranking criterion t_i, which is adopted.

Next, it is assumed that the average value of z_i is positive, i.e.,

4. See, for example, Ayub [1977] and Mangahas and Gamboa [1976].

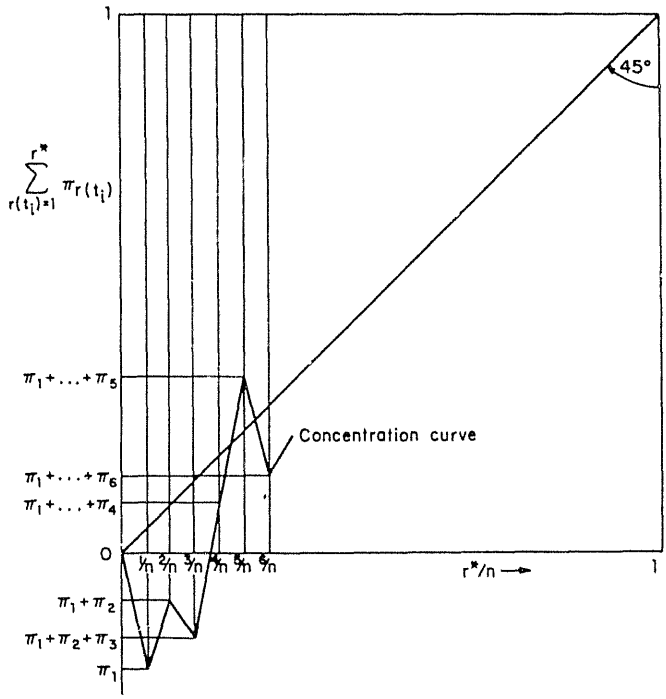

FIGURE I
Schematic Construction of a Concentration Curve

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i \geq 0. \tag{2}$$

It can be noted that this does not require that z_i be positive for all i.

Given $\bar{z}$, we can also define

$$\pi_i(z) = z_i/n\bar{z} \tag{3}$$

for each household. From (2) and (3) it is apparent that the sum over i of π_i is unity.

The concentration ratio of z_i with respect to t_i is obtained from a concentration curve. This curve graphs cumulative values of π_i against actual values of $r(t_i)/n$, when the cumulative sums are formed by taking families in the order indicated by $r(t_i)$. Figure I illustrates a concentration curve. The concentration ratio of z with respect to t, denoted as $C(z/t)$, is now defined as one *minus* twice the area under the concentration curve. It can be noted from the figure and the def-

inition of the curve that the latter need not be monotonic: a concentration ratio can be negative, or it can have a value greater than one if some z's are negative.

It can be seen from inspection of Figure I that the area under the concentration curve can be obtained as the sum of the areas of n vertical strips, each with width $(1/n)$, i.e., as[5]

$$(4) \quad \frac{1}{2}\frac{1}{n}\left(\pi_1 + \sum_{j=2}^{n}\left(\sum_{l=1}^{j-1}\pi_l + \sum_{l=1}^{j}\pi_l\right)\right) = \frac{1}{2}\frac{1}{n}\sum_{j=1}^{n}\pi_j\{1 + 2(n-j)\},$$

where the subscripts i refer to the rank of the family, i.e., to $r(t_i)$. From (4) the concentration ratio is

$$(5) \quad C\left(\frac{z}{t}\right) = 1 - \frac{1}{n}\sum_{j=1}^{n}\pi_j\{1 + 2(n-j)\},$$

which can be reduced to

$$(6) \quad C\left(\frac{z}{t}\right) = \frac{2}{n}\sum_{i=1}^{n}\pi_i\{r(t_i) - \bar{r}\} = 2\,\mathrm{cov}(\pi(z),r(t)),$$

where $\mathrm{cov}(a,b)$ is the covariance between variables a and b. Hence, from (3),

$$(7) \quad C(z/t) = 2\,\mathrm{cov}((z/n\bar{z}),r(t)) = (2/n\bar{z})\,\mathrm{cov}(z,r(t)).$$

In this framework a Gini coefficient is the concentration ratio of a variable with respect to itself, since in this case the concentration curve in Figure I is now the Lorenz curve. If y_i is the total income of the ith family, the Gini coefficient for family incomes is

$$(8) \quad G(y) = C(y/y) = (2/n\bar{y})\,\mathrm{cov}(y,r(y)).$$

The F-R-K study is concerned with the decomposition of $G(y)$ in terms of the factor components of y_i. If x_{ik} is the contribution of factor component k (e.g., wages) to the total income of the family y_i, then

$$(9) \quad y_i = \sum_{k=1}^{m} x_{ik} \quad \text{for } i = 1, \ldots, n,$$

where total income is made up of m factor components ($k = 1, \ldots, m$); and

$$(10) \quad \bar{y} = \sum_{k=1}^{m}\bar{x}_k,$$

5. When the concentration curve falls below the abscissa (i.e., is negative), then the area under the curve is also negative.

where $\bar{x}_k$ is the average, over all families, of x_{ik}

Substituting (9) in (8) now gives

$$G(y) = \left(\frac{2}{n\bar{y}}\right) \operatorname{cov}\left[\sum_k x_{ik}, r(y)\right] = \sum_{k=1}^{m} \phi_k C\left(\frac{x_k}{y}\right), \tag{11}$$

where

$$\phi_k = \bar{x}_k/\bar{y}. \tag{12}$$

From (10) and (12) it follows that the sum, over k, of ϕ_k is unity, so that result (11) expresses the Gini coefficient for total income y_i as the weighted average of the concentration ratios for factor incomes with respect to total incomes.

Result (11) is due to Rao [1969]. It provides an exact decomposition of $G(y)$, in which the weights ϕ_k are the shares of each factor type in total incomes. Condition (2) implies that these weights are nonnegative. However, the concentration ratios that appear in (11) can be negative as previously noted.

To examine the difference between factor concentration ratios and factor Ginis, it follows from (8) that the latter are given by

$$G(x_k) = (2/n\bar{x}_k) \operatorname{cov}(x_k, r(x_k)). \tag{13}$$

Using (7), this gives

$$\frac{C(x_k/y)}{G(x_k)} = \frac{\operatorname{cov}(x_k, r(y))}{\operatorname{cov}(x_k, r(x_k))} = R(y, x_k). \tag{14}$$

The ratio $R(y,x_k)$ will be referred to as a rank correlation ratio. It is not a rank correlation coefficient, but inspection of (14) suggests that this ratio will be unity only if

$$r(y) = r(x_k), \tag{15}$$

i.e., only if families have the same ranking with respect to factor income k as they have with respect to total income.

It can also be shown that unity is an upper bound for $R(y,x_k)$. This can be argued with respect to concentration curves. To obtain $G(x_k)$, families are ranked by the size of x_{ik}. Hence the rth point on the concentration curve depends on the r families for which x_{ik} is smallest. Some other ordering is involved in obtaining $C(x_k/y)$, namely, the ordering of families according to the size of y_i. The sum of x_{ik} cannot be smaller for the first r families according to this other ordering. Hence the concentration curve that yields $C(x_k/y)$ cannot lie below what yields $G(x_k)$, and therefore

$$C(x_k/y)/G(x_k) = R(x_k,y) \leq 1. \tag{16}$$

More generally, the argument implies that the concentration ratio of a variable cannot exceed its Gini coefficient, and hence that

$$\text{cov}(z,r(z)) \geq \text{cov}(z,r(t)) \quad \text{for all } t. \tag{17}$$

III. Grouped Data

For present purposes the discussion of grouped data can be restricted to instances where grouping results in an equal number of families in each group, irrespective of the grouping criterion that is used. This covers two special cases of interest, namely, the decile grouping of families as employed in F-R-K, and the case of only one family in each group, i.e., the ungrouped data case.

If families are grouped by size of income, then information is restricted to the average income within each group. Notation $\hat{y}_i$ will be used for the average income in the group within which y_i falls. Thus, $\hat{y}_i$ is defined for each family and can be used to define $r(\hat{y}_i)$ in each case.[6] Obviously $\hat{y}_i$ and $r(\hat{y}_i)$ will be the same for all families within a given group. Thus, if there are s groups of families, the concentration curve will be a connected series of s linear segments.[7] It will yield the Gini coefficient for grouped income data:

$$G(\hat{y}) = C(\hat{y}/\hat{y}) = (2/n\bar{y})\,\text{cov}(\hat{y},r(\hat{y})). \tag{18}$$

We can now note that[8]

$$\text{cov}(\hat{y},r(\hat{y})) = \text{cov}(y,r(\hat{y}_i)). \tag{19}$$

Hence, from (18),

$$\text{cov}(\hat{y},r(\hat{y})) \leqslant \text{cov}(y,r(y))$$

or

$$G(\hat{y}) \leqslant G(y), \tag{20}$$

6. Recall that in our earlier definition of $r(t_i)$ we adopted the convention that, if t_i is the same for two or more families, then for each of these families $r(t_i)$ takes on the average of the values it would have if there was an infinitesimal difference in t_i between the families.

7. This is equivalent to saying that the concentration curve is obtained by joining the origin and s of the n successive points on the concentration curve by straight lines. In the notation of Figure I, these s points correspond to the actual points at which $r^* = n/s, 2n/s, \ldots, (s-1)n/s, n$.

8. This follows directly from $\text{cov}(y,r(\hat{y})) = \text{cov}(\hat{y},r(\hat{y})) + \text{cov}(y-\hat{y},r(\hat{y}))$ and from noting that the second term on the right hand side must be zero. This is because this second term depends on covariances within groups for which one of the variables, viz. $r(\hat{y}_i)$, is in fact constant.

i.e., the Gini coefficient calculated from grouped data must always underestimate the true Gini.

The size of this underestimation has been explored by Gastwirth [1972], among others. His analysis and our own empirical results suggest that it is of marginal importance. It is obviously an element in the accuracy of the F-R-K statistics and is attributable to the use in that investigation of grouped data. However, our main concerns here are with problems that arise from the interaction of decomposition methods and grouped data. From this perspective the bias implied by (20) is not essential, as we shall show in Section IV.

When information is restricted to data for families grouped by size of total income, we may also know the average value of each factor income within each of the total income groupings. Let $\bar{x}_{ik}$ denote the average value of type k income for households in the total income group that includes household i. A typical element of $\bar{x}_{ik}$ is therefore the average wage income of families that fall in, say, the third size group according to total income. Then the sum over k of $\bar{x}_{ik}$ is the average of the combined income from all sources of families in a particular size group for total income, i.e.,

$$\hat{y}_i = \sum_{k=1}^{m} \bar{x}_{ik}. \tag{21}$$

Making this substitution in (19) yields

$$G(\hat{y}) = \sum_{k=1}^{m} \phi_k C\left(\frac{\bar{x}_k}{\hat{y}}\right). \tag{22}$$

This is an exact decomposition of the Gini coefficient for incomes $\hat{y}_i$, which is analogous to that previously obtained for ungrouped data as equation (11). Since (22) corresponds to the case where the only data available are grouped by income levels y_i, it follows that such grouping, of itself, does not preclude an exact decomposition.

Result (22) is the basic decomposition equation employed by F-R-K.[9] The concentration ratios $C(\bar{x}_k, \hat{y})$, which appear in it, are referred to in their analysis as pseudo-Ginis. However, we prefer to use here the longer but more precise name of "factor concentration ratios for data grouped by total income level."

IV. Problems of Accuracy and Interpretation

Result (22) in the previous section provides an exact decomposition of the Gini coefficient for total income when data are grouped

9. See Fei, Ranis, and Kuo [1978], Appendix equation (A2.11).

into size classes defined by the level of total income. This can be written as

$$G(y) = \sum_{k=1}^{m} \phi_k C\left(\frac{\tilde{x}_k}{\hat{y}}\right) + \{G(y) - G(\hat{y})\}, \tag{23}$$

where the term in braces on the right-hand side of (23) must be positive in view of the inequality (20). We have previously suggested that this term, which is discussed by Gastwirth [1972], is not large and has limited relevance to the problems that concern us. The next step in the present argument is to justify this position, given that our main concern is the extent to which it is legitimate to interpret the concentration ratios $C(\tilde{x}_k/\hat{y})$ as Gini coefficients for factor incomes.

To develop the argument, let $G(x_k)$ be the Gini coefficient of inequality among individual families for the kth type of factor income, e.g., property income. Clearly, if individual family data on property income were available, we could group families (into deciles, for example) according to the size of their property income. The individual data could now be discarded, and a Gini coefficient $G(\hat{x}_k)$ could be calculated from the grouped data, on the assumption that each family has an amount of property income equal to the average property income among families that fall within the same size class (decile) according to the level of property income. Separate calculations for each type of factor income ($k = 1,2, \ldots, m$) would now yield a series of statistics $G(x_k)$ and $G(\hat{x}_k)$, which, in view of result (20), would have the property

$$G(\hat{x}_k) \leq G(x_k) \quad \text{for } k = 1, \ldots, m. \tag{24}$$

It should be noted that calculating each $G(\hat{x}_k)$ requires a reranking of households according to the size of x_{ik}, and then grouping of the ranked data. These groupings will imply that for a given k there are no overlaps in the size of x_{ik} between groups. To compute the set of statistics $G(\hat{x}_k)$ for $k = 1,2, \ldots, m$ requires m distinct rerankings of the individual data. Clearly, such rerankings are impossible if individual data are not available. Hence, access to primary family data is necessary if the statistics $G(\hat{x}_k)$ are to be computed.

Simple manipulation of equation (23) yields the result

$$G(y) = \sum_{k=1}^{m} \phi_k G(x_k) + \sum_{k=1}^{m} \phi_k \epsilon_k + e, \tag{25}$$

where

$$\epsilon_k = C(\tilde{x}_k/\hat{y}) - G(\hat{x}_k) \quad \text{for } k = 1, \ldots, m \tag{26}$$

and

$$e = \{G(y) - G(\hat{y})\} - \sum_{k=1}^{m} \phi_k \{G(x_k) - G(\hat{x}_k)\}. \tag{27}$$

Hence, from (25), there are two types of error involved in assuming that the Gini coefficient for total income can be expressed as a weighted sum of the Gini coefficients for the separate types of factor income. The first type of error, ϵ_k for $k = 1, \ldots, m$, is the major source of error and is discussed in detail below. The second type of error e is less important for three reasons.

Components of the error e arise from using grouped data to calculate Gini coefficients. If the size of a variable z is the criterion used to group families, then the error in calculating the Gini coefficient for z from grouped data can be expressed as[10]

$$G(z) - G(\hat{z}) = \sum_{r=1}^{s} p_r \pi_r g_r, \tag{28}$$

where it is assumed that the population has been divided into s groups; p_r is the population proportion in the rth group; π_r is the share of group r in the total of z for the whole population; and g_r is the Gini coefficient of inequality for the distribution of z within group r. Hence, if the population is grouped into deciles according to the size of z, we have $s = 10$ and $p_r = 1/10$ for all r. Similarly, in this example, π_r will have an average value of $1/10$ while g_r will be bounded between 0 and 1 if z is a nonnegative variable. Hence the error term (28) has an order of magnitude of $1/10$. In fact, Gastwirth [1972] is able to impose much tighter bounds on this error term. These provide the first reason for regarding the term e in (27) as negligible.

The second reason for regarding e as negligible is that its determination as in (27) involves compensating errors. Each of the terms in braces in (27) must be positive, while the sum of the weights ϕ_k is unity. Hence, if decile grouping is always employed, the error e is the difference between two nonnegative terms each of the same order of magnitude. Hence it may be quite reasonable to ignore e in numerical work.

A third reason for ignoring e is that the problem we are interested in can be explored without reference to it. If available data are restricted to a grouped distribution according to total family income, then $G(y)$ is unknown, and analysis must necessarily be restricted to the decomposition of $G(\hat{y})$. Hence, taking (22) as the starting point, we can easily show that

10. See, for example, Pyatt [1976]

$$G(\hat{y}) = \sum_{k=1}^{m} \phi_k G(\hat{x}_k) + \sum_{k=1}^{m} \phi_k \epsilon_k, \tag{29}$$

i.e., a decomposition of $G(\hat{y})$ that does not involve errors of the type discussed by Gastwirth.

Our analysis in Section V will be based on equation (29). In terms of it, the interpretation problem using the F-R-K methodology is whether the concentration ratios $C(\bar{x}_k/\hat{y})$ are good approximations of the Gini coefficients $G(\hat{x}_k)$, i.e., whether the error terms ϵ_k are negligible. To explore this, we shall find it useful to express each ϵ_k as the sum of two terms, both of which are nonpositive. Beyond this, necessary and sufficient conditions can be obtained for each component of ϵ_k to be zero. Hence we can obtain conditions under which there is no interpretation problem.

At this stage our analysis requires a new concept, denoted $G(\bar{x}_k)$, which can be called a quasi-factor Gini. Given this, we want to prove that

$$C(\bar{x}_k/\hat{y}) \leqslant G(\bar{x}_k) \leqslant G(\hat{x}_k), \tag{30}$$

i.e. that the quasi-factor Gini lies between the concentration ratio $C(\bar{x}_k/\hat{y})$ and the Gini coefficient $G(\hat{x}_k)$. If (30) is valid, then it follows that

$$\epsilon_k = \epsilon_{1k} + \epsilon_{2k}, \tag{31}$$

where

$$\epsilon_{1k} = C(\bar{x}_k/\hat{y}) - G(\bar{x}_k) \leqslant 0 \tag{32}$$

and

$$\epsilon_{2k} = G(\bar{x}_k) - G(\bar{x}_k) \leqslant 0. \tag{33}$$

The variables $\bar{x}_{ik}$ have already been defined in Section II. They are defined for each k by grouping households according to the size of total income y_i, and then calculating the average value of x_{ik} within each of these groups. The groupings are therefore independent of k. They are the same grouping or clustering of families as is involved in computing the $\hat{y}_i$'s i.e., the average values of total income among groups of families defined with respect to the size of their total income. However, there is no reason why $r(\bar{x}_{ik})$ and $r(\hat{y}_i)$ should be the same: there is no reason, for example, why average transfer income should increase monotonically as we move upward between groups defined with respect to the size of y_i. Hence the rankings $r(\bar{x}_{ik})$ and $r(\hat{y}_i)$ are not necessarily the same, and results (17) and (18) are sufficient to establish the inequality (32).

This discussion of the difference between $r(\tilde{x}_{ik})$ and $r(\hat{y}_i)$ suggests the following condition:

FIRST INTERPRETATION CONDITION. A necessary and sufficient condition for $\epsilon_{1k} = 0$ is that the graph of $r(\hat{y}_i)$ *versus* $r(\tilde{x}_{ik})$ is monotonic nondecreasing.

To establish this first interpretation condition, we can argue as follows. The concentration curve that yields the pseudo-Gini $G(\tilde{x}_k)$ ranks families according to $r(\tilde{x}_{ik})$, i.e., according to the average level of x_{ik} for groups that result from the stratification of families by income levels. As previously noted, this must imply the same clustering or grouping as $r(\hat{y}_i)$. The only difference that can arise between $C(\tilde{x}/\hat{y})$ and $G(\tilde{x}_k) = C(\tilde{x}_k/\tilde{x}_k)$ is therefore in the ordering of these groups in the derivation of the concentration curve. The ordering for $G(\tilde{x}_k)$ is the one that minimizes the ordinate of the concentration curve for each value of the abscissa. Either any other ordering must result in a larger value of the ordinate, or $\tilde{x}_{ik}$ must have the same value for at least two groups. But if this latter possibility maintains, then, for two such groups, $r(\tilde{x}_{ik})$ is also the same. Hence the condition is not violated. Under the former possibility, the condition is violated and the concentration coefficient $C(\tilde{x}_k/\hat{y}_i)$ will be strictly less than the pseudo-Gini $G(\tilde{x}_k)$.[11]

To establish inequality (33), we can note that both $G(\tilde{x}_k)$ and $G(\hat{x}_k)$ are Gini coefficients calculated from grouped data. The difference between them must therefore lie in the grouping criterion, given that we assume an equal number of groups in all cases, and an equal number of households in each group. For $\tilde{x}_{ik}$, the grouping criterion is the size of y_i; while for $\hat{x}_{ik}$ the criterion is the size of x_{ik}. Hence the grouping criterion for computing $\hat{x}_{ik}$ is the one that minimizes the ordinate of the concentration curve (i.e., the Lorenz curve in this instance) for each value of the abscissa. Hence the concentration (Lorenz) curve that yields $G(\tilde{x}_k)$ can never lie below what yields $G(\hat{x}_k)$. From this, inequality (33) follows directly.

This argument leads to our second interpretation condition.

SECOND INTERPRETATION CONDITION. A necessary and sufficient condition for $\epsilon_{2k} = 0$ is that $r(\tilde{x}_{ik}) = r(\hat{x}_{ik})$ for all i.

This condition is clearly necessary, since if there is any difference in the weak ordering of families given by $r(\tilde{x}_{ik})$ and $r(\hat{x}_{ik})$, then the

11. It can be noted that, since $\tilde{x}_{ik}$ derives from groups defined by common values of $\hat{y}_i$, it is impossible to have the same value of $r(\tilde{x}_{ik})$ for two groups which have different values of $r(\tilde{x}_{ik})$.

concentration curve for $G(\bar{x}_k)$ must be above that for $G(\hat{x}_k)$ at at least one point. Since the latter concentration curve is invariably a lower bound for the former, it follows that $\epsilon_{2k} = 0$ only if there is no divergence in these weak orderings. Beyond this, if there is no difference in the weak orderings, then there will be no difference in $G(\bar{x}_k)$ and $G(\hat{x}_k)$. Hence sufficiency is established.

The two interpretation conditions given above can be combined to yield necessary and sufficient conditions for the interpretation of the concentration ratios $C(\bar{x}_k/\hat{y})$ (the F-R-K pseudo-Ginis), as Gini coefficients of form $G(\hat{x}_k)$. These are

Necessary: the graph of $r(\hat{y}_i)$ versus $r(\hat{x}_{ik})$ is monotonic nondecreasing

Sufficient: $r(\hat{y}_i) = r(\hat{x}_{ik})$ for all i.

Neither condition is as strong as the one that F-R-K discuss, namely, that the relationship between $\hat{x}_{ik}$ and $\hat{y}_i$ should be linear and have a positive slope: nonlinearity does not necessarily violate the sufficient condition.

V. The Interpretation Problem: Some Empirical Results

Given data grouped by size of total income y_i, the variables $\bar{x}_{ik}$ can be calculated, but the variables $\hat{x}_{ik}$ cannot. Hence F-R-K were able to examine the first interpretation condition that on ϵ_{1k}—but not the second. Generally, they found ϵ_{1k} to be small. However, there should be no presumption that this implies small or negligible values for ϵ_{2k}. To illustrate this point, some 1974 data for Taiwan have been analyzed with the results set out in Table I.

Table I shows that there is no interpretation problem of the first type for this data set, which refers to wage incomes. The average family wage incomes $\bar{x}_{iw}$ increase in strict monotonic fashion as we move through deciles of the total income distribution. Hence the first interpretation condition is satisfied entirely. However, the second interpretation condition is far from satisfied. If it were satisfied, then ranking into deciles by wage income level x_{iw} and total income level y_i would place all families on the main diagonal of Table I. In fact, 924 out of 5,256 families (17.6 percent) at most are to be found there. The reason why this figure has to be given as an upper limit, rather than an exact number, is because 201 families in the first two income deciles belong to the 18.1 percent of all families that have zero wage income. Whether all these 201 families should be assumed to lie on the diagonal is ambiguous. But meanwhile the fact that 18.1 percent of all

TABLE I

DISTRIBUTION OF FAMILIES AND AVERAGE WAGE INCOMES BY DECILES OF THE WAGE INCOME AND TOTAL INCOME DISTRIBUTIONS

TAIWAN: 1974

Deciles of the distribution of total income y_i	Deciles of the distribution of wage income x_{iw}											
	1st and 2nd											
	Families with zero wage income	Other families in the 2nd decile	3rd	4th	5th	6th	7th	8th	9th	10th	Total[a]	Average wage Income $\bar{x}_{iw}$[b]
1st	121	35	194	159	9						518	21.8
2nd	80	10	86	124	213	16					529	35.2
3rd	86	10	60	61	119	186	5				527	41.2
4th	109	16	45	35	44	129	146				524	44.0
5th	94	7	29	36	40	72	157	92			527	52.2
6th	81	5	37	27	30	49	96	169	19		513	57.4
7th	89	2	23	24	28	35	50	147	140		538	65.3
8th	108	5	24	20	16	19	37	61	207	29	526	69.8
9th	92	4	17	15	16	16	20	30	121	194	525	87.5
10th	91	2	11	13	17	9	15	24	40	307	529	135.8
Total[a]	951	96	526	514	532	531	526	523	527	530	5256	
Average wage income $\hat{x}_{iw}$[b]	zero	7.0	23.1	39.4	50.9	61.6	71.8	83.8	103.9	173.5		61.1

a. Totals vary between deciles because all families with the same (wage) income have been placed in the same cell of the table.
b. Units: thousands of won.

TABLE II

EXACT DECOMPOSITION OF TOTAL INCOME INEQUALITY USING FACTOR SHARES AND CONCENTRATION RATIOS, AND FACTOR GINI COEFFICIENTS FOR TAIWAN, 1964 TO 1976, BASED ON DATA GROUPED BY DECILES

	Wage income			Profit income			Agricultural income			All other income				
	Factor share ϕ_W	Concentration ratio $C(\hat{x}_W/\hat{y})$	Gini coefficient $G(\hat{x}_W)$	Factor share ϕ_π	Concentration ratio $C(\hat{x}_\pi/\hat{y})$	Gini coefficient $G(\hat{x}_\pi)$	Factor share ϕ_A	Concentration ratio $C(\hat{x}_A/\hat{y})$	Gini coefficient $G(\hat{x}_A)$	Factor share ϕ_X	Concentration ratio $C(\hat{x}_X/\hat{y})$	Gini coefficient $G(\hat{x}_X/\hat{y})$	$G(\hat{y}) = \sum_k \phi_k \times C(\hat{x}_k/\hat{y})$	$\sum_k \phi_k \times G(\hat{x}_k)$
							Agricultural families							
1964	0.230	0.120	0.523	0.063	0.263	0.451	0.635	0.413	0.469	0.072	0.532	0.984	0.345	0.517
1966	0.207	0.189	0.590	0.070	0.270	0.424	0.651	0.336	0.400	0.073	0.486	0.891	0.311	0.484
1968	0.300	0.249	0.583	0.080	0.241	0.419	0.495	0.277	0.434	0.125	0.538	0.852	0.298	0.530
1970	0.355	0.270	0.538	0.079	0.281	0.442	0.451	0.238	0.427	0.116	0.486	0.853	0.282	0.517
1972	0.416	0.269	0.497	0.077	0.257	0.434	0.395	0.235	0.456	0.112	0.497	0.858	0.280	0.516
1974	0.395	0.247	0.503	0.074	0.271	0.455	0.431	0.268	0.467	0.100	0.416	0.879	0.275	0.522
1976	0.389	0.248	0.503	0.087	0.281	0.439	0.388	0.258	0.461	0.137	0.364	0.816	0.271	0.525

						Nonagricultural families								
1964	0.514	0.259	0.479	0.107	0.446	0.678	0.021	0.077	0.823	0.358	0.431	0.705	0.337	0.588
1966	0.585	0.260	0.472	0.093	0.500	0.755	0.013	−0.312	0.874	0.309	0.424	0.705	0.326	0.576
1968	0.567	0.240	0.454	0.111	0.531	0.725	0.008	−0.194	0.878	0.314	0.425	0.730	0.327	0.574
1970	0.631	0.233	0.417	0.088	0.439	0.681	0.006	−0.071	0.880	0.275	0.382	0.746	0.290	0.534
1972	0.709	0.277	0.384	0.118	0.471	0.692	0.005	−0.007	0.902	0.168	0.281	0.803	0.299	0.493
1974	0.655	0.262	0.412	0.122	0.487	0.682	0.008	−0.025	0.900	0.215	0.346	0.767	0.305	0.525
1976	0.648	0.243	0.402	0.112	0.424	0.631	0.008	0.068	0.900	0.232	0.325	0.742	0.281	0.511
						All families								
1964	0.332	0.171	0.534	0.079	0.340	0.576	0.415	0.454	0.651	0.174	0.418	0.847	0.345	0.640
1966	0.443	0.235	0.541	0.084	0.429	0.679	0.251	0.350	0.755	0.221	0.420	0.788	0.321	0.660
1968	0.491	0.260	0.510	0.102	0.466	0.675	0.148	0.167	0.786	0.260	0.465	0.772	0.320	0.636
1970	0.535	0.268	0.486	0.085	0.390	0.620	0.161	0.121	0.752	0.220	0.433	0.792	0.291	0.608
1972	0.644	0.295	0.430	0.109	0.447	0.670	0.092	0.079	0.826	0.156	0.336	0.819	0.299	0.554
1974	0.605	0.273	0.446	0.113	0.469	0.668	0.089	0.129	0.854	0.193	0.372	0.790	0.302	0.574
1976	0.591	0.264	0.445	0.107	0.407	0.610	0.091	0.095	0.831	0.211	0.354	0.762	0.283	0.565

families have zero wage income is quite lost when families are grouped by income level. More generally, the data in Table I imply a concentration ratio (or pseudo-Gini) of 26.4 percent. The true Gini for wages using decile groups is 44.3 percent.[12] Therefore there is a major problem in interpreting pseudo-Ginis as true Ginis even when there is no interpretation problem of the first type.

Since both types of interpretation problem generally can arise, it follows that the gap between $C(\bar{x}_k/\hat{y})$ and $G(\hat{x}_k)$ must be the only measure of reliability of the F-R-K methodology in the sense that, if this gap is large, then the decomposition

$$G(\hat{y}) \doteq \sum_{k=1}^{m} \phi_k G(\hat{x}_k) \tag{34}$$

must be subject to substantial errors. Table II sets out some results for the period 1964 to 1976 on the size of the gap.

Numerically, the results in Table II are not strictly comparable to those presented in Fei, Ranis, and Kuo [1978]. There are two reasons for this. First, F-R-K used published survey results. The data in Table II use individual family survey results. Not all the original questionnaires could be recaptured for our purposes, so that results for the earlier years especially are open to doubt.[13] Second, the surveys used by F-R-K and here have changed in coverage over the years, notably in the representation of Taipei city from 1970 onward when the sample size was doubled. An adjustment for this has been made by F-R-K, while the present results refer to original, unadjusted data.

With these reservations, Table II shows some interesting and significant gaps. Since the concentration ratios $C(\bar{x}_k/\hat{y})$ must always be smaller than $G(\hat{x}_k)$ ($\epsilon_k \leqslant 0$), it follows that

$$G(\hat{y}_i) = \sum_k \phi_k C\left(\frac{\bar{x}_k}{\hat{y}}\right) \leq \sum_k \phi_k G(\hat{x}_k). \tag{35}$$

The final columns of Table II show the difference between the exact decomposition (22) and the decomposition (34), which is implied by interpreting concentration ratios as Gini coefficients. The gap is substantial, so that the F-R-K interpretation of concentration ratios (pseudo-Ginis) as true Ginis is hardly appropriate.

12. It can be noted that these numbers differ slightly from those in Tables II and III below, because the Table I data are taken from a tape that excludes certain families, because data for them are incomplete in respects that do not concern the issues under discussion here. We are indebted to Pravin Visaria for the Table I tabulation.

13. The sample recovery fractions are as follows:

1964	1966	1968	1970	1972	1974	1976
41.6%	80.6%	99.7%	98.0%	99.6%	98.6%	100.0%

Table II illustrates the fact that concentration ratios can be negative. Such an observation implies an inverse relationship between a factor income of a given type and total income: the concentration ratio is zero for a factor if all income groups receive an equal amount of income of the given factor type. Table II also illustrates the possibility that trends in concentration ratios can be quite different from those for Gini coefficients. Two extreme examples are provided by the data on wages and agricultural incomes for all families. With respect to agricultural income, the concentration ratio falls from 0.45 to 0.10 over the period (with a minor break in monotonicity for 1972). In contrast, the Gini coefficient rises from 0.65 to 0.83 (again with minor breaks in monotonicity). Similarly, the Gini coefficient trend for wages is U-shaped with a minimum of 0.43 in 1972: the trend of the concentration ratio has an inverse U-shape with a maximum in 1972 of 0.30. It is apparent from these examples that the level and trend of the concentration ratio $C(\tilde{x}_k/\tilde{y})$ may be misleading as a guide to the level and trend of the factor Gini $G(\tilde{x}_k)$.

It is not difficult to provide an explanation of such results. A high value for the Gini coefficient $G(\tilde{x}_k)$ is a necessary but insufficient reason for having a high value of $C(\tilde{x}_k,\tilde{y})$. Suppose, for example, that there is a high Gini for agricultural income, mainly due to the fact that many families have zero agricultural incomes, so that such income is concentrated among a few farm families. Then if average income levels among farm families are similar to those among nonfarm families, the importance of agricultural income as an element of total family income will be largely independent of the size of total income, irrespective of inequality among farm families. Hence the concentration ratio will be small: the size of total income gives little or no guide to the size of agricultural income in a mixed population of farm and nonfarm families in this case.

VI. Exact Decomposition Using Individual Family Data

When individual family data are available, the results in Section II above can be used to obtain an exact decomposition. This is based on the results in equations (11) and (14), which combine to give

$$G(y) = \sum_k \phi_k R(y,x_k) G(x_k), \tag{36}$$

i.e., an exact decomposition of the Gini coefficient for total income $G(y_i)$ into separate components for each type of factor income. The component of the decomposition corresponding to each type of factor

TABLE III

EXACT DECOMPOSITION OF TOTAL INCOME GINI AND OF CONCENTRATION RATIOS INTO RANK CORRELATION EFFECTS AND FACTOR GINI EFFECTS USING INDIVIDUAL FAMILY DATA, TAIWAN, 1964 TO 1976

	Wage income			Profit income			Agricultural income			All other income			Total income
	Concentration ratio $C(x_W/y)$	Rank correlation ratio $R(y,x_W)$	Gini coefficient $G(x_W)$	Concentration ratio $C(x_\pi/y)$	Rank correlation ratio $R(y,x_\pi)$	Gini coefficient $G(x_\pi)$	Concentration ratio $C(x_A/y)$	Rank correlation ratio $R(y,x_A)$	Gini coefficient $G(x_A)$	Concentration ratio $C(x_X/y)$	Rank correlation ratio $R(y,x_X)$	Gini coefficient $G(x_X)$	Gini $G(y) = \sum \phi_k \times R(y,x_k) \times G(x_k)$
							Agricultural families						
1964	0.122	0.229	0.532	0.270	0.583	0.463	0.420	0.879	0.478	0.553	0.538	1.028[a]	0.351
1966	0.191	0.318	0.601	0.277	0.632	0.438	0.344	0.843	0.408	0.494	0.533	0.926	0.319
1968	0.250	0.423	0.591	0.245	0.568	0.431	0.283	0.640	0.442	0.562	0.635	0.885	0.305
1970	0.276	0.504	0.548	0.288	0.632	0.456	0.241	0.554	0.435	0.504	0.570	0.884	0.287
1972	0.275	0.545	0.505	0.264	0.589	0.448	0.240	0.515	0.466	0.509	0.574	0.887	0.280
1974	0.252	0.493	0.511	0.279	0.592	0.471	0.273	0.572	0.477	0.426	0.469	0.908	0.286
1976	0.253	0.495	0.511	0.288	0.634	0.454	0.262	0.555	0.472	0.376	0.448	0.840	0.276

						Nonagricultural families							
1964	0.264	0.542	0.487	0.462	0.665	0.695	0.084	0.106	0.846	0.452	0.619	0.730	0.348
1966	0.263	0.548	0.480	0.511	0.658	0.777	−0.306	−0.337	0.908	0.437	0.580	0.753	0.333
1968	0.244	0.529	0.461	0.545	0.726	0.751	−0.195	−0.213	0.916	0.438	0.585	0.749	0.335
1970	0.236	0.557	0.424	0.449	0.641	0.700	−0.077	−0.084	0.919	0.394	0.516	0.764	0.297
1972	0.283	0.722	0.392	0.480	0.676	0.710	−0.023	−0.024	0.971	0.288	0.350	0.824	0.306
1974	0.268	0.638	0.420	0.502	0.715	0.702	−0.008	−0.008	0.985	0.354	0.450	0.786	0.313
1976	0.247	0.604	0.409	0.436	0.672	0.649	0.064	0.065	0.987	0.334	0.440	0.759	0.287
						All families							
1964	0.174	0.320	0.543	0.348	0.587	0.593	0.462	0.696	0.664	0.438	0.498	0.879	0.354
1966	0.238	0.432	0.551	0.439	0.626	0.701	0.360	0.466	0.773	0.430	0.531	0.810	0.328
1968	0.264	0.509	0.519	0.483	0.691	0.699	0.175	0.217	0.806	0.477	0.601	0.794	0.329
1970	0.273	0.553	0.494	0.399	0.625	0.638	0.121	0.158	0.767	0.447	0.550	0.813	0.297
1972	0.301	0.686	0.439	0.457	0.664	0.688	0.085	0.100	0.850	0.343	0.407	0.842	0.305
1974	0.280	0.615	0.455	0.484	0.702	0.689	0.133	0.151	0.880	0.379	0.467	0.811	0.309
1976	0.269	0.595	0.452	0.418	0.667	0.627	0.097	0.113	0.857	0.362	0.464	0.780	0.289

a. $G(x_k)$ calculated via cov $(x_k, r(x_k))$ can exceed unity if some values of x_k are negative.

income is the product of three terms: (i) the share of the factor in total income; (ii) a rank correlation ratio as defined in equation (14); and (iii) the Gini coefficient for the distribution of income of the given factor type. Table III sets out this exact decomposition using the same data source as Table II.[14] (Values for ϕ_k are the same as in Table II and therefore are not repeated in Table III.)

Apart from the modifications of concentration ratios and Gini coefficients that arise from the use of individual as opposed to grouped family data, the new information in Table III concerns the rank correlation ratios $R(y,x_k)$. These ratios show some interesting trends, notably the decline among all families and among agricultural families in the rank correlation ratio for total income *versus* agricultural income. The implication of these results is that the ranking of families by agricultural income has declined in importance as a determinant of ranking by total income. This is perhaps what one would expect, given the declining importance of agricultural income as a share of total income, which can be seen from Table II. It can also be noted that the Gini coefficient for agricultural income among agricultural families has tended to increase since the late sixties. This would be consistent with a greater tendency to move toward other forms of remunerative activity (e.g., wage employment) by members of those families that are at the lower end of the agricultural income distribution.

From the point of view of statistical decomposition, the results in Table III tell a complete story of the trends in total income Gini coefficients in terms of factor components. However, the mapping from the statistical magnitudes to the concepts of economic theory is not straightforward. Such a mapping is provided in F-R-K, but under the assumption that the correlation effects $R(x_k,y)$ are stable over time and approximately unity. Our new results show that this is too restrictive, and point to the need for an interpretation of the correlation effects in terms of economic concepts. We have no such interpretation to offer at this juncture. At best, we can suggest that the correlation effects are potentially associated with some dynamic aspects of development, not least with migration out of the farm sector and into wage employment. This would be consistent with the observation that wages grow in importance as an explanation of (i.e., in their correlation with) total income; and with a narrowing of the differential in incomes between farm and nonfarm families, which tends

14. It can be noted that one of the Gini coefficients in Table III exceeds unity. This is possible, given that our definitions do not exclude the computation of a Gini coefficient for a variable that takes on negative values for some families.

to reduce the correlation between total income and income from agriculture. Such phenomena are consistent with our observations and sensible from a theoretical perspective. In going beyond them, it may be that a marriage of the statistical approach in this paper and the interpretation of Gini coefficients in Pyatt [1976] would take us further toward a sound theoretical interpretation for the correlation effects. But such developments must wait for the future; for the present we leave these questions open.

DEVELOPMENT RESEARCH CENTER, WORLD BANK
THE INSTITUTE OF THE THREE PRINCIPLES OF THE PEOPLE,
ACADEMIA SINICA, TAIPEI
YALE UNIVERSITY

REFERENCES

Ayub, M., "Income Inequality in a Growth-Theoretic Context: The Case of Pakistan," Ph.D. thesis, Yale University, 1977.

Fei, J. C. H., G. Ranis, and S. W. Y. Kuo, "Growth and the Family Distribution of Income by Factor Components," this *Journal*, XCII (Feb. 1978), 17–53.

Fields, G. S., "Income Inequality in Urban Colombia: A Decomposition Analysis," *Review of Income and Wealth*, Series 25, No. 3 (Sept. 1979).

Gastwirth, J. L., "The Estimation of the Lorenz Curve and Gini Index," *Review of Economics and Statistics*, LIV (Aug. 1972), 306–16.

Mangahas, M., and E. Gamboa, "A Note on Decomposition of the Gini Ratio by Family and by Type of Income," *Philippine Review of Business and Economics*, XIII (Dec. 1976), 97–130.

Pyatt, G., "On the Interpretation and Disaggregation of Gini Coefficients," *Economic Journal*, LXXXVI (June 1976), 243–55.

Rao, V. M., "Two Decompositions of Concentration Ratio," *Journal of the Royal Statistical Society*, Series A, CXXXII, Part 3 (1969), 418–25.

THE WORLD BANK
Headquarters:
1818 H Street, N.W.
Washington, D.C. 20433, U.S.A.

European Office:
66, avenue d'Iéna
75116 Paris, France

Tokyo Office:
Kokusai Building,
1-1 Marunouchi 3-chome
Chiyoda-ku, Tokyo 100, Japan

The full range of World Bank publications, both free and for sale, is described in the *World Bank Catalog of Publications*, and of the continuing research program of the World Bank, in *World Bank Research Program: Abstracts of Current Studies*. The most recent edition of each is available without charge from:

PUBLICATIONS UNIT
THE WORLD BANK
1818 H STREET, N.W.
WASHINGTON, D.C. 20433
U.S.A.