Does Geographical Aggregation Cause Overestimates of the Returns to Schooling?

World Bank Reprints

No. 276. Sweder van Wijnbergen, "Interest Rate Management in LDCs," Journal of Monetary Economics


No. 278. Oli Havrylyshyn and Martin Wolf, "Recent Trends in Trade among Developing Countries," European Economic Review


No. 280. Walter Schaefer-Kehnert and John D. Von Pischke, "Agricultural Credit Policy in Developing Countries," translated from Handbuch der Landwirtschaft und Ernährung in den Entwicklungsländern (includes original German text)

No. 281. Bela Balassa, "Trade Policy in Mexico," World Development


No. 283. Anne O. Krueger, "Trade Policies in Developing Countries," Handbook of International Economics


No. 288. Hollis B. Chenery, "Interaction between Theory and Observation in Development," World Development


No. 291. Danny M. Leipziger, "Lending versus Giving: The Economics of Foreign Assistance," World Development


No. 294. Barend A. de Vries, "International Ramifications of the External Debt Situation," The AMEX Bank Review Special Papers

DOES GEOGRAPHICAL AGGREGATION CAUSE OVERESTIMATES OF THE RETURNS TO SCHOOLING?*

Nancy Birdsall and Jere R. Behrman

Rates of returns to schooling are usually estimated from cross-section national sample or census data on income and schooling. Reported estimates of rates of return average over 20 per cent for primary schooling in developing countries, and are interpreted as justifying substantial investments in schooling. Interpretation of such estimates is usually qualified by comment on possible upward biases due to omitted variables like ability, motivation and school quality. The observation has also been made that a high return to schooling in a cross-section does not necessarily imply a high return over time; the former may reflect scarcity returns for the more schooled that would not necessarily persist if schooling opportunities were expanded. In contrast, no attention has been paid to biases which can arise from aggregating over space to obtain estimates from national cross-section samples.

In this paper we investigate the possibility of biases due to such geographical aggregation. In Section I we outline a simple method for estimating a national rate of return to schooling based on disaggregation of the sample by each individual's region of origin and destination. Using data from the 1970 Brazilian census, we show that the estimated rate of return to schooling is much lower with this method than with the standard procedure, and that our modified estimated is statistically

* This paper was prepared as part of the World Bank Research Project 672-21, 'Studies on Brazilian Distribution and Growth'. We are grateful to Punam Chuan for efficient programming to Robin Sickles for a useful discussion of statistical biases, George Psacharopoulos, Richard Sabot, T. Paul Schultz, Paul Taubman and Jeffrey G. Williamson for valuable comments on earlier related studies and to referees. None of these individuals nor the World Bank is responsible for any view expressed here. The authors share equally the responsibility for this paper.

1 The World Bank (1980), for example, partly on the basis of Psacharopoulos' (1981) review of existing earnings function estimates, advocates greatly increased investments in primary schooling in the developing countries. Colclough (1982) presents a more recent and quite parallel argument.

2 For recent discussion of these issues, see the studies mentioned in note one and Behrman (1983); Behrman et al. (1980) and Taubman (1977). For estimates of biases in estimated schooling returns in developing countries due to omitted family-background related ability and motivation and school quality, see Behrman and Birdsall (1983), Behrman and Wolfe (1984), Boissiere et al. (1983), and Wolfe and Behrman (1982).

superior. In Section II we consider various explanations for the apparent upward bias in the standard estimate. In a concluding section we comment on the implications of our findings for total schooling investments. Throughout, we abstract from other possible biases in the estimated returns to schooling in order to focus on the possible effects of geographical aggregation.

I. STANDARD AND DISAGGREGATED ESTIMATES OF THE RETURNS TO SCHOOLING IN BRAZIL

The standard approach is to estimate a semilogarithmic earnings function from cross-section data:

\[ \ln Y = a + rS + cE + dE^2 + U \]  

(1)

where \( Y \) is nominal earnings, \( S \) is grades of schooling, \( E \) is experience (to represent post-schooling investment), and \( U \) is a disturbance term. In the Mincerian (1974) interpretation \( r \) is the private rate of return to the private cost of foregoing labour market participation in order to attend school, and \( a \) is the logarithm of the no-experience, no-schooling earnings level. The essence of the Mincerian derivation is illustrated in the simple case in which an individual chooses between income streams 1 and 2, both with a fixed span of work life and with no post-schooling change in human capital. Income stream 2 is higher because of the extra schooling acquired at a private opportunity cost of delaying the start of the receipt of earnings while in school. In equilibrium, an individual is indifferent between the two income streams, so their present discounted values are equal, with \( r \) being the appropriate discount rate. By letting income stream 1 be that obtained with no schooling, the semilogarithmic relation between income and schooling in equation (1) is obtained from the equilibrium condition (ignoring the experience terms). Equation (1) is estimated under the (often implicit) assumption that the estimate of \( r \) is not contaminated by omitted variable bias and simultaneity bias.

We estimate equation (1) for males aged 15 to 65, using the Brazilian 1/100 Public Use sample of households of the 1970 census. This provides information on individual income and years of schooling. Paid labour force experience is not reported, so we use potential experience, defined as the number of years an individual has been aged 15 or older and not in school. Most males participate more or less continuously in the labour force, so for males potential experience is a good proxy for actual experience. We consider only males because women are less

As in many such studies, we use income to represent earnings. There has been some controversy about the income figures in the census since they imply a significantly lower national income than do the national accounts. We discuss possible implications of using these income figures in Section II.
likely to have worked continuously, and to avoid the complication of selectivity regarding which females participate in the labour force.

In order to examine the effect of geographical aggregation on estimates of equation (1) we divide the sample into six geographical areas: rural and urban areas for each of three major regions -- the relatively prosperous and industrial southeast, the relatively poor and agricultural northeast, and the combined frontier and central states. For each individual we take into account both the region in which he went to school, his 'origin', and the region where he now resides, his 'destination'.

Table 1 indicates the distribution of the 4,113 members of the sample among the 36 origin-destination combinations considered, and shows mean income and mean grades of schooling for each of the 36 origin-destination combinations. The numbers along the diagonal in the top portion of the table represent those individuals who have not migrated. Almost one in five (18.5 per cent) of the adult males in the sample had migrated and so were in a geographical area in 1970 different from that in which they had received schooling. The most common move was from rural to urban areas (7.8 per cent of the total sample), with the southeast urban area the largest gainer in absolute terms (from 31.0 per cent by origin to 37.7 per cent by destination). Mean income differences are greatest between urban and rural areas, with mean income in rural destinations (origins) only 31 per cent (35 per cent) of that in urban ones. The same is true for schooling differences, with mean schooling in rural destinations (origins) only 27 per cent (30 per cent) that of urban ones. Among the three regions, mean income and schooling are highest for the southeast and lowest for the northeast.

The first column in Table 2 gives standard estimates of equation (1) using this sample. The estimated rate of return on schooling is 20.5 per cent. The second column presents estimates of our modified version of equation (1). In the modified version, both the constant \( a \) and the rate of return \( r \) can be different for each of the 36 origin-destination combinations. Each is allowed to vary as a function of six dichotomous variables that reflect different origins and destinations:

\[
\begin{align*}
  a &= a_0 + a_1 D OR + a_2 D ON + a_3 D OF + a_4 D DR + a_5 D DN + a_6 D DF \\
  r &= r_0 + r_1 D OR + r_2 D ON + r_3 D OF + r_4 D DR + r_5 D DN + r_6 D DF
\end{align*}
\]  

(2a)

(2b)

where the first letter refers to a dichotomous variable \( D \) with a value of one in the indicated state and zero otherwise, the second letter refers to origin \( O \) or destination \( D \), and the third letter refers to rural \( R \), northeast \( N \) or frontier and central \( F \). With this specification the

\footnote{We estimate the modified relation with ordinary least squares procedures though under some (but not all) assumptions about the cause of possible differences in parameters across areas, simultaneity might be a problem (e.g. the simultaneity bias in the determination of schooling). However, with this and most similar data sets, good instruments with which to use simultaneous estimation techniques are not available.}
### TABLE 1A

**Distribution of Sample Among Six Geographical Areas, by Origin (Where Schooled) and by Destination (Where Income Received)**

<table>
<thead>
<tr>
<th>Origin</th>
<th>Rural</th>
<th>Urban</th>
<th>Rural</th>
<th>Urban</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>South-</td>
<td>North-</td>
<td>South-</td>
<td>North-</td>
</tr>
<tr>
<td>Destination</td>
<td>east</td>
<td>east</td>
<td>east</td>
<td>east</td>
</tr>
<tr>
<td>Rural</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Southeast</td>
<td>840</td>
<td>44</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>Northeast</td>
<td>2</td>
<td>692</td>
<td>0</td>
<td>35</td>
</tr>
<tr>
<td>Frontier and Central</td>
<td>23</td>
<td>31</td>
<td>179</td>
<td>9</td>
</tr>
<tr>
<td>Urban</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Southeast</td>
<td>201</td>
<td>44</td>
<td>1</td>
<td>1179</td>
</tr>
<tr>
<td>Northeast</td>
<td>0</td>
<td>49</td>
<td>1</td>
<td>368</td>
</tr>
<tr>
<td>Frontier and Central</td>
<td>4</td>
<td>10</td>
<td>13</td>
<td>25</td>
</tr>
<tr>
<td>Column total and per cent</td>
<td>1,070</td>
<td>870</td>
<td>194</td>
<td>1,277</td>
</tr>
</tbody>
</table>

### TABLE 1B

**Mean Monthly Income by Origin-Destination Combinations**

<table>
<thead>
<tr>
<th>Origin</th>
<th>Rural</th>
<th>Urban</th>
<th>Rural</th>
<th>Urban</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Southeast</td>
<td>146</td>
<td>147</td>
<td>260</td>
<td>183</td>
</tr>
<tr>
<td>Northeast</td>
<td>65</td>
<td>88</td>
<td>149</td>
<td>317</td>
</tr>
<tr>
<td>Frontier and Central</td>
<td>149</td>
<td>156</td>
<td>151</td>
<td>339</td>
</tr>
<tr>
<td>Urban</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Southeast</td>
<td>329</td>
<td>290</td>
<td>501</td>
<td>440</td>
</tr>
<tr>
<td>Northeast</td>
<td></td>
<td>140</td>
<td>804</td>
<td>314</td>
</tr>
<tr>
<td>Frontier and Central</td>
<td>206</td>
<td>202</td>
<td>404</td>
<td>379</td>
</tr>
<tr>
<td>Column means</td>
<td>180</td>
<td>108</td>
<td>156</td>
<td>487</td>
</tr>
</tbody>
</table>
estimates of $a_0$ and $r_0$ are for individuals with both origin and destination in the urban southeast, the most common origin and destination combination for the sample, with 28.7 per cent of the total.\footnote{Though this specification is parsimonious, it still results in considerable multicollinearity and high standard errors on the point estimates. The bivariate correlations among the additional variables in the modified specification range from $-0.22$ to $0.84$, with a number in the upper end of this range.}

This modification of the standard specification does make a difference. An $F$ test of the imposition of the restrictions that reduce the modified specification to the standard one strongly rejects the imposition of those restrictions at the 1 per cent level.\footnote{$F$ is 40.9 as compared with a critical value of 2.3 for ten restrictions and over 4,000 degrees of freedom in the denominator.}

Tables 3A and 3B show the implied point estimates for $a$ and $r$ for each of the 36 origin–destination combinations for which there are at least seven observations. The column means are the averages for all individuals from each of the six origins, weighted by the number of individuals from each origin, and the row means are the averages for all individuals in the six destinations, similarly weighted. Both the

---

**Table 1C**

*Mean Years of Schooling by Origin-Destination Combinations*

<table>
<thead>
<tr>
<th>Origin</th>
<th>Rural</th>
<th>Urban</th>
<th>Frontier and Central</th>
</tr>
</thead>
<tbody>
<tr>
<td>Destination</td>
<td>South-east</td>
<td>North-east</td>
<td>Central</td>
</tr>
<tr>
<td>Rural</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Southeast</td>
<td>1.7</td>
<td>1.0</td>
<td>-</td>
</tr>
<tr>
<td>Northeast</td>
<td>0.5</td>
<td>0.4</td>
<td>-</td>
</tr>
<tr>
<td>Frontier and Central</td>
<td>1.8</td>
<td>0.6</td>
<td>1.0</td>
</tr>
<tr>
<td>Urban</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Southeast</td>
<td>3.0</td>
<td>1.8</td>
<td>3.0</td>
</tr>
<tr>
<td>Northeast</td>
<td>-</td>
<td>1.3</td>
<td>4.0</td>
</tr>
<tr>
<td>Frontier and Central</td>
<td>1.0</td>
<td>1.0</td>
<td>3.2</td>
</tr>
<tr>
<td>Column means</td>
<td>2.0</td>
<td>0.6</td>
<td>1.1</td>
</tr>
</tbody>
</table>
TABLE 2

Standard and Modified Estimates of Return to Schooling for Adult Brazilian Males in 1970

<table>
<thead>
<tr>
<th>Right-side variables</th>
<th>Standard estimates</th>
<th>Modified estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schooling</td>
<td>0.205 (37.1)</td>
<td>0.169 (22.3)</td>
</tr>
<tr>
<td>Experience</td>
<td>0.153 (29.1)</td>
<td>0.143 (28.2)</td>
</tr>
<tr>
<td>Experience^2</td>
<td>-0.0026 (21.8)</td>
<td>-0.0024 (21.1)</td>
</tr>
<tr>
<td>Constant</td>
<td>2.73</td>
<td>3.32</td>
</tr>
</tbody>
</table>

Consant adjustments

Origin:
- Northeast: 0.431 (4.4)
- Frontier and Central: 0.042 (0.3)
- Rural: -0.112 (1.4)

Destination:
- Northeast: -0.645 (6.2)
- Frontier and Central: 0.141 (1.0)
- Rural: -0.532 (6.7)

Rate of return adjustments

Origin:
- Northeast: -0.050 (2.3)
- Frontier and Central: -0.001 (0.0)
- Rural: -0.026 (1.2)

Destination:
- Northeast: 0.039 (1.6)
- Frontier and Central: -0.036 (1.2)
- Rural: -0.039 (1.7)

R^2: 0.37
SE: 1.24

T-statistics are in parentheses.

constants and rates of returns are higher for urban origins and destinations than for rural ones. The highest rate of return is 16.9 per cent for the southeast urban origin and destination. In general most of the rates in Table 3B are well below the 20.5 per cent obtained using the standard specification. The rate of return based on the weighted average of all 36 cells is 12.7 per cent, only three-fifths of the return obtained using the standard specification.

Which rate of return is the correct one for Brazil as a whole? The answer depends on the nature and extent of the various types of what we call 'geographical aggregation' biases discussed below. If there are no
TABLE 3A
Constants Implied for 36 Origin and Destination Combinations by Modified Estimates in Table 2

<table>
<thead>
<tr>
<th>Origin</th>
<th>Rural</th>
<th>Urban</th>
<th>Frontier and Central</th>
<th>Frontier and Central</th>
<th>Weighted row means</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>South-</td>
<td>North-</td>
<td>South-</td>
<td>North-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>east</td>
<td>east</td>
<td>east</td>
<td>east</td>
<td></td>
</tr>
<tr>
<td>Rural</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Southeast</td>
<td>15</td>
<td>22</td>
<td>16</td>
<td>25</td>
<td>15.5</td>
</tr>
<tr>
<td>Northeast</td>
<td>-</td>
<td>12</td>
<td>-</td>
<td>13</td>
<td>12.0</td>
</tr>
<tr>
<td>Frontier and Central</td>
<td>17</td>
<td>26</td>
<td>17</td>
<td>29</td>
<td>18.8</td>
</tr>
<tr>
<td>Urban</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Southeast</td>
<td>25</td>
<td>38</td>
<td>28</td>
<td>43</td>
<td>29.1</td>
</tr>
<tr>
<td>Northeast</td>
<td>-</td>
<td>20</td>
<td>-</td>
<td>22</td>
<td>21.7</td>
</tr>
<tr>
<td>Frontier and Central</td>
<td>-</td>
<td>44</td>
<td>30</td>
<td>49</td>
<td>36.1</td>
</tr>
<tr>
<td>Weighted column means</td>
<td>17</td>
<td>15.1</td>
<td>17.9</td>
<td>27.4</td>
<td>21.8</td>
</tr>
</tbody>
</table>

TABLE 3B
Percentage Rates of Return to Schooling Implied for 36 Origin and Destination Combinations by Modified Estimates in Table 2

<table>
<thead>
<tr>
<th>Origin</th>
<th>Rural</th>
<th>Urban</th>
<th>Frontier and Central</th>
<th>Frontier and Central</th>
<th>Weighted column means</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>South-</td>
<td>North-</td>
<td>South-</td>
<td>North-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>east</td>
<td>east</td>
<td>east</td>
<td>east</td>
<td></td>
</tr>
<tr>
<td>Rural</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Southeast</td>
<td>10.4</td>
<td>5.4</td>
<td>13.0</td>
<td>8.0</td>
<td>10.3</td>
</tr>
<tr>
<td>Northeast</td>
<td>-</td>
<td>9.3</td>
<td>-</td>
<td>11.9</td>
<td>9.4</td>
</tr>
<tr>
<td>Frontier and Central</td>
<td>6.8</td>
<td>1.8</td>
<td>6.7</td>
<td>4.4</td>
<td>6.3</td>
</tr>
<tr>
<td>Urban</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Southeast</td>
<td>14.3</td>
<td>9.3</td>
<td>16.9</td>
<td>11.9</td>
<td>16.0</td>
</tr>
<tr>
<td>Northeast</td>
<td>-</td>
<td>13.2</td>
<td>-</td>
<td>15.8</td>
<td>15.6</td>
</tr>
<tr>
<td>Frontier and Central</td>
<td>-</td>
<td>5.7</td>
<td>10.6</td>
<td>8.3</td>
<td>11.7</td>
</tr>
<tr>
<td>Weighted column means</td>
<td>11.2</td>
<td>9.0</td>
<td>7.1</td>
<td>16.6</td>
<td>12.7</td>
</tr>
</tbody>
</table>

Note: Point estimates based on less than seven observations are indicated by a dash.
such biases, the standard estimate of 20.5 per cent is correct. If the dichotomous variables are only representing differential intercepts and differential rates of return across areas, the best estimate is the weighted average of 12.7 per cent. We return to the question of the best estimate at the start of Section III. But the difference between 20.5 and 12.7 per cent suggests that geographical aggregation bias can be considerable, and that its causes warrant further exploration.

II. WHY POSITIVE AGGREGATION BIAS?

We consider six reasons for positive bias in $r$ associated with geographical aggregation, and indicate how each should be reflected in the differences in returns among individuals in the various origin-destination combinations.

1. Geographical Variation in Prices is Likely to be Associated with Average Schooling Levels

The true relation between schooling and income is in terms of real, not nominal income:

$$\ln \left( \frac{Y}{P} \right) = a^* + r^* S + c^* E + d^* E^2 + U^*$$  

(3)

where $P$ is the appropriate deflator. If geographical differences in price deflators are ignored and if price deflators and schooling are correlated across space, omitted variable bias results:

$$\lim r^* = r^* + b_{12}$$  

(4)

where $b_{12}$ is the regression coefficient in the 'auxiliary regression' of the excluded $\ln P$ on $S$.

The sign of this bias depends on the sign of the association between $\ln P$ and $S$. In most countries schooling tends to be greater in more

---

8 A referee has suggested that, if the standard estimate is the preferred one, the alternatives may be lower due to greater random measurement error in the more disaggregate estimates. But in such a case the random measurement error in schooling would have to be quite large, a noise to signal ratio averaging over 60 per cent to reconcile the assumed true standard estimate with the weighted average one. We see no reason to expect a measurement error anywhere near this magnitude. Moreover, in such a case it is not clear why the modified version of equation (1) would be preferable statistically to the standard version. Therefore, while there may be more measurement error in the disaggregated estimate than in the standard one, we doubt that such error is a major factor in the differences among the estimates.

9 It might seem prima facie that the rate of return obtained from all of the modified estimates by setting all dichotomous variables equal to zero is of interest for an alternative national estimate. But this is just an estimate for the excluded category in the definitions of the dichotomous variables, which is an arbitrary choice. In our case the excluded category is the origin and destination in the urban southeast, for which case we obtain an estimate of 16.9 per cent. This is almost certainly an upper bound estimate because migration selectivity on unobserved ability and motivation leads to higher ability and more motivated individuals in this area than on the average in the nation.

10 We assume here that $\ln P$ is uncorrelated with $E$ and $E^2$. 
urban areas for a host of reasons: the ‘urban bias’ of governments, spatial economies of scale, and the existence of private schooling due to higher income levels. In most countries, prices are also correlated with urbanization, in part because of higher rents for housing space. Equation (4) thus implies that the return to schooling will be overestimated if geographical price variation is ignored.

There is little doubt that price variation is important in Brazil. Thomas (1982, p. 39) estimates considerable price variation in Brazil in the mid 1970's:

'Estimated cost of living indices vary widely within Brazil, ranging roughly from 150 or more in the cities of Sao Paulo, Rio de Janeiro and Brasilia (compared to a national average of 100) to about 100 in the majority of small towns, and 90 or less in the rural areas.'

Unfortunately his estimates are for data collected five years after the 1970 census, they cannot be matched up with our areas, and no better spatial cost-of-living indices exist. We cannot use his indices to control directly for regional price differentials in our estimates.

However, in our preferred estimates the dichotomous variables for the constants control for inter-area price variations. If the inter-area price differentials dominate the estimated pattern of constants across areas, the coefficients of these dichotomous variables, considering both the destination and origin variables, should be higher for the areas with higher costs-of-living. The estimates in Table 3 are generally consistent with this pattern. Costs-of-living are higher in urban than in rural areas, and all but one of the ten highest constant estimates are for urban destinations. Costs-of-living are higher in the southeast and the frontier and central areas than in the northeast, and (controlling for whether the destination is urban or rural) all of the constants are higher for the southeast and the central and frontier destinations than for the northeast destination. The addition of these dichotomous variables for the constants thus reduces any omitted variable bias. The exclusion of any such controls for inter-area price differentials in the standard specification normally used is likely to cause upward bias in the estimate of schooling impact in samples with inter-area price differentials.

Thomas suggests there are also considerable cost-of-living differences within areas, which means that even our modified estimates are subject to upward bias. In principle this bias could be further reduced by subdividing the national sample into more areas, but as this would add further parameters and more problems with multicollinearity, we do not do so.

11 See Thomas (1980a, b and 1982). Higher prices in urban areas for market-purchased goods, however, may be partially offset by greater availability of public services.
2. Underreporting of Earnings in Rural Areas and Inclusion of Unearned Income in Urban Areas means that the Positive Association Between Income and Schooling could be Exaggerated

Income reported by individuals in the 1970 census only accounts for about 60 per cent of income as measured in the Brazilian national accounts. To the extent that earnings are more likely to be underreported in rural areas (where much payment is in kind) and schooling levels are lower in rural areas, the rate of return is overestimated in an aggregate estimate due to omitted variable bias (in a way analogous to that caused by omission of a price deflator). Of course much underreporting of income is also likely in urban areas, particularly among those at the top of the income distribution. This could cause an underestimate of schooling returns if the unreported income is earned income. But to the extent that this hidden income is not labour but property income, its omission from what is ideally an earned income-schooling relation is an advantage.

The negative coefficients on the rural dichotomous variables (both the constants and the rates of return for both origin and destination variables) are consistent with systematic underreporting of earnings in rural areas and systematic inclusion of unearned income in urban areas, as well as with a higher cost-of-living in urban areas.

3. Higher Income Causes as well as Results from Higher Schooling and Regional Income Patterns are Serially Correlated Over Time, so there may be Upward Simultaneity Bias

Higher family income tends to cause higher schooling for children because public expenditures on schooling are partially locally financed, because the distribution of public schooling expenditures favours those with more income (e.g. Selowsky, 1979), and because those households with higher income have lower private marginal utility costs for foregone child earnings while children are in school and better access to capital markets. There is also likely to be intergenerational serial correlation in regional average incomes due to persistence across regions in natural advantages, capital stocks, and relative prices. If there is

---

12 If unearned income is random across geographical areas, it causes no bias.
13 For further discussion of underreported income in the 1970 Brazil census and estimates of the extent of underreporting across regions and income groups, see Lluch (1982).
14 Even if there are not variations in schooling across space, the last of these possibilities may cause simultaneity bias in standard estimates of schooling returns. For a study of the determination of schooling of current children in Brazil, see Birdsall (1984). For studies in other Latin American contexts, see Birdsall (1980) and Wolfe and Behrman (1983), both of which report strong income effects in urban areas.
15 Even if there is strong serial correlation in household income, there may not be serial correlation in relative regional incomes if migration is sufficiently equalizing. However, despite enormous internal migration in most countries, many regional income differentials appear to persist for decades.
persistence in regional average incomes, then the dependence of an adult’s schooling on his childhood family income means there is probably simultaneity bias in the standard estimate of the effect of schooling on adult income.

The direction of bias depends on the size of the positive effect of family income on child schooling, the amount of intergenerational serial correlation in income across areas, and the income return to schooling \((r)\). The calculations in the appendix, using assumptions for these values based on empirically observed relations in Brazil, suggest that this simultaneity bias is positive for our sample, causing an upward bias in the standard estimated return to schooling. Such an upward simultaneity bias is more likely in the standard cross-section estimates the stronger is the intergenerational income correlation across geographical areas.

4. Migratory Costs are not Taken into Account

In the Mincerian interpretation, as noted above, \(r\) in equation (1) is the return to the investment of foregoing market earnings in order to attend school. But in estimates of equation (1) from cross-section data, part of the estimated return to schooling is the result of the more-educated migrating to areas with higher rewards for schooling given geographical labour market disequilibria. Since such migration involves costs that are not incorporated into the Mincerian derivation, the estimate of \(r\) once again is an overestimate of the return to schooling. In fact it is a return to the opportunity cost of time spent in school plus the cost of migration. If migration costs are relevant, estimated constants will be high in destination areas relative to origin areas because of the failure to net out the migration costs in the estimation of the returns in the destination. (The appendix gives a simple, more formal illustration of how ignoring migration in the derivation of equation (1) can lead to omitted variable bias which overstates the true return to schooling.)

The estimates in Table 3 are for the most part consistent with this prediction. For the largest migratory flow, from the rural southeast to the urban southeast, the constant is higher for those who migrate than for those who stay in the rural southeast (25 versus 15). The same pattern prevails for the second largest flow (northeast urban to southeast urban, 43 versus 22) and for the fourth largest flow (the rural to urban northeast, 20 versus 12). The only other large flow is from southeast urban to southeast rural. The expected pattern is not obtained in this case, but we think that for this group, much of the flow actually represents suburbanization, i.e. small shifts in residence but not necessarily in workplace, which are not migration in the usual sense of the word and in any case represent short moves at low cost.
5. Geographical Labour Market Disequilibrium is Likely to be Combined with a Positive Association Between Schooling and Unobserved Complementary Productive Inputs

Another important consequence of migration costs is that such costs, including information acquisition costs, may preclude short-run geographical labour market equilibrium. If there is a systematic positive correlation between schooling and other productive inputs, the persistence of geographical disequilibrium can cause upward bias in standard estimates of the rate of return to schooling. This possibility can be illustrated most vividly by considering the bias in the estimated rate of return to schooling that would occur if a sample from a rich, highly-schooled country with many complementary productive inputs (e.g. Belgium) were combined with a sample from a poor low-schooled country with few complementary productive inputs (e.g. Afghanistan). Figure 1 illustrates this case. The two solid lines are the true regressions for the two countries, with that for ‘Belgium’ substantially above that for ‘Afghanistan’ because of the differences in complementary inputs. These lines are not parallel, under the assumption that the true rates of return to schooling differ in the two countries. All observations from Belgium are assumed to be bunched around 14 years of schooling, and all observations for Afghanistan are assumed to be bunched around four years of schooling. There is assumed to be complete labour market disequilibrium in that there are no factor movements between the two countries. The estimated rate of return to schooling from the combined
'Belganistan' sample in the standard procedure is given by the broken line. This estimate has a strong upward bias because schooling serves as a proxy for the unobserved complementary productive inputs. This bias would disappear if there were sufficient factor movements (either of labour or of the complementary productive inputs) to eliminate the labour market disequilibrium.

Casual observation suggests that in Brazil the combination of the rural northeast and the urban southeast (see Table 1) results in a Belganistan-type sample, with slow adjustment for disequilibrium reflected in continuing migration. A resulting upward bias in the rate of return to schooling should be reflected in our modified estimates by positive coefficients for the dichotomous variables that represent areas that attract migrants, and negative coefficients for areas that migrants leave. The predicted pattern is obtained for the largest disequilibrium as judged by migratory flows: between rural and urban areas. All of the rural dichotomous variables have negative coefficient estimates. On a regional level there is less support for this prediction. The greatest disequilibrium is between the northeast and the southeast, and only half of the coefficient estimates for the northeast dichotomous variables are negative.16

6. The Private Cost of Schooling is Probably Higher in Rural than in Urban Areas

An important assumption in the Mincerian earnings function derivation is that foregone labour market income is the private cost of schooling. This assumption is likely to understate the true private opportunity cost of withdrawing children from own-farm activities in rural areas. This can be the case even if rural labour markets function fairly well, since transaction and supervision costs can result in children being more productive in own-farm activities than they would be working elsewhere.17 Further, transportation costs for sending children to school are probably higher in rural than in urban areas.18 If areas for which the private opportunity costs are understated are combined with others for

16 The pattern of estimates consistent with labour market disequilibrium could also arise if there is selectivity in migration. If the educated are more able and ability is not measured, we have the now-familiar upward bias in the estimated rate of return to schooling. For example, see Behrman et al. (1980), Behrman and Wolfe (1984c), Da Vanzo and Hosek (1981), and Taubman (1977). If migrants are also more able, we have more than the average bias in the net gainer regions and less in the net loser regions. The pattern of estimates in Table 3B is largely consistent with the possibility that migration selectivity is important. For migrants the estimated rates of return are generally as high or higher for the southeast urban destination as for any other destination, and urban immigrants within each of the three regions receive higher rates of return than rural nonmigrants.

17 Pollak (1982) presents a provocative discussion of the managerial advantages of family firms and farms due to transaction costs.

18 See Birdsall (1984) for evidence that school attendance is much lower in rural than urban areas of Brazil because schooling is less available.
which the Mincerian assumption regarding opportunity costs is more satisfactory, the resulting weighted average estimated returns will be upwardly biased. Rural origin dichotomous variables should have positive coefficient estimates to compensate for the higher than assumed schooling costs, assuming there are no other biases.

In this case the empirical results do not support the prediction. The coefficient estimates of the rural origin dichotomous variables are negative. This is undoubtedly due to the greater relative impact of other biases that work in the other direction, including price variation, migration costs, and the presence of complementary inputs, each of which as shown above, tends to increase coefficients in urban compared to rural areas, in part because urban areas tend to be destination areas.

III. CONCLUDING REMARKS

We argue that estimates of the rate of return to schooling from cross-national samples are likely to be upwardly biased. Geographical aggregation can cause such bias for a number of reasons: omitted regional prices, systematic underreporting of earnings and inclusion of unearned income, simultaneity bias due to the role of income in the determination of schooling, migration costs, geographical labour market disequilibrium, and systematic underrepresentation of the private cost of schooling.

These possibilities are explored for adult males in Brazil by controlling for the geographical origins in which individuals went to school and the geographical destinations in which they now earn income. Our estimates have patterns of intercept and rate of return shifts that are generally consistent with all of these possible sources of geographical aggregation bias, except for the last one. We have no way to identify the relative importance of each because, except for the last, the patterns we predict due to each source of bias are similar. Clearly the last is less important than others which offset it. What we do show is that the combination of these biases is positive and substantial.

The estimate of the national rate of return to schooling, using the weighted average of the ones for geographical areas that incorporates the impact of the dichotomous variables, is 12.7 per cent, about three-fifths of the value of 20.5 per cent obtained in the standard procedure. Our best estimate would actually be below 12.7 per cent (rather than in the range between 12.7 and 20.5 per cent), because we do not correct for intra-area price variations that may be substantial.

It may appear that we find substantial upward bias because we use data from a large country. However, though large, Brazil is probably more integrated by transportation and communication networks than many smaller and poorer developing nations. In Nicaragua for example, a country with only 1.5 per cent of Brazil’s area and 2.1 per cent of Brazil’s population, tests for differences among earnings relations for
regions identified by the degree of urbanization show significant differences across three such regions, suggesting that a rate of return based on an aggregated sample would be similarly biased upward.¹⁹

Our conclusion raises the question whether rates of return based on standard estimates are not overly optimistic about the impact of schooling on earned income. Even if they are, of course, it does not necessarily follow that developing countries are devoting too much of their scarce resources to investment in schooling, since much of the return to schooling may be in nonmonetary form — in terms of better health or a more literate citizenry. Still, the impact of schooling on earned income (and by implication, on labour market productivity) should be an important factor in social investment decisions. The summaries of these estimates by Colclough (1982), Psacharopoulos (1981), and the World Bank (1980) probably overstate the true effect of schooling on earnings and on productivity. To avoid such problems, better procedures and data must be developed to control for geographical aggregation bias, and standard estimates must be reinterpreted in light of the possibility of important geographical aggregation bias as well as of the more commonly acknowledged biases.

Date of Receipt of Final Manuscript: October 1983.

¹⁹ See Behrman and Wolfe (1984a,b). These studies do not discuss the issues raised in this paper.

REFERENCES


APPENDIX

1. *Derivation of Bias due to Simultaneity Bias between Schooling and Income*

Consider a simple model in which all variables are measured in deviation terms: In income \((y)\) depends on schooling \((s)\) with a stochastic term \((u)\), schooling depends on In income at the time when current adults were children \((y^*)\) with a stochastic term \((v^*)\), and there is intergenerational serial correlation of \(\rho\) between \(y\) and \(y^*\).

\[
\begin{align*}
    y &= rs + u \quad \text{(A1)} \\
    s &= m^*y^* + v^* \quad \text{(A2)} \\
    y &= \rho y^* + e \quad \text{(A3)}
\end{align*}
\]

Substitution of equation (A3) into (A2) gives a simultaneous system between (A1) and (A4) (with \(m = m^*/\rho\) and \(v = v^* - em^*/\rho\)):

\[
\begin{align*}
    s &= my + v \quad \text{(A4)}
\end{align*}
\]

If ordinary least square procedures are used to estimate the income equation in (A1) from \(n\) observations:

\[
\begin{align*}
    \hat{r} &= (\Sigma sy)/\Sigma s^2 = r + \left[ \left( \frac{1}{n} \right) \Sigma su \right] / \left[ \left( \frac{1}{n} \right) \Sigma s^2 \right] \\
    \text{p lim } \hat{r} &= r + \text{p lim } \left[ \left( \frac{1}{n} \right) \Sigma su \right] / \text{p lim } \left[ \left( \frac{1}{n} \right) \Sigma s^2 \right] \\
\end{align*}
\]

Substitute equation (A1) into (A4) and solve for \(s\):

\[
\begin{align*}
    s &= (mu + v)/(1 - mr) \quad \text{(A7)}
\end{align*}
\]

Substitute equation (A7) into the numerator and denominator of the second right-side term in (A6) and rewrite in terms of variances and covariances of disturbance terms to obtain equation (A8) (since \(\text{p lim } [(1/n) \Sigma su] = (m\sigma_u^2 + \sigma_{uv})/(1 - mr)\) and \(\text{p lim } [(1/n) \Sigma s^2] = (m^2\sigma_s^2 + \sigma_s^2 + 2m\sigma_{uv})/(1 - mr)\)):

\[
\begin{align*}
    \text{p lim } \hat{r} &= r + (1 - mr) \frac{(m\sigma_u^2 + \sigma_{uv})}{m^2\sigma_s^2 + \sigma_s^2 + 2m\sigma_{uv}}
\end{align*}
\]

where \(m\) is the coefficient of In \(Y\) in the schooling determination relation (with a correction for intergenerational serial correlation in In regional income), \(\sigma_s^2\) is the variance of the disturbance in the schooling determination relation, and \(\sigma_{uv}\) is the covariance between the disturbances in the schooling determination and the income determination relations.

Under the fairly weak assumption that \(\sigma_{uv}\) is not too negative, the direction of the simultaneity bias in equation (A8) depends on the sign.
of $(1 - mr)$. This bias becomes algebraically more positive as the intergenerational serial correlation in income increases.

The following very crude calculation suggests that this bias is positive for the Brazilian sample used in the text. If the income elasticity of schooling implied by equation (A2) were 1.2, the intergenerational serial correlation in incomes across areas ($\rho$) were 0.7, the value of $r$ were 0.127 (see Section 1), and schooling was at the sample mean in Table 1C, then $(1 - mr)$ would be $0.37 > 0$, so the bias would be positive.

2. Derivation of Bias due to Ignoring Migration Costs in the Standard Model

As indicated at the start of Section I, the simplest Mincerian semilogarithmic derivation is the outcome of the equilibrium condition for equating the present discounted value of two income streams associated with $s$ and 0 years of schooling (given a fixed earning life of $q$ years, no change in post-schooling human capital and a fixed discount rate $r$):

$$1 = \frac{V_s}{V_0} = \frac{\int_s^{s+q} Y_s \exp(-rt) dt}{\int_0^q Y_0 \exp(-rt) dt} = \frac{Y_s}{r \exp(-rs)} \frac{1 - \exp(-rq)}{1 - \exp(-rq)}$$

(A9)

Therefore:

$$Y_s = Y_0 \exp(rs) \quad \text{or} \quad \ln Y_s = \ln Y_0 + rs \quad \text{(A10)}$$

Assume that to reap the higher income stream associated with schooling, however, an additional cost must be incurred (say a migration cost, though any cost will have a similar effect). Let $C$ be the present value of that cost. Then the numerator in the equilibrium condition must be modified, so $C$ is subtracted from the numerator in equation (A9):

$$1 = \frac{V_s}{V_0} = \frac{Y_s \exp(-rs)}{Y_0} - \frac{rC}{Y_0[1 - \exp(-rq)]}$$

(A11)

Therefore:

$$Y_s = Y_0 \exp(rs) + rC \exp(rs)/[1 - \exp(-rq)] \quad \text{(A12)}$$

If equation (A10) is estimated instead of equation (A12), that is equivalent to omitting the second term in equation (A12). This biases upward the estimated schooling coefficient since the omitted term is positively associated with the included one [both include $\exp(rs)$] and has a positive coefficient ($C$ is positive by assumption and $r/[1 - \exp(-rq)]$ must be positive if $V_0$ is positive).
Issues of the World Bank Reprint Series are available free of charge from the address on the bottom of the back cover.