INDEXING HOUSEHOLD INVESTMENT IN FORMAL EDUCATION OF SCHOOL-AGE CHILDREN IN DEVELOPING ECONOMIES

Dov Chernichovsky
Population & Human Resources
Development Economics Department
1818 H Street, N.W.
Washington, D. C. 20433

January, 1977

INDEXING HOUSEHOLD INVESTMENT IN FORMAL EDUCATION OF SCHOOL-AGE CHILDREN IN DEVELOPING ECONOMIES*

Dov Chernichovsky
World Bank

I. Introduction

Primary education, which has become a way of life for children in developed areas, is still irregular in many developing areas. This implies that education in these areas is largely a matter of parents' or household's choice rather than the individual's choice.

Since one's education is a key predictor of his earning capacity, studying the relationship between children's primary schooling and their household environment is of major significance.[1/] This can improve our understanding of the determinants of schooling and the related size distribution of incomes in developing economies. Relevant studies with relatively available data would be facilitated by developing an index to measure parents' investment in their school-age children.

The objective of this paper is to define and illustrate such an index. The index proposed is based on the often observed age-specific distribution of grades among school-age children in developing areas. First, we estimate a central tendency in children's schooling based on individual traits. This central tendency, which is informative by itself, is then used to standardize each child's schooling and to define a household index. Finally, the use of the index is illustrated with survey data. While the empirical results are of interest in their own right, the discussion in this paper focuses on the properties of the proposed index.

---

* The views expressed in this paper do not necessarily represent those of the World Bank.

1/ Jacob Mincer, Schooling, Experience, and Earnings (New York: National Bureau of Economic Research, 1974).

II. The Standard

The determinants of a child's schooling, in a given market , may be categorized into individual traits, household characteristics, and community schooling opportunities. The first category includes primarily age, sex, ability, and health. The second incorporates interrelated variables such as household wealth, parents' education and attitudes, ethnic group, family size, and distance to school. The third category includes resources such as facilities and teachers per pupil, fees, and enforcement of attendance.[2/]

Assuming as constant the variables in the third category and, for the moment, assuming as statistically independent the variables in the other categories, a child's level of schooling can be estimated by an explicit relationship of

$$S_i = f(T_i) + U_i, \quad (i = 1...N), \tag{1}$$

where S_i is the i^{th} child's observed level of schooling, and T_i is a vector of his individual traits. U_i is a residual term accounting for variation in S_i that is attributed to his household's characteristics, H_i, and to unobserved and random elements, V_i:

$$U_i = g(H_i) + V_i \text{ [3/]}. \tag{2}$$

2/ For an analytical discussion that interrelates these categories in an economic framework, see Gary S. Becker, Human Capital and Personal Distribution of Income: An Analytical Approach (W.S. Woytinsky Lecture No. 1, Dept. of Economics, Institute of Public Administration, University of Michigan, 1967).

3/ Because the third category variables are community variables, they are assumed to be fixed for households in a given community. These variables should also be used when comparing households across communities, or over time in a developing community.

Relationship (2) is not to be substituted into (1) because most of the households with school-age children are likely to have more than one child in the relevant age group, say 6-14. This poses the problem addressed in this paper. The solution to this problem necessitates, first, standardizing the levels of schooling of children of different ages, and then, aggregating children's schooling within households.

The standard is based on estimating relationship (1) -- to establish a central tendency for the schooling of children in a given community on the basis of their traits, primarily age and sex. Health and ability indices are to a large extent determined by the household environment (H_i) and their measurement is often controversial. This may suggest excluding them from T_i, unless available data and preliminary estimates suggest otherwise.

When T_i comprises age and sex, a compact explicit functional relationship for estimating (1) that is used for illustrative purposes, is

$$S_i = \beta_0 + \beta_1 [A(a_i)] + \beta_2 D_i \ (=1|\text{boy}, =0|\text{girl}) + \beta_3 [A(a_i) \times D_i] + U_i \quad (i = 1...N), \qquad (3)$$

where $A(a_i)$ is some positive and monotonic function of the i^{th} child's age, a_i, and D_i is a dummy variable for his or her sex.[4/]

The transformation of a_i is suggested to allow for a potential decrease in the children's marginal propensity (with respect to age) to attend school. This possibility should be tested because child labor is

---

4/ The advantages of this estimation procedure are that it provides statistics that summarize within and across group variations in schooling, and it suggests a single functional relationship. This procedure is especially advantageous when more variables are included in T_i and when the number of observations is relatively small in particular age-sex cells. Other procedures that compute within age- and sex-group location statistics, and that then test for discrimination between boys and girls within age groups, are equally appropriate.

common in developing areas; as children grow up, their productive capacities increase and consequently, the opportunity costs of their schooling increase also. The coefficients on the age variable, $\hat{\beta}_1$ and $(\hat{\beta}_1 + \hat{\beta}_3)$, can also be influenced in a developing community by increasing schooling opportunities for younger children. The statistical significance and the levels of $\hat{\beta}_2$ and $\hat{\beta}_3$ indicate whether parents, on the average, discriminate in the education of their sons and daughters, and when they do, the nature of this discrimination.

The statistical properties of U_i must be considered to evaluate the estimated coefficients and to discuss the proposed index. The term U_i consists of two elements: one non-stochastic, $g(H_i)$, and one stochastic, V_i. We assume that these elements are statistically independent, COV $[g(H_i), V_i] = 0$, and that V_i is normally distributed with an expected value of 0, $E(V_i) = 0$.[5/] Consequently, U_i has the expected value of $g(H_i)$ and is normally distributed only if $g(H_i)$ is also normally distributed. Since U_i depends partially on a vector of observed variables, its distribution may be inferred from observing the marginal distributions within cells, provided that there are a sufficient number of observations within each cell. A priori, U_i can assume any distribution. Consequently, normality can be assumed when no other information negates it. Otherwise, we must resort to some means of normalizing U_i or to robust tests of significance.[6/]

5/ The first of these assumptions may be strong when ability and health are excluded from T_i, and are indeed determined by household characteristics. We should consider this possibility while interpreting the estimated effects on schooling of household characteristics.

6/ Henry Theil, Principles of Econometrics (New York: John Wiley & Sons, Inc., 1971), pp. 615-622.

The expected value of U_i is unlikely to be zero because $g(H_i)$ is a function of a vector of socio-economic variables that do not have zero means. This, however, does not bias the estimated coefficients when the independent variables in equation (1) take fixed values and are independent of U_i. Such is the case when T_i includes just age and sex.

The variance-covariance terms of U_i must be considered to assess the efficiency of the estimated coefficients, if indeed U_i is normally distributed. $Var(U_i)$ is likely to be a positive function of age because S_i has a wider potential range of variation at higher ages, for a given vector of observed independent variables, and a vector of unobserved and stochastic determinants of U_i. Hence, it is desirable to test and to correct, if found, for heteroscedasticity when estimating equation (1), particularly when the sample size is small.[7/] The covariance term, COV (U_i, U_j), is likely to be positive for any two children from the K^{th} household, because U_i and U_j $(i \neq J;\ i, j \in K)$ are determined by identical household variables. Hence, it is also desirable to test and to correct for autocorrelation, when found, by estimating equation (1) with one child per household, when the loss in degrees of freedom is relatively low. For this purpose it is desirable to select, in a household with more than one child in the relevant age group, either the oldest child or the oldest of the boys and the oldest of the girls. This suggested solution to the autocorrelation problem is based on two assumptions: a. Parents tend to minimize discrimination in the schooling of any two children at

---

7/ Ibid, pp. 244-249.

least of the same sex; and b. older children's schooling reveals better parents' preferences concerning children's education.8/

III. The Individual's Index

The index (Y_i) for the i^{th} child's schooling is defined as the ratio between his observed level of education (S_i) and the predicted level ($\hat{S}_i$) which is given by estimating equation (2) or a more general relationship as implied by (1). Hence,

$$Y_i = \frac{S_i}{\hat{S}_i} = 1 + \frac{U_i}{\hat{S}_i} . \qquad (4)$$

$\hat{S}_i$ is now fixed for all children with identical traits; $\hat{S}_i = \hat{S}_j$ when $T_i = T_j, (i \neq j)$. When $S_i = 0$, it should be substituted by $S_i = \theta$, $0 < \theta < 1$, to measure $S_i = 0$ consistently with positive levels of schooling. Hence, each child's observed schooling is standardized by a norm "typical" of his peer group as defined by common individual traits. Ceteris paribus, the basic properties of the index are:

a. It is a positive (and linear) function of the level of schooling;9/

8/ The first assumption does not hold when parents "specialize" by assigning different children to different activities, one of which is education; then COV $(U_i, U_j) < 0$. This argument bears also on the second assumption, when parents, for example, educate only older sons.

9/ $\frac{dY}{dS} > 0$, $(\frac{d^2Y}{dS^2} = 0)$.

b. It assigns a lower score to a given level of schooling at older age;[10/]

c. It assigns a lower score to a given increase in the level of schooling at older age.[11/]

Given its properties, the index should be evaluated by its predictive power: the consistency of a higher score with a higher probability of more schooling attained eventually, and with a subsequent higher earnings capacity. By these criteria the index has some inevitable weaknesses that result from the nature of the schooling process in many developing areas. The index performs relatively well under a regime where age of entry is uniform and the population of school-age children comprises only children who never attended school, dropouts, repeaters, and normal regular attenders.[12/] The first of the above properties is consistent with measuring schooling in absolute terms, by years of schooling, within age groups.

10/ $\frac{dY}{da} < 0$, when $S \geq \theta$ and constant, since $\hat{S}$ is a positive and monotonic function of age.

11/ $\frac{d^2Y}{dSda} < 0$, when $S \geq \theta$ and $\frac{d\hat{S}}{da} > 0$.

12/ The various subgroups--children who never attended school, dropouts, late entrants, repeaters, and normal regular attenders -- can be studied as separate phenomena, depending on research objectives and availability of data. This index is not considered vis-a-vis this option because it is designed to measure household investment in education, and a household may have children in various subgroups. Furthermore, many surveys collect data on "last grade completed." This does not allow for the separation of the subgroups.

The second and third properties are meaningful in the following sense. When two children of different ages but of the same (positive) level of schooling, or increment in it, are compared, the younger always has a positive probability to have a higher level of education than the older when he reaches the latter's age.[13/] Under this regime the index may still assign a higher score to a child who dropped out a year before than to one who repeated two years, although the latter may eventually achieve more schooling.

However, usually late entrants are also included in the population of school-age children. This introduces some distortion in the index. Around the entry point, say age six and first grade, the index performs relatively poorly as a predictor; it may assign higher scores to children who drop out after the first or the second grade than to children who are out of school but enter school later.

Beyond the entry point the performance of the index improves, on the average, by age groups as final decisions become increasingly apparent. The group that never attended school becomes more definite, because the probability of late entry falls with age. On the other hand, the cumulative number of dropouts, who terminated their formal education, increases. This out-of-school population will have an age-specific lower

---

13/ Note that in a cross-section of children the index may assign different values to children of different ages but with an identical schooling pattern, say regular schooling. Suppose that regular schooling is denoted by a linear function of age, $S = H(a)$, and the standard by a function $S'=g(a)$ according to equation (1). The index remains invariant for any two children of different ages if within the relevant range $dS/S = dS'/S'$. This holds if g is also linear and $g(a^o) = H(a^o)=0$, $a^o > 0$. Generally the relationships $dS/S > dS'/S'$ holds since regular schooling is most likely an upper bound for schooling, and $g''(a) \leq 0$. Hence, the index gives a "premium" to parents who maintain a child in school on a regular basis over a longer period of time. The consequences of this are discussed later on.

average score than the in-school population. This is consistent with higher average present and future schooling for the in-school population than for the out-of-school population.[14/] However, the index may still be misleading in individual cases by failing to discriminate between some recent dropouts and late entrants who are in school, or by assigning a higher score to a dropout than to a late entrant. The probability of such distortions is minimal considering that dropouts usually leave school after repeating a grade or two.

IV. The Household Index

The K^{th} household's schooling index, Y_k, is defined as the arithmetic mean of the individual indices of N_k school-age children in the household:

$$Y_k = \frac{\sum_{i}^{N_k} \frac{S_i}{\hat{S}_i}}{N_k} = 1 + \frac{\sum_{i}^{N_k} \frac{U_i}{\hat{S}_i}}{N_k} \qquad (k = 1...M).[15/] \tag{5}$$

Y_k has the same mathematical properties as the index for the individual child. It can be used as a dependent variable to explain the variation in schooling per child across M households, by estimating an explicit relationship of

$$Y_k = Y(X_k) + \omega_k, \tag{6}$$

where X is a vector of variables that are hypothesized to affect household investment in children's schooling, and ω_k is a linear combination of V_i, which was discussed on page 2:

14/ This statement is consistent with the macro evidence that improved retention at earlier, rather than at later, stages of schooling, brings about higher numbers of students in the later stages. John K. Folger and and Charles B. Nam, Education of the American Population, 1960, Bureau of the Census monograph (Washington, D.C.: Government Printing Office, 1967), p. 67.

15/ It is possible to use the geometric mean, which is commonly used for indices. This, however, complicated any inference about the variance of Y_k.

$$\omega_k = \frac{\sum_{1}^{N_k} \frac{v_i}{\hat{S}_i}}{N_k}, \tag{7}$$

Hence, ω_k is normally distributed with

$$E(\omega_k) = 0 \tag{8}$$

since $E(v_i) = 0$

and $\hat{S}_i$ is fixed for all children with identical traits,

$$\text{and} \quad \text{Var}(\omega_k) = \left[\frac{1}{N_k^2}\left(\sum_{i}^{N_k} \frac{\text{Var}(v_i)}{\hat{S}_i^2} + 2\sum_{i<j}^{N_k} \frac{\text{Cov}(v_i, v_j)}{\hat{S}_i \hat{S}_j}\right)\right], \tag{9}$$

$(i \neq j)$.

A priori, we can reasonably assume Var (ω_k) to be constant. Var (v_i), which increases with age, is deflated by $\hat{S}_i^2$, which is also a positive function of age. The deflating effect of N_k^2, which varies among households, is at least partially offset by the additional positive variance-covariance terms that are associated with the increments in N_k.[16/] Nevertheless, we may wish, after estimating relationship (6), to test whether the estimated residuals ω_k are independent of N_k and of the mean age of school-age children.[17/]

16/ The covariance term in the bracket stands for the previous assumption that non-measured and stochastic effects on schooling of any two children of the same household are probably positively correlated.

17/ As indicated in footnote 13, the individual's index assigns higher values for children attending school on a regular basis. Consequently, it may assign higher values to the household where the mean age of school-age children is higher, ceteris paribus. This causes a problem only if mean age is correlated to any of the explanatory variables in (6).

C. An Illustration

The data were obtained from a cross-sectional household survey that was carried out during 1967-68 in Ankodia, a village in the Indian state of Gujarat.[18] One hundred sixty-one households reported a total of 324 children between ages 6-14, the compulsory schooling ages in India. Table 1 shows the age-specific means and standard deviations of schooling levels, along with the number of children, in each age group. These statistics indicate the irregular schooling pattern in this village.

TABLE 1

Number of Observations (N_a), Mean Levels and Standard Deviations (S.D.) of Schooling by Age

Age	6	7	8	9	10	11	12	13	14	Total
N_a	24	28	48	39	41	34	44	30	36	324
Mean	.500	1.143	1.854	2.410	2.780	3.676	4.623	4.800	5.770	3.071
S.D.	.722	1.238	1.237	1.332	1.993	2.306	2.672	3.033	2.768	2.567

Table 2 shows the estimated coefficients for equation (3). The semi-logarithmic functional relationship provided the best fit after testing both for a linear relationship and for a quadratic one.[19] This may indicate either a decreasing marginal propensity to attend school or better schooling opportunities for younger age groups or both.

18/ For a summary report, see R.M. Patel, Ankodia: Change in Economic Life of a Tobacco Village (Agro-Economic Research Center, Vallabh Vidyanagar, 1970).

19/ The difference between the linear relationship and the one reported has been only marginal.

The differential in schooling between boys and girls is of relatively limited statistical significance in the first equation. However, the estimates are statistically inefficient because of autocorrelation, which is indicated at 0.01 level of significance by the Durbin-Watson statistic, and heteroscedasticity, which is implied by the standard-deviations shown in table 1.

The second equation is based only on the population of elder children, to correct for autocorrelation. Although the Durbin-Watson statistic indicates an inconclusive test, this is an improvement over the first equation. A comparison between these two estimated equations indicates that, on the average, parents apparently do not discriminate between their older and younger children in this village.

To correct for heteroscedasticity (and to restore the degrees of freedom lost in equation 2) the first equation was re-estimated, dividing it by the $(Ln_e \text{ age})^2$ the best deflator of the ones that were tested. The estimated coefficients are shown in the third equation.[20/] They are indeed more efficient than the earlier estimates and indicate a difference between schooling of boys and of girls. While children of both sexes seem to start school about the same age, boys have a higher propensity, on the average, to continue in school and subsequently achieve higher levels of education.

The schooling indices for each child and for each of the 161 households were defined according to the third equation coefficients. Children with no schooling were assigned S=0.5. Table 3 shows the regression coefficients for relationship (6) with the household index as the dependent

---

20/ The R^2, F, and Durbin-Watson statistics are not reported for equation (3) because they are inapplicable, since the intercept of this equation was suppressed during estimation.

TABLE 2

Regression Coefficients
Children's Schooling as a Dependent Variable
(t statistics in parentheses)

Equation No.	Population	Constant	Ln(age)	Sex (Boy=1)	(Sex)x (Ln age)	R^2	F	Durbin-Watson
1	All children	-8.84	5.09 (8.25)	-3.43 (-1.63)	1.76 (1.93)	.36	61.03	1.29
2	Oldest children	-9.49	5.37 (4.92)	-4.21 (1.03)	2.01 (1.19)	.28	20.66	1.50
3	All children	-7.86	4.62 (9.22)	-3.86 (-2.32)	1.96 (2.58)	*	*	*

TABLE 3

Regression Coefficients
Household Education Index as a Dependent Variable
(t statistics in parentheses)

Equation No.	Constant	Household Income	Father's Years of Schooling	Mother Literate (=1)	Father a Civil Servant (=1)	$\left(\frac{\text{No. of children who died}}{\text{No. ever born}}\right)$	Mean Ages of Children 6-14	Adjusted R^2	F
1	0.81902	0.00003 (2.796)	0.04932 (2.424)	0.22637 (2.094)	0.54790 (3.755)	-0.49033 (-2.118)		0.28	13.49
2	1.66059	0.00003 (3.033)	0.06172 (3.092)	0.22739 (2.178)	0.47657 (3.347)	-0.39255 (-1.743)	-0.09022 (-3.498)	0.33	14.09

variable.[21] The coefficients are consistent with prior notions advocated by the theory of human capital. Income and parents' education have a positive and statistically significant effect on children's education as measured. The same effect holds, ceteris paribus, for households headed by civil servants, who have better access to schooling than others and are vulnerable to compulsory schooling laws. As mentioned earlier, some other factors may be related to these findings. High-income, high-education parents may have healthier and more able (better prepared for school) children than low-income, low-education groups. Hence, the income and education variables may also account for these intervening factors.[22] These factors may be controlled, however, by the ratio of the number of children who died to the ones ever born. This ratio, which has a negative effect on investment in schooling, may be a good proxy for health conditions that have a negative effect on schooling. It may also indicate parents' reluctance to invest in children in an environment where children have a lower survival probability.[23]

Of special interest are the effects of two variables: the number of children 6-14, N_k, and these children's mean age, since these variables are associated with the definition of the household index. The effect of N_k

---

21/ Potential biases that may result from applying the wrong functional relationship, or simultaneity, are not considered here.

22/ Earlier unreported estimates indicated that neither younger, pre-school siblings nor older, post-school ones have an effect on household investment in primary education. However, older siblings contribute to household income. See Dov Chernichovsky "Fertility Behavior in Developing Economies: An Investment Approach" Ph.D. dissertation, City University of New York, 1975). pp. 85-87.

23/ For a theoretical formulation of this argument, see Ibid, pp.49-53.

is not reported because of persistent statistical insignificance of the estimated coefficient in preliminary estimates. The coefficient on this variable, however, may be subject to a specification bias because it may be determined simultaneously with children's education. It is also slightly biased downward because of the definition of the index. For example, a household with more children of the same mean age and pattern of schooling is assigned a slightly lower index.[24/]

The second equation in table 3 shows that mean age has a statistically negative effect on household investment in schooling -- as defined. This estimated effect, which primarily improves the fit of the relationship, may combine the increasing costs of educating older children and the improved schooling opportunities for younger ones in a developing community.[25/]

N_k and mean age are both uncorrelated with the residuals of equation 2 in table 3. The simple correlation between the residuals and N_k is -.0021, and between the residuals and mean age, -.0063. Hence, the estimated coefficients are unbiased except for possible specification biases. Tests indicate that heteroscedasticity is associated with mean age.[26/] Hence, the efficiency of the estimated coefficients of equation 2 (table 3) can be improved.

This illustration and the related discussion indicate that the proposed index performs well as a measure for household investment in the formal education of school-age children.

24/ A household with a 10-year old boy in the 4th grade is assigned an index of 1.1658. Another household with two boys, ages 9 and 11 in the 3rd and 5th grades, correspondingly, is assigned an index of 1.1639. This results from the concavity of the estimated semi-logarithmic $\hat{S}$ function.

25/ Inclusion of the standard deviation of children's ages as a measure of dispersion did not yield any significant results.

26/ The test for heteroscedasticity consisted of rearranging the observations by N_k, and by mean age in an increasing order. In the case of the first, only one central observation was removed to allow for variation in N_k at its lower levels. The F statistic for equality in the residual variance was .93. The same procedure with mean age, removing 41 central observations, produced an F statistic of 2.89, indicating a decreasing [illegible] with mean age.