1-66

INSTITUTE FOR RESEARCH ON POVERTY

DETERMINANTS OF SCHOOL ENROLLMENT AND RELATIVE PROGRESS IN SCHOOL

John Conlisk

DISCUSSION PAPERS

THE UNIVERSITY OF WISCONSIN, MADISON, WISCONSIN

Preliminary draft, not to be
quoted without permission of
the author.

DETERMINANTS OF SCHOOL ENROLLMENT
AND RELATIVE PROGRESS IN SCHOOL

John Conlisk

This research was supported by funds granted to the Institute for Research on Poverty at the University of Wisconsin by the Office of Economic Opportunity, pursuant to the provisions of the Economic Opportunity Act of 1964. The conclusions are the sole responsibility of the author.

This paper presents and estimates a simple model explaining school enrollment rates and relative progress in school variables. Demographic variables describing age, color, sex, rural-urban status, education of parents, and income of parents, are used as exogenous explanatory variables. The data came from a special report on education of the 1960 Census [4].

The school enrollment rate of a group within the school age population will be measured as the fraction of the group enrolled in school. Relative progress in school for a group of students will be measured as the fraction who are, in terms of grades completed, ahead of their age group (roughly, the fraction who have skipped grades) minus the fraction who are behind their age group (roughly, the fraction who have flunked grades). Since flunking is far more frequent than skipping, this relative progress rate might be descriptively termed the negative of a "net flunk rate." Table 1 presents the cut off points used by the Census Bureau in deciding when a child is ahead of his age group and when behind it. The value of the relative progress rate for a group of students is a rough measure of the cumulative school performance of the group. It is only a rough measure for at least two reasons. First, the standard of performance for a given child in determining skipping and flunking is the average ability of his classmates; and this varies greatly from school to school. Second, neither skipping nor flunking is an automatic consequence of a superior or inferior performance by a student (given parents and educators reluctance to allow the potentially bad social and psychological side effects). Nonetheless, the relative progress rate has the great advantage of being available on the complete scale of the U.S. Census.

School enrollment rates are economically important at the macro level in determining the rate of investment in human capital; and they are important at the micro level in determining the future size distribution of income. Relative progress rates are important partly because they roughly measure the quality of the human capital produced and partly because they are a partial determinant of enrollment rates.

Section I presents the model, section II the estimates, and Section III a concluding remark.

Table 1. Cut-off Points in Defining Relative Progress Rate

	Year in which Enrolled*		
Age	Behind Age Group	With Age Group	Ahead of Age Group
7	none	1 and 2	3 or more
8	1 or less	2 and 3	4 or more
9	2 or less	3 and 4	5 or more
10	3 or less	4 and 5	6 or more
11	4 or less	5 and 6	7 or more
12	5 or less	6 and 7	8 or more
13	6 or less	7 and 8	9 or more
14	7 or less	8 and 9	10 or more
15	8 or less	9 and 10	11 or more
16	9 or less	10 and 11	12 or more
17	10 or less	11 and 12	13 or more
18	11 or less	12 and 13	14 or more
19	12 or less	13 and 14	15 or more

Source: Page IX of [3].

* The numbers 1 to 8 refer to the eight years of grade school, 9 to 12 to the four years of high school, and 13 and up to college.

I. The Model

The unit of observation is a group of young people, all of the same age; the model traces the group's school behavior over time. Let r_t be the enrollment rate of a group of age t young people; and let p_t be the relative progress rate of those in the group who are in school. [Recall that p_t = (fraction of those enrolled who are ahead of their age group) - (fraction of those enrolled who are behind their age group).] It must be assumed that the group in question is small relative to the entire population of that age; otherwise there would be no standard relative to which to measure the progress variable p_t. Let $\underline{x}$ be a column vector of demographic variables describing the group's characteristics, where $\underline{x}$ includes variables such as color, sex, rural-urban status, education of parents, and income class of parents. It is assumed that $\underline{x}$ does not change with the group's age t. Partly, this assumption is justified by the genuine constancy of most of the $\underline{x}$-variables listed; and partly, the assumption is forced on the model by the limitations of the data used below. The model is as follows --

$$r_t = a_t + \underline{\beta}_t'\underline{x} + \gamma_t p_{t-1} + u_t \qquad (1)$$

$$(\gamma_t > 0)$$

$$\Delta p_t = a_t + \underline{b}_t'\underline{x} + c_t p_{t-1} + v_t \qquad (2)$$

(sign $\underline{b}_t$ same for all t) $(c_t \geq 0)$

Here α_t, $\underline{\beta}_t$, γ_t, a_t, $\underline{b}_t$, and c_t are parameters, $\underline{\beta}_t$ and $\underline{b}_t$ being column vectors with the same dimension as $\underline{x}$. Thus, the parameters may vary with the group's age t. The variables u_t and v_t are random error terms. Some assumptions about parameter values are put in parentheses under the parameters.

Equation (1) states that, at age t, the group's enrollment rate is equal to a linear function of the demographic variables $\underline{x}$ and the previous period's progress rate p_{t-1}, plus an error term. The variable p_{t-1} is included in equation (1) with a positive coefficient, since students who are doing well in school one period seem more likely to continue their education the next period.

Equation (2) determines $\Delta p_t = p_t - p_{t-1}$. Since p_t measures cumulative past performance, then Δp_t measures current performance. Equation (2) thus states that current performance is a linear function of the demographic variables $\underline{x}$ and lagged past performance p_{t-1}, plus an error term. The coefficient c_t of $p_{t-1}$is assumed non-negative; because a negative coefficient would indicate that a better past performance results in a worse current performance, which seems unreasonable. A judgment about the reasonablenees of the assumption that the sign of (every element of) $\underline{b}_t$ is the same for all t must wait till $\underline{x}$ is precisely defined in the next section. However, the sense of the assumption is to make statements like the following. If being non-white has a negative effect on school performance at age t, all else equal, then it will have a negative effect at all ages. Or, if having uneducated parents has a negative effect on school performance at age t, all else equal, then it will have a negative effect at all ages.

The data to be used below in estimating the model consists of a sample of cross section observations on p_t, r_t, and $\underline{x}$ for each of a number of different age groups of young people, all observed in 1960. Since the observations are all at one point in time, no lagged variables are available. This means that (1) and (2), which contain the lagged variable p_{t-1}, cannot be estimated as they are. Fortunately, they can be solved to forms in which p_{t-1} no longer appears.

Equation (2) is a linear, first-order difference equation, complicated by an error term and by parameters which change with t. Its solution is --

$$p_t = A_t + \underline{B}_t'\underline{x} + V_t \tag{3}$$

where --

$$\begin{aligned} A_t &= a_t + \sum_{i=2}^{t} [a_{i-1} \prod_{j=i}^{t} (c_j+1)] \\ \underline{B}_t &= \underline{b}_t + \sum_{i=2}^{t} [\underline{b}_{i-1} \prod_{j=1}^{t} (c_j+1)] \\ V_t &= v_t + \sum_{i=2}^{t} [v_{i-1} \prod_{j=i}^{t} (c_j+1)] \end{aligned} \tag{4}$$

The solution assumes that all groups of equal-aged young people start out even in the sense that $p_o = 0$ (an assumption which is surely true, if the origin $t = 0$ is pushed back far enough). The solution may be checked by substituting it back in equation (2). Substituting (3) in (1) gives --

$$r_t = (\alpha_t + \gamma_t A_{t-1}) + (\underline{\beta}_t' + \gamma_t \underline{B}_{t-1}')\underline{x} + (u_t + \gamma_t V_{t-1}) \tag{5}$$

Since there are no lagged variables in (3) and (5), these equations can be fit to the available cross section samples. Since the explanatory variables $\underline{x}$ are exogenous, ordinary least squares, or regression analysis, is an appropriate estimation technique; it will be used below. The various samples are for various age groups; hence the fits will give estimates of A_t, $\underline{B}_t$, $\alpha_t + \gamma_t A_{t-1}$, and $\underline{\beta}_t + \gamma_t \underline{B}_{t-1}$ for various values of t. The model yields some predictions about these sets of estimates.

It follows from (4) that --

$$\begin{aligned} \underline{B}_t &= \underline{B}_{t-1} + c_t \underline{B}_{t-1} + \underline{b}_t \\ V_t &= V_{t-1} + c_t V_{t-1} + v_t \end{aligned} \tag{6}$$

Since, by assumption, sign $\underline{b}_t$ is the same for all t and $c_t \geq 0$ for all t, it follows from (4) that sign $\underline{B}_t$ is the same for all t. These facts, plus the first of equations (6), imply that $|\underline{B}_t| > |\underline{B}_{t-1}|$ for all t. Finally, the second of equations (6) implies that the variance of V_t will be greater than the variance of V_{t-1} for all t, assuming no substantial negative covariances among the v_t,which seems reasonable. Thus, the following predictions may be made about the various age-group fits of equation (3) --

a. The coefficients (except the constant term) will have the same signs in each fit. (Sign $\underline{B}_t$ will be the same for all t.)

b. The absolute values of the coefficients (except the constant term) will get larger for more advanced age groups. ($|\underline{B}_t| > |\underline{B}_{t-1}|$ for all t.)

c. The error variance of the equation will get larger for more advanced age groups. [$\mathrm{Var}(V_t) > \mathrm{var}(V_{t-1})$.]

Very briefly and heuristically, these predictions may be rationalized as follows. Since p_t is a <u>cumulative</u> measure of school performance, then the associated coefficients and error variance in equation (3) may also be expected to *cumulate*; and this is essentially all the predictions say.

No such simple predictions can be made about the coefficients and error variance of (5). Nonetheless, since $\underline{B}_t$ and V_t are components of these coefficients and error variance, a tendency toward a similar pattern would not be surprising.

II. Estimates of the Model

A. The Data

The data used in the regressions presented below are taken from Table 5 of one of the 1960 Census special reports on education [4]. This table, constructed from a 5 per cent sample of the total U.S. population, gives data on school enrollment and relative progress in school of children living with one or both parents. The enrollment data are presented for each of seven age groups -- 5 years, 6 years, 7-9 years, 10-13 years, 14-15 years, 16-17 years, and 18-19 years. The progress data are not reported for the first two age groups because 5 and 6 year olds have not yet had time to establish a skip-flunk pattern. The age groups stop at age 19 because, after that age, too few children are still living with their parents; and thus the Census, which is taken on a family-by-family basis, does not contain data on both the children and their parents' education and income.

For each of the age groups, the data is cross-classified by---

a. 2 color categories

b. 2 sex categories

c. 3 rural-urban categories

d. 3 education of parents categories

e. 4 income of parents categories

Thus, for each age group, there are 2x2x3x3x4 = 144 mutually exclusive cells. Each cell will have a sub-population of children in it who are demographically homogeneous in the sense that they are all of the same color, same sex, same rural-urban status, etc. For each cell, the progress and enrollment variables p_t and r_t were calculated; and the 144 cells serve as the 144 observations for each."of the age group regressions presented below."

The $\underline{x}$ vector of explanatory variables for each observation, or cell, is a set of nine dummy variables describing the color, sex, etc. of the cell, as follows --

$$\underline{x} = \begin{pmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_9 \end{pmatrix}, \text{ where } \begin{cases} x_1 = 1 \text{ for non-whites, 0 for whites} \\ x_2 = 1 \text{ for females, 0 for males} \\ x_3 = 1 \text{ for persons living outside a central city but not on a farm, 0 otherwise} \\ x_4 = 1 \text{ for persons living on a farm, 0 otherwise} \\ x_5 = 1 \text{ if parent (father if living, otherwise mother) has 0 to 7 years of schooling, 0 otherwise otherwise.} \\ x_6 = 1 \text{ if parent has 8 to 11 years of schooling, 0 otherwise} \\ x_7 = 1 \text{ if family income is under \$3000, 0 otherwise} \\ x_8 = 1 \text{ if family income is from \$3000 to \$4999, 0 otherwise} \\ x_9 = 1 \text{ if family income is from \$5000 to \$6999, 0 otherwise} \end{cases}$$

Thus, the observation, or cell, being referred to when all the dummy variables are zero ($\underline{x} = \underline{0}$) is the one for white males living in a central city, whose parents have a high school or better education and a $7000 or better income. The value of the dependent variable for this "benchmark" group is equal to the constant term (plus a random error); that is, the constant term is a benchmark value. Thus, the coefficient of a particular dummy variable may be thought of as the deviation from the benchmark value caused by the characteristic associated with that dummy variable.

B. The Use of Weighted Regressions

All regressions in this paper are weighted regressions, using cell sizes as weights. The weighting, which is a standard procedure for such grouped-data situations, is meant to correct for heteroskedasticity of the regression error terms. The formal assumption is that the error variance σ_i^2 for the i-th observation (the i-th cell of children) in a regression is given by $\sigma_i^2 = \sigma^2/w_i$, where w_i is the cell size and σ^2 is the constant error variance associated with an individual child (a cell of one). This assumption leads, via standard least squares theory, to a weighted regression with the w_i the weights. (See, for instance, [2,pp.231-36].) The rationale for assuming $\sigma_i^2 = \sigma^2/w_i$ is that the error term for the i-th cell is a sample average of the error terms for the children in that cell; and the variance of a sample average is equal to the underlying population variance (σ^2 in this case) divided by the sample size (w_i in this case).

The validity of the relation $\sigma_i^2 = \sigma^2/w_i$ for a given regression may be checked using the results of the corresponding unweighted regression. This was done for the regressions presented in this paper. The estimated residuals from the unweighted regressions did indeed tend to run larger in absolute size for observations with small cell sizes, as $\sigma_i^2 = \sigma^2/w_i$ would predict. Hence the weighted regressions seem preferable to the unweighted. However, two things suggest that the exact form of the relation $\sigma_i^2 = \sigma^2/w_i$ is incorrect. First, the negative relation between the absolute size of the unweighted regression residuals and the cell sizes tended to wash out and even reverse for fairly large cell sizes (suggesting that the larger cells encompass more diverse types of children). Second, the estimated error variances from the weighted regressions are estimates of σ^2; and they were impossibly large. Since the dependent variables p_t and r_t never exceed one in absolute value, their error variances cannot exceed one; whereas the various weighted regression estimates of σ^2 exceeded one in every case, often by a great deal.

These facts suggest an error relationship of the form of $\sigma_i^2 = f(w_i)$, where $f(w_i)$ is negatively related to w_i over most of its range, but where $f(w_i)$ is not of the precise form $f(w_i) = \sigma^2/w_i$. No attempt was made here to estimate $f(w_i)$. This does not, however, mean that the weighted regression coefficients presented are bad estimates. The weighting does make a rough correction for heteroskedasticity. The coefficients are unbiased (barring other statistical difficulties), since heteroskedasticity does not cause bias. And finally, the bias in the estimates of coefficient standard deviations is in general upward; hence taking the reported standard deviations at face value leaves one on the "safe" side of the bias. (In a weighted regression, the weighting discounts the observations which are likely to have the largest residuals; therefore, weighting generally increases the R^2 and reduces the estimated standard deviations.)

The sum of squares minimized by a weighted regression is $\Sigma_i w_i (y_i - y_i^*)^2$ where y_i and y_i^* are the actual and predicted values of the dependent variable for the i-th observation. This suggests the following R^2 formula, where $\bar{y}$ is the sample mean of the y_i --

$$R^2 = 1 - \Sigma_i w_i (y_i - y_i^*)^2 / \Sigma_i w_i (y_i - \bar{y})^2$$

All R^2's reported below are computed according to this formula.

C. Regression Fits of Equation (3)

Table 2 presents a series of age group regressions of the dependent variable p_t on the explanatory variables $\underline{x}$ [that is, fits of equation (3)]. The R^2's indicate a fairly good level of explanation; and every category of explanatory variable (color, sex, urban-rural status, education of parents, and income of parents) is highly significant. Furthermore, the coefficient values bear out the predictions stated at the end of section I. With few exceptions, the coefficients of the successive age group regressions do in fact have the same signs and do in fact get larger in absolute value. The prediction at the end of section I about the successive error variances cannot be tested with these results, because the relationship $\sigma_i^2 = f(w_i)$, discussed in the preceding few paragraphs, is not known.

Inspection of the coefficients of individual variables suggest the following comments --

1. The parental education and income variables may be thought of as measuring the quality of students' education outside the school, which complements education in the school. The signs of the coefficients of these variables are negative because the benchmark group, with respect to which the dummy variables are defined, has parents in the highest education and income categories. Parental education appears to be more important than parental income.
2. The positive significance of the female dummy indicates that girls tend to do better in school than boys; and the coefficients are substantial in size. This is a surprisingly strong result in view of the mixed evidence from psychologists on sex differences in children's abilities. [1,pp.9-10].

Table 2. Regressions with the Progress Rate p_t the Dependent Variable

Age Group	Coefficients and (in Parentheses) Coefficient Standard Deviations of										R^2	Dependent Variable Mean
	Constant	Non-White Dummy	Female Dummy	Rural-Urban Dummies		Parents' Schooling Dummies		Parents' Income Dummies				
				Not Farm or Central	Farm	0-7 Years	8-11 Years	Less Than $3000	$3000 to $5000	$5000 to $7000		
7-9	.0458 (.0033)	.0236 (.0041)	.0158 (.0025)	-.0317 (.0029)	-.0268 (.0051)	-.0507 (.0038)	-.0046 (.0029)	-.0392 (.0041)	-.0112 (0036)	-.0049* (.0032)	.82	.010
10-13	.0320 (.0075)	-.0245 (.0097)	.0512 (.0057)	-.0456 (.0066)	-.0219* (.0113)	-.1463 (.0085)	-.0315 (.0066)	-.1034 (.0093)	-.0403 (.0083)	-.0165 (.0075)	.88	-.073
14-15	.0326 (.0092)	-.0480 (.0119	.0772 (.0070)	-.0634 (.0081)	-.0260 (.0130)	-.1966 (.0101)	-.0483 (.0083)	-.1342 (.0110)	-.0571 (.0103)	-.0287 (.0095)	.91	-.120
16-17	-.0091* (.0089)	-.1028 (.0118)	.0801 (.0068)	-.0458 (.0078)	-.0084* (.0125)	-.1784 (.0097)	-.0578 (.0080)	-.1287 (.0108)	-.0652 (.0100)	-.0317 (.0091)	.91	-.167
18-19	-.0844 (.0095)	-.1729 (.0118)	.0825 (.0073)	-.0701 (.0082)	-.0556 (.0133)	-.2558 (.0103)	-.1101 (.0089)	-.1623 (.0116)	-.0973 (.0108)	-.0530 (.0099)	.95	-.369

* Coefficient less than 1.96 times its estimated standard deviation. (The value 1.96 is the critical value for a standard t-test using either a .05 level and a two-tail test or a .025 level and a one-tail test.)

3. The coefficients of the non-white dummy are negative and significant, as would be expected. This non-white effect is measured with other variables held equal. It should be noted that other variables are typically not equal for non-white children, who are very likely to have low parental education and income also working against them. Similar all-else-not-equal considerations apply to judgments about the orders of magnitude of all the coefficients.
4. The coefficients for the two rural-urban dummies have the same sign and very rough order of magnitude in the various regressions; this is perhaps because the not-farm-or-central-city residence category is made up largely of rural-type population (country towns, small cities, and rural non-farm), which is similar in character to farm population.

D. Regression Fits of Equation (5)

Table 3 presents regressions of the enrollment rate r_t on the explanatory variables $\underline{x}$.[1] The R^2's indicate a fairly good level of explanation and the coefficients are in general highly significant. It may be expected that the various institutional constraints faced by the various age groups of children will influence the regressions. The five-year-olds are too young to fall under the compulsory school attendance laws; and those that do attend school typically do so at their own, rather than the public's, expense. This is also true to some extent for the six-year-olds. For the 7-9 and 10-13 year-olds, however school is compulsory and free. For the 14-15 and 16-17 year-olds, schooling is typically still free; but the compulsory attendance laws either no longer apply or are more difficult to enforce; and the opportunity costs from other occupations start to rise. Finally, the 18-

Table 3. Regressions with the Enrollment Rate r_t the Dependent Variable

Age Group	Coefficients and (in parentheses) Coefficient Standard Deviations of										R^2	Dependent Variable Mean
	Constant	Non-White Dummy	Female Dummy	Rural Urban Dummies		Parents' Schooling Dummies		Parents' Income Dummies				
				Not Farm or Central City	Farm	0-7 Years	8-11 Years	Less than $3000	$3000 to $5000	$5000 to $7000		
5	.669 (.007)	.0441 (.0091)	.0053* (.0056)	-.1474 (.0064)	-.2659 (.0118)	-.1372 (.0087)	-.0458 (.0064)	-.1636 (.0093)	-.1241 (.0081)	-.0696 (.0073)	.94	.401
6	.942 (.005)	.0103* (.0064)	.0034* (.0039)	-.0499 (.0044)	-.1113 (.0081)	-.0998 (.0061)	-.0247 (.0045)	-.1061 (.0065)	-.0689 (.0056)	-.0355 (.0050)	.93	.793
7-9	.984 (.001)	-.0078 (.0012)	.0005* (.0007	.0032 (.0009)	-.0034 (.0015)	-.0152 (.0011)	-.0034 (.0009)	-.0158 (.0012)	-.0055 (.0011)	-.0023 (.0010)	.90	.970
10-13	.983 (.001)	-.0070 (.0011)	.0010* (.0007)	.0056 (.0009)	.0052 (.0013)	-.0163 (.0010)	-.0053 (.0008)	-.0134 (.0011)	-.0042 (.0010)	-.0015* (.0009)	.90	.972
14-15	.973 (.002)	-.0085 (.0028)	.0014* (.0016)	.0083 (.0019)	.0089 (.0030)	-.0564 (.0023)	-.0181 (.0019)	-.0365 (.0026)	-.0117 (.0024)	-.0029* (.0022)	.92	.938
16-17	.910 (.005)	-.0037* (.0068)	.0178 (.0039)	.0295 (.0045)	.0583 (.0072)	-.1706 (.0056)	-.0724 (.0046	-.0764 (.0062)	-.0336 (.0058)	-.0127 (.0052)	.93	.824
18-19	.624 (.009)	.0385 (.0117)	-.0777 (.0072)	.0159 (.0081)	.0089* (.0131)	-.2534 (.0101)	-.1827 (.0087)	.0082* (.0115)	.0047* (.0107)	.0183* (.0097)	.87	.474

* Coefficient less than 1.96 times its estimated standard deviation. (The value 1.96 is the critical value for a standard t-test using either a .05 level and a two-tail test or a .025 level and a one-tail test).

19 year-olds are of beginning college age; and schooling is typically no longer free.

Inspection of the coefficients suggests the following comments --

1. For the four age groups 7-9, 10-13, 14-15, and 16-17 (which face roughly the same free-compulsory school situation), a rough pattern of cumulating coefficient values is observed in the successive regressions. This is the same pattern as observed for the p_t-regressions of Table 2; and the theoretical rationale suggested in section I may apply here as well.
2. By far the most important explanatory variables are the education of parents dummies, particularly for the important last three age groups, which cover the years when more than half the students drop out of school.
3. The coefficients of the income dummies behave predictably for all except the last age group, where they become insignificant. This is a puzzling result, since income would seem to be particularly important for the age group which is first facing college expenses.
4. The non-white dummy is generally significant, taking a positive sign for the youngest and oldest age groups, and a negative sign otherwise. Since the youngest and oldest age groups bear much of their schooling cost personally, this sign pattern suggests that,

[1] Closely related regressions may be found in Chapters 24 and 25 of [3]. In that study, the dependent variable is an index of years of schooling completed; the observations are for individuals rather than groups; and the list of explanatory variables is much more detailed.

other variables constant, non-whites may be more willing than whites to sacrifice other expenditures for school expenditures. Perhaps this is because a non-white is, relative to his social context, richer than a white with the same education and income; and thus he is better able to afford extra educational expenditures for his children. (Another hypothesis is that non-whites in the 18-19 age group have a higher enrollment rate, other variables constant, because proportionally more of them have fallen behind scholastically and are still finishing high school. A test of this hypothesis can be gotten by adding the relative progress variable p_t as an additional explanatory variable in the regressions. If, after controlling on p_t, the sign pattern of the non-white dummy still remains, then the hypothesis is only a partial explanation at best. This turns out to be the case, as the regressions of Table 4 below will show.)

5. The female dummy is significant for only the 16-17 and 18-19 age groups, with a positive and negative coefficient for the two groups, respectively. The positive sign for the 16-17 age group (terminal high school years) is perhaps due to a girl's lesser impatience to quit school and get a job; while the negative sign for the 18-19 age group (beginning college years) is perhaps due to society's relative reluctance to invest a college education in a prospective housewife.

6. The two rural-urban status dummies have the same sign and general order of magnitude in the various regressions. This is the same pattern as observed on Table 2, and the same suggested rationale applies here. The negative significance of these dummies for the

5 and 6 age groups is perhaps due to the difficulty of getting pre-school age rural children to a kindergarten or other pre-school. A convincing rationale for the positive significance of these dummies for the older age groups seems difficult to find.

E. Supplementary Regressions

In the original statement of the model in section I, the enrollment rate r_t was assumed to depend partly on the lagged progress variable p_{t-1} as follows -- $r_t = \alpha_t + \underline{\beta}'_t\underline{x} + \gamma_y p_{t-1} + u_t$, where γ_t was hypothesized to be positive. Since data on the lagged variable p_{t-1} was unavailable, a solution form was found which expressed r_t as a function of $\underline{x}$ alone [equation (5)]. Unfortunately, in finding this solution form, the ability to test the hypothesis $\gamma_t > 0$ was lost; and the regressions of Table 3 do not in fact provide such a test. However, there is another equation for r_t available from the model, one which does not involve lagged variables and does not lose the ability to test the hypothesis $\gamma_t > 0$. Solving equation (2) for p_{t-1} as a function of p_t and $\underline{x}$, and substituting this result in equation (1) gives --

$$(7) \qquad r_t = [\alpha_t - \gamma_t a_t/(c_t+1)] + [\underline{\beta}'_t - \gamma_t \underline{b}'_t/(c_t+1)]\,\underline{x} + [\gamma_t/(c_t+1)]\,p_t + [u_t - \gamma_t v_t/(c_t+1)]$$

which includes no lagged variables and is thus estimable with the available data. Table 4 presents regressions of this form. (Since, in the model, p_t is determined independently or r_t, then ordinary least squares, or regression analysis, is still an appropriate estimation technique.) In terms of these regressions, the hypothesis $\gamma_t > 0$ becomes the hypothesis that the coefficient of p_t is greater than zero. In four of the five regressions, the coefficient of p_t is indeed significantly positive (by a standard t-test at any conventional significance level). This provides rough confirmation of the hypothesis $\gamma_t > 0$.

Table 4. Supplementary Regressions with the Enrollment Rate r^t The Dependent Variable

Age Group	Coefficients and (in parentheses) Coefficient Standard deviations of											R^2
	Constant	Non-White Dummy	Female Dummy	Rural-Urban Dummies		Parents' Schooling Dummies		Parents' Income Dummies			Progress Rate p_t	
				Not Farm or Central City	Farm	0-7 Years	8-11 Years	Less than $3000	$3000 to $5000	$5000 to $7000		
7-9	.977 (.001)	-.0117 (.0012)	-.0021 (.0007)	.0084 (.0010)	.0077 (.0014)	-.0069 (.0015)	-.0027 (.0007)	-.0094 (.0013)	-.0037 (.0009)	-.0015* (.0008)	.162 (.022)	.93
10-13	.981 (.001)	-.0054 (.0010)	-.0022 (.0007)	.0085 (.0008)	.0066 (.0011)	-.0072 (.0015)	-.0034 (.0007)	-.0069 (.0013)	-.0017* (.0009)	-.0005* (.0008)	.062 (.009)	.93
14-15	.970 (.002)	-.0030* (.0026)	-.0074 (.0020)	.0155 (.0020)	.0118 (.0027)	-.0342 (.0040)	-.0127 (.0019)	-.0214 (.0033)	-.0053 (.0023)	.0004* (.0020)	.113 (.018)	.94
16-17	.913 (.005)	.0216 (.0077)	-.0019* (.0051)	.0408 (.0046)	.0604 (.0066)	-.1266 (.0095)	-.0581 (.0049)	-.0446 (.0081)	-.0176 (.0060)	-.0049* (.0050)	.247 (.045)	.94
18-19	.630 (.012)	.0510 (.0188)	-.0836 (.0101)	.0210 (.0100)	.0129* (.0140)	-.2349 (.0241)	-.1747 (.0128)	.0199* (.0180)	.0118* (.0136)	.0222 (.0108)	.072* (.0854)	.87

*. Coefficient less than 1.96 times estimated standard deviation. (The value 1.96 is the critical value for a standard t-test using either a .05 level and a two-tail test or a .025 level and a one-tail test.)

The most important explanatory variables in the various regressions presented are generally the education of parents variables. These variables refer to the father's education, if he is living, otherwise the mother's. It is informative to refit the regressions, using separate variables for the father's education and the mother's education; the point is to see if one or the other parent exerts a greater influence on the child. The required data may be found on Table 4 of the same 1960 Census special report on education [4]. This table presents, for each age group of children living with both parents, enrollment and relative progress data cross-classified by --

a. 2 color categories

b. 2 sex categories

c. 3 rural-urban categories

d. 10 education of father and mother categories

Thus, there are 2X2X3X10 = 120 mutually exclusive cells, which serve as the observations for the regressions presented on Table 5. Since no data on the incomes of parents were available, these regressions are only roughly comparable to the previous regressions; and results are reported only for the two age groups 16-17 and 18-19. In these regressions, the education of parents data were translated into two quantitative variables, defined as years of schooling completed by mother and by father; Table 6 shows how the translation was made.

Table 5 suggests that the educations of a child's father and mother are of roughly equal importance in determining p_t and r_t. Though the coefficient of the father's education variable is larger in all four regressions on Table 5, the differences are not substantial. They could easily be due to specification bias; the father's education variable may be picking up much of the effect of the excluded income variables.

Table 5. Regressions Measuring Seperate Effects of Father's and Mother's Educations

Dependent Variable and Age Group	Coefficients and (in Parentheses Coefficient Standard Deviations of							R^2	Dependent Variable Mean
	Constant	Non-white Dummy	Female Dummy	Rural-Urban Dummies		Years of Schooling Completed by			
				Not Farm or Central City	Farm	Father	Mother		
Progress Rate									
16-17	-.436 (.018)	-.1185 (.0146)	.0773 (.0080)	-.0571 (.0092)	-.0472 (.0140)	.0177 (.0013)	.0164 (.0014)	.88	-.131
18-19	-.709 (.017)	-.1882 (.0129)	.0780 (.0076)	-.0837 (.0086)	-.1030 (.0132)	.0269 (.0013)	.0225 (.0013)	.95	-.299
Enrollment Rate									
16-17	.538 (.012)	-.0029* (.0094)	.0167 (.0051)	.0207 (.0060)	.0305 (.0091)	.0166 (.0009)	.0134 (.0009)	.90	.864
18-19	.089 (.026)	.0854 (.0193)	-.0814 (.0115)	.0129* (.0129)	.0064* (.0198)	.0270 (.0020)	.0157 (.0020)	.84	.545

* Coefficient less than 1.96 times estimated standard deviation. (The value 1.96 is the critical value for a standard t-test using either a .05 level and a two-tail test or a .025 level and a one-tail test.)

Table 6. Translation of Census Parental Education Categories into Quantitative Regression Variables

Years of Schooling Completed			
Categories on Census Table		Values Assigned to Regression Variables	
Father	Mother	Father	Mother
0-7	0-7	5	5
0-7	8 and up	5	11
8-11	0-7	9.5	5
8-11	8-11	9.5	9.5
8-11	12 and up	9.5	14
12	0-11	12	7.5
12	12	12	12
12	13 and up	12	15
13 and up	0-12	15	8.5
13 and up	13 and up	15	15

III. A Concluding Remark

Judging by coefficient estimates, R^2's, and the like, the model seems fairly successful in explaining enrollment rates and relative progress rates. However, this success of the $\underline{x}$ variables in explaining r_t and p_t is probably more discouraging than encouraging from a policy viewpoint. The $\underline{x}$ variables measure characteristics of children's home environment which are almost completely outside the control of the children themselves; so the strength of $\underline{x}$'s explanatory power is in a sense a measure of a lack of equal opportunity the children face. Furthermore, the $\underline{x}$ variables are mostly outside the control of policy makers who might wish to influence p_t and r_t; so the strength of $\underline{x}$'s explanatory power is in a sense also a measure of the difficulty of policy formulation.

Madison, Wisconsin
August, 1966

References

1. Gerald S. Lesser, Gordon Fifer, and Donald H. Clark, "Mental Abilities of Children from Different Social-Class and Cultural Groups," Monographs of the Society for Research in Child Development, Vol. 30, No. 4, serial No. 102, 1965 (University of Chicago Press).

2. Arthur S. Goldberger, Econometric Methods (New York: Wiley, 1964).

3. James N. Morgan, Martin H. David, Wilbur J. Cohen, and Harvey E. Brazer, Income and Welfare in the United States (New York: McGraw-Hill, 1962).

4. United States Census of Population: 1960, Subject Reports, School Enrollment, Final Report PC(2) - 5A, (Washington, D.C.: U.S. Government Printing Office, 1964).