R E P O R T R E S U M E S

ED 012 091 EA 000 369

DETERMINANTS OF SCHOOL ENROLLMENT AND SCHOOL PERFORMANCE.

BY- CONLISK, JOHN
WISCONSIN UNIV., MADISON

EDRS PRICE MF-$0.09 HC-$1.08 27P.

DESCRIPTORS- AGE GROUPS, *MODELS, RURAL URBAN DIFFERENCES, *PARENTAL BACKGROUND, FAMILY INCOME, ACCELERATION, GRADE REPETITION, STATISTICAL ANALYSIS, *STUDENT ENROLLMENT, *ACADEMIC PERFORMANCE, MADISON

DEMOGRAPHIC VARIABLES (X) DESCRIBING AGE, COLOR, SEX, RURAL-URBAN STATUS, EDUCATION OF PARENTS, AND INCOME OF PARENTS ARE USED AS EXOGENOUS VARIABLES TO EXPLAIN SCHOOL ENROLLMENT RATES (R)--THE FRACTION OF A GROUP WITHIN THE SCHOOL AGE POPULATION ENROLLED IN SCHOOL--AND RELATIVE PROGRESS (P)--THE FRACTION OF A GROUP OF STUDENTS WHO ARE AHEAD OF THEIR AGE GROUP MINUS THE FRACTION WHO ARE BEHIND. A MODEL IS DEVELOPED AND TESTED STATISTICALLY, USING DATA OF ONE OF THE 1960 CENSUS SPECIAL REPORTS ON EDUCATION. THE RESULTS SHOW THAT THE X VARIABLES, ESPECIALLY THE PARENT'S EDUCATION VARIABLE, ARE SUCCESSFUL IN EXPLAINING R AND P. THESE X VARIABLES ARE, HOWEVER, ALMOST COMPLETELY OUTSIDE THE CONTROL OF THE CHILDREN THEMSELVES SO THAT TO SOME EXTENT THIS IS A MEASURE OF A LACK OF EQUAL OPPORTUNITY. IN ADDITION, THESE X VARIABLES ARE OUTSIDE THE CONTROL OF POLICY MAKERS WHO MIGHT WISH TO INFLUENCE P AND R. (HW)

Determinants of School Enrollment and School Performance*

John Conlisk
Institute for Research on Poverty
University of Wisconsin

Introduction

This paper analyzes 1960 Census data on school enrollment and school performance of age 5 to 19 children in the United States. School enrollment here refers to whether a child is or is not enrolled in school; and school performance refers to whether an enrolled child is behind, with, or ahead of his age group in years of schooling completed. Demographic variables describing age, color, sex, rural-urban status, education of parents, and income of parents are used to explain variation in school enrollment and performance across the school age population. Since these explanatory variables are almost completely outside the control of the children themselves, their explanatory power measures the lack of equal educational opportunity the children face. Since the explanatory variables are also largely outside the short-run control of would-be policy-makers, their explanatory power also indicates to some extent the difficulty of educational policy to improve school enrollment and performance. Nonetheless, the importance of the parental income variable is somewhat encouraging evidence for income-supplementing policies for the poor. If supplements to poor parents' incomes

*The research reported here 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. Thanks are due to Harold Watts for helpful comments on an earlier draft of the paper.

tend to improve their children's school enrollment and performance, then the supplements will tend to have desirable second generation effects on poverty.

The school enrollment variable used in the theoretical model below is defined as the probability that a child of given characteristics will be enrolled in school. The school performance variable used is based on two probabilities--(i) the probability that an enrolled child has skipped ahead of his age group in years of school completed, and (ii) the probability that he has flunked or otherwise fallen behind his age group in years completed. For simplicity, the two probabilities will not be treated separately; rather the model will deal with their difference (the first minus the second). Thus, the school performance variable will be a "net skip probability" (or the negative of a "net flunk probability," which is perhaps more descriptive, since flunks far outnumber skips). A simple two equation model will be specified which makes the enrollment variable and the school performance variable depend on the demographic explanatory variables (age, sex, etc.); and the model will be fit via ordinary regression analysis to the 1960 Census data.

Table 1 presents the cut-off points used by the Census in deciding when a child is behind or ahead of his age group in years of schooling completed. It should be stressed that the school performance variable to be used is only a very rough indicator, 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 and systematically from school to school. Second, neither skipping nor flunking is an automatic consequence of a superior or inferior performance by a child. Nonetheless, any school performance data collected on the complete scale of the U. S. Census seems to deserve at least as much attention as it is given here.

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 [4]. *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.

1. The Model

Let r_t be the probability that a child of age t is enrolled in school; and, given that he is enrolled, let p_t be the difference of (i) the probability that he is ahead of his age group and (ii) the probability that he is behind his age group. Let $\underline{x}$ be a column vector of demographic variables describing the child's characteristics, where $\underline{x}$ includes variables for color, sex, rural-urban status, education of parents, and income 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. The model is as follows--

$$r_t = \alpha_t + \underline{\beta}_t'\underline{x} + \gamma_t p_{t-1} + u_t \qquad (\gamma_t > 0) \tag{1}$$

$$\Delta p_t = a_t + \underline{b}_t'\underline{x} + c_t p_{t-1} + v_t \qquad (\text{sign } \underline{b}_t \text{ same for all t; } c_t \geq 0) \tag{2}$$

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 next to the equations.

The model traces out the values of r_t and p_t for a child as his age t progresses. Equation (1) states that, at age t, the child's enrollment probability is a linear function of the demographic characteristics $\underline{x}$ and the previous period's performance p_{t-1}, plus an error. The performance variable p_{t-1} is included in equation (1) with a positive coefficient, since students who have done well in school in the past seem more likely to continue their education.

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 appropriateness 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 Census data used below in fitting the model is all measured at one point in time (1960). Hence lagged values of variables are not available, and the model cannot be fit as it stands. However, since equation (2) is simply a first order linear difference equation in p_t (complicated by an error term and by parameters which change with t), it can be solved for p_t as a function of $\underline{x}$. The solution is--

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

where--

$$A_t = a_t + \sum_{i=2}^{t} [a_{i-1} \prod_{j=i}^{t}(c_j+1)]$$

$$\underline{B}_t = \underline{b}_t + \sum_{i=2}^{t} [b_{i-1} \prod_{j=i}^{t}(c_j+1)] \tag{4}$$

$$V_t = v_t + \sum_{i=2}^{t} [v_{i-1} \prod_{j=i}^{t}(c_j+1)]$$

It is assumed in this solution that a child starts out at age t = 0 even with his age group in terms of skipping ahead or flunking behind; that is, it is assumed that $p_o = 0$. 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 Census data. Since the explanatory variables $\underline{x}$ are exogenous,

ordinary least squares, or regression analysis, is an appropriate estimation technique; it will be used below. Since the coefficients in (3) and (5) are specific to various age levels (various values of t), the equations will be fit for each of a series of 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--

$$\underline{B}_t = \underline{B}_{t-1} + c_t \underline{B}_{t-1} + \underline{b}_t$$

(6)

$$V_t = V_{t-1} + c_t V_{t-1} + v_t$$

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. [$Var(V_t) > var(V_{t-1})$.]

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 <u>cumulate</u>; 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.

2. The Data

The data come from one of the 1960 Census special reports on education [4]. Table 5 of this report, constructed from a five percent sample of the total U. S. population gives data on school enrollment for each of seven age groups of children--5 years, 6 years, 7-9 years, 10-13 years, 14-15 years, 16-17 years, and 18-19 years. Data on skip-flunk patterns are also presented for each of the age groups except the 5-year-olds and 6-year-olds, who 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 matched data on children and parents.

For each of the age groups, the children are cross-classified by--

a. 2 racial categories
b. 2 sex categories
c. 3 rural-urban categories
d. 3 education of parents categories
e. 4 income of parents categories

Then, for each age group there are 2X2X3X3X4 = 144 mutually exclusive cells containing data on enrollment and skip-flunk patterns. The number of children in a given cell will be referred to as the cell size. These 144 cells serve as the 144 observations in the regressions of r_t on $\underline{x}$ and p_t on $\underline{x}$ for each age group (each value of t). Though r_t is an unknown probability and p_t a difference of unknown probabilities, r_t and p_t can nonetheless be approximated with good accuracy for a given one of the 144 cells of a given age group by the following empirical definitions--

$$r_t = \frac{\text{students enrolled in school}}{\text{total students}}$$

$$p_t = \frac{\text{students enrolled and ahead of their age group}}{\text{total students enrolled}} - \frac{\text{students enrolled and behind their age group}}{\text{total students enrolled}}$$

The ratios on the right of these equalities are sample proportions; and it is well known that the error variance in measuring a true probability P by a sample proportion is P(1-P)/m, where m is the sample size (here the cell size) on which the measure is based. Since the cell size m is almost without exception very large for all the 144 cells, or observations, of a given regression, then r_t and p_t will be gotten with good accuracy by these measures.

The vector $\underline{x}$ of explanatory variables to be used in the regressions of r_t on $\underline{x}$ and p_t on $\underline{x}$ is defined by the following series of zero-one, or dummy variables.

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

The vector $\underline{x}$ of dummy variables may thus take on 144 possible values corresponding to the 144 cells for a given age group.[1] It follows from the zero-one nature of

---

[1] In making this count note that the x_i come in groups. The five groups (x_1), (x_2), (x_3,x_4), (x_5,x_6), and (x_7,x_8,x_9) represent color, sex, rural-urban status, parental education, and parental income, respectively. No more than one x_i in a given group can take on the value one for a given observation. Taking account of this constraint, the five groups listed may take on 2, 2, 3, 3, and 4 possible values, respectively. Thus, the complete vector $\underline{x}$ may take on 2X2X3X3X4 = 144 values.

the x_i that the regressions using $\underline{x}$ as the vector of explanatory variables will be equivalent to five-way analyses of variance where the five classifications are color, sex, rural-urban status, parental education, and parental income.

The observation described by the condition that all the x_i 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 expected value of the dependent variable for this observation is simply the constant term of the regression. It is convenient to think of this observation as a benchmark observation, and to think of the constant term as a benchmark value. Then the coefficient of a given x_i in a regression may be thought of as a deviation from the benchmark value caused by the characteristic associated with that variable.

3. The Use of Weighted Regression

Preliminary versions of the regressions to be presented below indicated a serious heteroskedasticity (unequal error variance) problem. It was found that the absolute values of the 144 residuals for a given regression tended to be negatively related to the corresponding cell sizes. That is, the residual error variance appeared to be negatively related to the cell size. Partly this problem might be traced to a decline of the measurement error in the dependent variables r_t and p_t as the cell size increases (as discussed in the last section). However, this appeared not to be a sufficient explanation; much of the problem appeared to be due to a genuine heteroskedasticity in the error terms of the underlying model. An easy way to take account of a negative relation between the error variance and the cell size is to assume the relation takes the exact form $\sigma_i^2 = \sigma^2/w_i$, where σ_i^2 is the error variance of the i-th observation in the regression, σ^2 is a constant, and w_i is the cell size of the observation. This relation leads, via standard least squares theory, to a straightforward weighted regression with the w_i the weights (see for instance [2, pp. 231-36]). All regressions reported below will be such weighted regressions. This method of handling heteroskedasticity is a compromise with computational ease, since a relation between σ_i^2 and w_i other than $\sigma_i^2 = \sigma^2/w_i$ might well be more faithful to the data, though more difficult computationally. However, it is comforting to recall that heteroskedasticity by itself does not cause bias in estimated regression coefficients.

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 weighted 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.

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

Coefficients and (in Parentheses) Coefficient Standard Deviations of

Age Group	Constant	Non-White Dummy	Female Dummy	Rural-Urban Dummies		Parents' Schooling Dummies		Parents' Income Dummies			R^2	Dependent Variable Mean
				Not Farm or Central	Farm	0-7 Years	8-11 Years	Less than $3000	$3000-$5000	$5000-$7000		
7-9	.0458 (.0033)	.0236 (.0041)	.0158 (.0025)	-.0317 (.0029)	-.0268 (.0051)	-.0507 (.0038)	-.0046 (.0029)	-.0392 (.0041)	-.0012 (.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 (.0018)	.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.)

4. Regression Fits of Equation (3)

Table 2 presents a series of age group regressions of the progress rate p_t on the explanatory variables $\underline{x}$ [that is, fits of equation (3)]. Every category of explanatory variable (color, sex, urban-rural status, education of parents, and income of parents) is highly significant; and the coefficient values bear out the predictions stated at the end of section I. That is, 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 that there would be an increase with age in the variance of the error term of equation (3) did not seem reasonably testable with the regressions reported here, due to the serious heteroskedasticity problem discussed in the last section. The R^2's indicate a fairly good level of explanation. However, the R^2's should be interpreted with caution. They measure the regression's ability to predict <u>probabilities</u> involved in a child's school behavior, not the behavior itself. The behavior itself for a single child may still be quite unpredictable (just as an accurate knowledge of the probability of flipping heads with a coin does not imply an ability to predict accurately the outcome of a single flip). On the other hand, in making predictions about a sizable group of children, an accurate notion of probabilities does imply an ability to predict accurately the percentages of the group that will behave in given ways.

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

1. The parental income and education variables have a positive and highly significant influence on school performance p_t. (The coefficients themselves are negative because the benchmark group, with respect to which the dummy variables are defined, has parents in the highest income and education category.) The importance of this intergenerational effect may be explained a number of ways. Partly, the parental variables may measure the quality of home education,

which complements school education in determining school performance. Partly, the parental variables may stand as proxies for attitudinal variables which determine how hard children try in school. Partly, the parental variables may stand as proxies for innate intelligence of parents, which is to some extent genetically bequeathed to children. And so on. The diversity of effects for which the parental variables may stand makes it hard to say what effect income supplements to poor parents would have on their children. Taking the coefficients at simple face value suggests that a poverty policy which pushed all families in the less than $3000 income category into the $3000 to $5000 category would have a significant impact on the children's school performance.

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. (See for instance [1, pp. 9-10] and references there.)

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.

The systematic pattern of cumulating coefficient sizes on the regressions of Table 2 makes it possible to combine all five regressions into a single simplified regression based on all 5(144) = 720 observations. Suppose as a first simplification that the three rural-urban status categories are reduced to two--central city and rural (defined as not central city). This simplification, which is suggested by the similarity of coefficients of the two rural-urban dummies on Table 2, allows rural-urban status to be handled by a single dummy, call it x_{RUR},

which equals 0 for central city children, one otherwise. As a second simplification, suppose education and income of parents are handled by continuous variables E and Y instead of by two sets of dummy variables, where E is the number of years of schooling completed by a child's parent (father if living, otherwise mother) and Y is the income of the child's family.[3] With these simplifications, the relation between p_t and the explanatory variables for a given age group might be specified as--

$$p_t = \beta_o + \beta_1 x_{NW} + \beta_2 x_{FEM} + \beta_3 x_{RUR} + \beta_4 \ln(E) + \beta_5 \ln(Y)$$

Here x_{NW} and x_{FEM} are the dummy variables for non-white-ness and female-ness; and E and Y have been included in logarithmic form because preliminary results suggested it. It is known from Table 2 that the coefficients β_i in such a relation get larger for successive age groups of children. Suppose this cumulative effect is approximated by making each of the β_i a linear function of age, call it A, so that $\beta_i = \beta_{i0} + \beta_{i1}A$. Then the relation becomes--

$$p_t = (\beta_{00} + \beta_{01}A) + (\beta_{10} + \beta_{11}A)x_{NW} + (\beta_{20} + \beta_{21}A)x_{FEM}$$
$$+ (\beta_{30} + \beta_{31}A)x_{RUR} + (\beta_{40} + \beta_{41}A)\ln(E) + (\beta_{50} + \beta_{51}A)\ln(Y)$$

It may be seen that there are 12 β_{ij} to be estimated in this relation and that, if one multiplies through the parentheses, there will be just enough terms to estimate the 12 β_{ij} by ordinary regression analysis. Such a regression was in fact fit, where the values (8, 11.5, 14.5, 16.5, 18.5) were assigned to A for each of the five age groups respectively; and where the observations in the regression were weighted by the corresponding cell sizes.

---

[3] Since the parental education data come in discrete categorizations, the following assumptions were used in constructing E. A parent with 0-7 years of schooling was assigned an E-value of E = 4; a parent with 8-11 years was assigned E = 9.5; and a parent with 12 or more years was assigned E = 13. Similarly, in constructing Y, the parental income categories 0-3000, 3-5000, 5-7000, and 7000+ were assigned Y = 1500, Y = 4000, Y = 6000, and Y = 10000 respectively.

The regression result, based on 720 observations, is--

$$p_t = -[.119 + .0985\ A] + [.0605 - .0162\ A]\ x_{NW} \quad (.050)\ (.0071) \quad (.0018)\ (.0017)$$

$$+ [.00512 + .00758\ A]\ x_{FEM} - [.0309 + .0019\ A]\ x_{RUR} \quad (.00712)\ (.00102) \quad (.0080)\ (.0011)$$

$$+ [.0205 + .0158\ A]\ \ln(E) + [.0168 + .00514\ A]\ \ln(Y) \quad (.0090)\ (.0012) \quad (.0060)\ (.00084)$$

The R^2 for this regression is .87, which compares fairly well with a pooled R^2 of .94 for all five regressions on Table 2.

5. Regression Fits of Equation (5)

Table 3 presents regressions of the enrollment rate r_t on the explanatory variables $\underline{x}$.[4] 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 18-19 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 age groups which face the same free-compulsory school situation (7-9, 10-13, 14-15, and 16-17), 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 at the end of section 1 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] Closely related regressions may be found in Chapters 24 and 25 of [3]. There 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.

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

Coefficients and (in parentheses) Coefficient Standard Deviations of

Age Group	Constant	Non-White Dummy	Female Dummy	Rural Urban Dummies		Parents' Schooling Dummies		Parents' Income Dummies			R^2	Dependent Variable Mean
				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)	-.013⅓ (.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* (.0015)	.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 of a .025 level and a one-tail test).

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, 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 the same, 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, it suggests that 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 the general order of magnitude in the various regressions. This is the same pattern as observed in 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.

6. Supplementary Regressions

The paper will be concluded with several additional regressions bearing on some minor points. In the original statement of the model in Section I, the enrollment probability r_t was assumed to depend partly on the lagged progress variable p_{t-1} as follows--$r_t = \alpha_t + \underline{\beta}_t'\underline{x} + \gamma_t p_{t-1} + u_t$, 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--

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

which includes no lagged variables and is thus estimable with the available data. Table 4 presents regressions of this form. (Since p_t is determined in the model independently of r_t, then ordinary least squares, or regression analysis is still an appropriate estimation technique.) In terms of these regressions, the hypothesis that $\gamma_t > 0$ becomes the hypothesis that the coefficient of p_t is greater than zero (under the apparently safe assumption that the presumably positive parameter c_t is at least greater than -1.) 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 $\gamma_t > 0$.

The most important explanatory variables in the various regressions presented were usually the education of parents variables. These variables refer to the

Table 4. Supplementary Regressions with r_t the Dependent Variable

Coefficients and (in parentheses) Coefficient Standard Deviations of

Age Group	Constant	Non-White Dummy	Female Dummy	Rural-Urban Dummies		Parents' Schooling Dummies		Parents' Income Dummies			Progress Rate p_t	R^2
				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)	.163 (.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)	-.0046 (.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* (.085)	.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.)

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 school performance 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. 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 variable.

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

Coefficients and (in Parentheses) Coefficient Standard Deviations of

Dependent Variable and Age Group	Constant	Non-white Dummy	Female Dummy	Rural-Urban Dummies		Years of Schooling Completed by		R^2	Dependent Variable Mean
				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)	.0730 (.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

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).