University of Wisconsin-Madison

Institute for Research on Poverty

Discussion Papers

Anita A. Summers
Barbara L. Wolfe

IMPROVING THE USE OF EMPIRICAL RESEARCH AS A POLICY TOOL: AN APPLICATION TO EDUCATION

DP #580-79

Improving the Use of Empirical Research as a Policy Tool:

An Application to Education

Anita A. Summers
Associate Chairperson
Public Management Unit
Wharton School
University of Pennsylvania

Barbara L. Wolfe
Assistant Professor
Departments of Economics and Preventive Medicine
Institute for Research on Poverty
University of Wisconsin-Madison

November 1979

We are indebted to Lawrence H. Summers for his many suggestions and interest, and to the School District of Philadelphia for its willingness to allow us access to the detailed data.

ABSTRACT

In economics, as in other fields, different, equally competent researchers give different answers to policy-makers, leading to caution and disaffection with empirical research. A major underutilized tool is replication. We explore the replication process in attempting to shed light on the question--which school resources help the learning of junior high school age students?

Three forms of replication are used to test the "theory" derived from mining the data on 553 eighth grade students in 42 schools in Philadelphia. We examine (1) a larger, much less disaggregated data set (1541 students in 52 schools), (2) a randomly selected, stored sample (465 students in 50 schools), and (3) a modification of the larger sample, using a missing variable technique to develop estimators for the unsatisfactorily aggregated data.

Interestingly enough, most "answers" were consistent throughout the replications. No matter how one looks at the data, the conclusion that schools are not impotent emerges: organizational and teacher characteristics affect learning growth, and the negative effects of race and income can be traced through their impact on the absorption of particular school inputs.

We conclude that education production coefficients can contribute more to the policy reviews associated with budget cutbacks and declining enrollments if efforts are made to improve their reliability by using pupil-specific, longitudinal data and by replicating. Furthermore, encouragement needs to be given--in education, as in other fields where large quantities

of public resources are spent--to developing large, readily available data bases which are disaggregated, longitudinal, and have wide geographic coverage. Such data would encourage the use of the important tools of replication.

Improving the Use of Empirical Research as a Policy Tool: An Application to Education

The process by which knowledge is accumulated in economics is not clear. As in other fields, different, equally competent researchers give different answers to policy-makers. Why do the answers differ so? In many research areas--education, fertility, and history are examples--the existing body of theory does not carry the investigators very far. With a set of priors which are not firmly established the researcher goes to an examination of the data to further illuminate relationships. The data are mined, which means, of course, that standard errors and t-statistics do not really provide discriminatory guidelines for distinguishing the stable relationships. Such relationships can only be established by documenting their robustness--by replication. Despite references to replication in standard texts, however, it remains a greatly underutilized tool.[1] There is a need to encourage its use and explore the techniques and interpretation of multiple sets of results.

In this article, we take a coherent look at the equivalent of many studies addressing the problem--which school resources help learning? The objective is threefold: (1) to use very rich data to shed new light on what makes a difference to the junior high school age student, who has been less studied; (2) to apply a missing variable technique for addressing the sometimes encountered problem of having rich data for a small sample, and poorer data for a large sample from the same population; (3) to lay out the winnowing process of analyzing sets of results from several experiments--the number of reproduced results may be smaller than the number of significant results from one experiment, but the confidence in them can be greater.

Why has this sort of procedure been done so rarely? Sometimes there may have been an insufficient number of observations. (In the literature of educational production functions, however, there are generally observations to burn!) More fundamentally, and curiously enough, it hasn't been regarded as essential to common scientific practice in the social sciences. It means, of course, that the investigator must deal with the hazards of "explaining" different sets of results. If the results are different enough, the investigator may perish, not publish.[2] And the economist's model would suggest that the incentives don't exist--tenure points are given for the value added by conclusions from new populations or new questions from old populations, not for conclusions from repeated samples from the same population.

The consequence is that policy-makers receive empirical results that are less than firmly rooted. This, in part, explains why the world frequently doesn't behave as predicted when policies are implemented. And disaffection for empirical research, appropriately enough, sets in.[3]

In this article, therefore, the process of replication in the field of education production functions is explored. More specifically, data for a large number of junior high school students in the Philadelphia School District are explored with the development of a hypothesized relationship and three replications. Comparisons are made between the policy recommendations that flow from the prereplication and postreplication results. And the robust recommendations for increasing students' learning are identified.

Section I describes the samples and model estimation procedures for the initial experiment with a pupil-specific[4] data set, and for three forms of replication. A summary of the major findings from these four experiments is presented in Section II. The differences in policy recommendations--the winnowing effect of replication--are discussed in Section III. Some concluding remarks constitute Section IV.

1. DESCRIPTION OF SAMPLES AND MODEL ESTIMATION PROCEDURES

We had access to a very rich data base. The number of observations was large, and the data were available on a disaggregated and longitudinal basis. The quality of these data allowed for a better exploration of the question--what helps junior high school age students learn?--than has been done previously. The misleading answers coming from aggregated data (using school averages, rather than pupil-specific measures of teacher characteristics, for example) have been avoided: and it was possible to explore the interactions of pupil characteristics with school inputs [Summers and Wolfe, 1977].

All of the data used in this study are drawn from 1970/71 pupil files for eighth grade pupils in the Philadelphia School District. These pupils attended junior high, K-8, and middle schools. A two-year education history was compiled for each pupil, including achievement test scores in the sixth and eighth grades, and schools attended. Socioeconomic information was collected, including race and sex; and an estimated family income was matched to each pupil by using his or her address.[5] The dependent variable chosen is the change in a composite achievement test score over the two-year period, sixth to eighth grades; it is the same for all samples. This change formulation permits the prediction of the effect on pupil learning of changes in educational input.

Four separate data sets were used to explore the replication process: (A) a relatively small, pupil-specific data set; (B) a large, less disaggregated sample; (C) another sample drawn randomly from the less pupil-specific data set; and (D) a constructed data set based on (B), but "corrected" for the aggregate nature of the data--a variant of the missing data problem. The level of detail, particularly teacher quality and attendance information, varies significantly by sample. In each case, the same basic equation is estimated, though some of the variables differ in level of aggregation. The use of (A) to form the hypothesized relationships, which are then tested by (B), (C), and (D), is an attempt to accomod the standard scientific procedure to the realities of the discipline.

A. Small, Pupil-Specific Sample

The initial results are based on data for 553 eighth grade pupils who attended 42 schools. The schools were randomly selected and the pupils were randomly selected from these schools. Detailed data were obtained from files kept within the school; these included individual pupil attendance records, test scores, family backgrounds, and the specific teachers each pupil had. This information was then matched to a teacher file, so that each pupil's data file has, for his or her individual English, Math, and Social Studies teachers,[7] a fairly detailed set of measurable characteristics of quality: a rating of the undergraduate institutions attended,[8] National Teacher Examination scores, levels of education completed, and years of experience as a teacher. The pupil information was also matched to information on the school(s) attended:

this included information on the principal's characteristics, racial composition, percentage of high and low achievers, enrollment, age and condition of the facilities, and class size by grade.

This sample is the "ideal" one in terms of data--the one on which 453 regressions were run to arrive at the equation of "best fit," the one which, in the absence of an agreed-upon body of theory to test, provides the hypothesized relationship against which replications can be checked. The calculated results of the "best fit" equation, our choice of the appropriate specification, are listed in Table 1, Column (2). A number of functional forms for each of the variables have been examined. Dummy variables and other nonlinearities were explored. Interactions between pupil characteristics and school inputs (the relationship between class size and pupil's family income, for example) were extensively examined, consistent with the generally accepted theory of educators that students of different abilities react differently to the characteristics of their schools, their peers, their teachers, and their programs. All the findings were checked and rechecked against alternative specifications for robustness--only those variables which remained strong throughout the mining process were retained. Normally, this is the equation from which policy conclusions flow. However, we regard these results as Step One--the development of a hypothesized relationship between the change in achievement from sixth to eighth grades and many inputs.

B. Large Sample Involving Less Pupil-Specific Data

Another sample was randomly drawn from the Philadelphia School District's computerized record files; after a small number were discarded

TABLE 1 - REGRESSION RESULTS FOR FOUR EIGHTH GRADE SAMPLES OF PHILADELPHIA SCHOOL DISTRICT STUDENTS, 1969-71[a]

(t-values in parentheses)

Variables (1)	Pupil-Specific Small Sample (2)		Large Sample (3)		Stored Data from Large Sample (4)		Tests of Equality (3) and (4) (5)	Modified Small Sample (6)		Large Sample with Estimators (7)		Tests of Equality (6) and (7) (8)
1. Sex	.63	(.73)	.58	(1.22)	1.20	(1.24)	.57	.42	(.48)	.67	(1.36)	-.25
2. (1) x Low Achievement	-2.77	(-2.41)	-1.54	(-2.31)	-2.80	(-2.11)		-2.38	(-2.04)	-1.11	(-1.37)	
Low Ach. = 1	-2.14	(-2.58)	-.96	(-1.78)	-1.60	(-1.61)	-.56	-1.97	(-2.33)	-.44	(-.65)	-1.40
3. 2nd Generation Native Born	4.52	(2.13)	b		b			b		b		
4. Third-Grade Score	.52	(2.41)	b		b			b		b		
5. $(4)^2$	-.004	(-1.49)	b		b			b		b		
6. Sixth-Grade Score	-.38	(-2.38)	.04	(1.26)	.07	(1.15)	.43	-.32	(-1.93)	-.01	(-.03)	-1.21
7. Race (Black = 1)	-.36	(-.12)	-1.39	(-2.62)	-.66	(-.61)	.61	.27	(.09)	.40	(.12)	-.03
8. Income	-.21	(-1.10)	.28	(3.10)	.44	(2.37)	.77	-.15	(-.77)	.17	(1.49)	-1.40
9. Unexcused Absences[c,f]	.30	(1.54)	.10	(1.86)	.11	(1.06)	.13	.33	(1.71)	.18	(.53)	
10. (9) x (6)[f]	-.01	(-2.69)	e		e			-.01	(-2.84)	-.004	(-.63)	
(6) = 30	-.02	(-.20)						.0003	(.004)	.06	(.35)	-.30
(6) = 50	-.23	(-5.15)						-.22	(-4.96)	-.03	(-.25)	-1.76
(6) = 80	-.54	(-4.16)						-.56	(-4.23)	-.15	(-.68)	-1.58
11. Residential Moves	2.12	(.69)	e		e			3.55	(1.15)	-.35	(-.24)	
12. (11) x (6)	-.08	(-1.27)						-.11	(-1.72)	.002	(.08)	
(6) = 30	-.31	(-.24)						.24	(.19)	-.28)	(-.38)	-.35
(6) = 50	-1.92	(-3.04)						-1.97	(-3.05)	-.24	(-.55)	-2.22✓
(6) = 80	-4.35	(-2.01)						-5.28	(-2.42)	-.18	(-.22)	-2.19✓
13. Rating of Social Studies Teacher's College[f]	-3.06	(-1.50)	e		e			-3.49	(-1.68)	-.58	(-.22)	
14. (13) x (6)[f]	.06	(1.59)						.07	(1.82)	.02	(.46)	
(6) = 30	-1.28	(-1.27)						-1.42	(-1.39)	.04	(.03)	-.81
(6) = 50	-.09	(-.17)						-.05	(-.09)	.45	(.48)	-.46
(6) = 80	1.69	(1.50)						2.02	(1.77)	1.07	(.77)	.52
15. Rating of English Teacher's College[d,f]	-2.10	(-3.02)	-1.13	(-1.70)	1.78	(1.49)	2.13✓	-1.99	(-2.81)	-.24	(-.28)	
16. (15) x (6)[d,f]	.04	(3.53)	.03	(5.02)	.02	(1.88)		.04	(3.21)	.03	(4.03)	
(6) = 30	-.96	(-2.07)	-1.14	(-.25)	2.49	(2.44)	2.25✓	-.94	(-1.98)	.62	(.85)	-1.78
(6) = 50	-.21	(-.53)	.51	(.94)	2.96	(3.08)	2.21✓	-.24	(-.59)	1.20	(1.76)	-1.82
(6) = 80	.93	(1.93)	1.50	(2.67)	3.66	(3.69)	1.90	.82	(1.68)	2.06	(3.16)	-1.53
17. English Teacher's Experience[d,f]	-.54	(-.78)	-3.70	(-2.72)	-5.29	(-1.92)	-.52	-1.05	(-1.56)	-.31	(-.17)	-.38

TABLE 1 (CONTINUED)

18. English Teacher's Experience, >3[d,f]	.14	(.18)	4.15	(2.97)	6.19	(2.20)	.65	1.13	(1.55)	.82	(.44)	.15
19. English Teacher's Experience, >10	3.93	(2.50)	b		b			b		b		
20. Math Teacher's Experience[d,f]	.16	(1.34)	.60	(3.76)	.12	(.40)	1.40	.19	(1.60)	.16	(.57)	.10
21. Social Studies Teacher's National Exam Score[d,f]	.11	(2.01)	e		e			.11	(1.84)	.02	(.27)	.91
22. Math Teacher's Education Beyond B. A.[d]	e		-2.69	(-3.69)	-.79	(-.52)	1.13	e		e		
23. Math Teacher's Race = Pupil's Race[f]	1.85	(1.89)	e		e			1.17	(1.18)	3.85	(1.10)	-.74
24. Percent Black Teachers	-.17	(-3.72)	e		e			-.16	(-3.54)	-.01	-.34	-2.76✓
25. Remedial Education Expenditures	.03	(1.77)	.02	(4.26)	.03	(2.90)	.72	.03	(1.61)	.02	(3.06)	
26. (25) x Low Achievement	-.04	(-1.72)	e		e			-.04	(-1.87)	-.002	(-.15)	
Low Ach. = 1	-.01	(-.38)						-.02	(-.67)	.02	(1.71)	-1.43
27. Class Size ≥ 32	-8.29	(-3.12)	e		e			-8.37	(-3.11)	1.95	(1.11)	-3.21✓
28. (27) x (8)	.66	(2.10)						.60	(1.88)	-.23	(-1.32)	
(8) = 5	-4.99	(-4.05)						-5.37	(-4.30)	-.79	(.85)	(-3.95)✓
(8) = 10	-1.69	(-1.74)						-2.37	(-2.42)	-.36	(-.77)	(-1.86)
(8) = 15	1.62	(.70)						.63	(.27)	-1.51	(-1.45)	.84
29. Attending a K-8 School	4.52	(3.28)	3.31	(3.54)	6.83	(3.39)	1.58	4.80	(3.43)	4.36	(5.00)	-.26
30. Percent Black Students	.07	(1.66)	e		e			.07	(1.56)	.01	(.42)	
31. (7) x (30)	-.04	(-.51)						-.08	(-.95)	.03	(.59)	
(7) = Black	-.03	(.50)						-.01	(-.08)	.04	(.93)	-.58
32. Percent Black Students ≥ 50	-.21	(-1.99)	e		e			-.24	(-2.21)	-.02	(-.20)	
33. (7) x (32)	.32	(2.34)						.40	(2.89)	-.06	(-.66)	
(7) = Black	.11	(1.19)						.16	(1.75)	-.08	(-1.13)	2.08✓
34. Relative Income Change in School Feeder Area, 1960-70	e		-8.71	(-2.39)	-22.66	(-3.07)	-1.70	e		e		
35. School Enrollment	e		-.08	(-1.83)	-.03	(-.31)	.58	e		e		
Constant	13.59		9.00		-3.32			27.80		-2.54		
$\bar{R}^2$	.32		.32		.30			.29		.35		
$\hat{F}$	9.03		42.73		12.93			8.94		30.65		

<u>Note</u>: Dependent Variable: Eighth Grade Iowa Test of Basic Skills (Composite) Minus Sixth Grade Score.

[a] Sources of data are listed in the Appendix.

[b] No comparable data are available for the other samples.

[c] In the samples described by columns (3) and (4), the variable used is Average Daily Attendance in the pupil's school.

[d] In the samples described by columns (3) and (4), the variables used are the average quality characteristic of the relevant group of teachers in the pupil's school.

[e] Not significant at the t=.05 level. Variable not included in equation.

[f] Variables for which estimators were developed.

✓ Null hypothesis test reveals significantly different results.

because of incomplete data, these consisted of 1,541 eighth grade pupils in 52 schools. Step Two involved testing the hypothesized relationship established in Step One, the result of torturous experimentation with the small, pupil-specific sample, on this virgin data.

In many respects, the data were of the same quality: Information on individual pupil test scores was matched to data that were available on tapes of individual records. These included data on sex, race, address (to match family income estimates with block income estimates), and schools attended. As in the small sample, these data were matched with characteristics of the schools and their principals.

But, as is the reality in social science experimentation, some data differed: information about the teachers each pupil had was not available on the tapes. Instead, pupil data were matched with the averages of the teachers in the school in the relevant grade and subject. Thus, Mary Smith's scores could not be matched with the number of years of experience of Mr. Jones, her eighth grade Math teacher (the detail of data available in the small sample), but with the average years of experience of the eighth grade Math teachers in her school. Pupil-specific data on attendance were also unavailable: Mary Smith's scores could not be matched with her unexcused absences, but only with the average daily attendance of her school(s).

An equation, paralleling the "best-fit" equation of the small sample--using the best directly available substitutes for pupil-specific data--is our first replication or testing of the hypothesized relationship. The results of the calculations are in Table 1, Column (3). The use of aggregated data--means, rather than individual observations--does not, of course, result in biased estimates. The errors are uncorrelated, so it is entirely legitimate to compare the results of the sample using some averages as data with the original experimental sample.

C. Stored Sample From Large Sample

This sample is made up of 465 eighth grade pupils from 50 schools. These were drawn randomly from the sample taken from the Philadelphia School District pupil files and, figuratively, were stored away in the bank vault. The hypothesized relationship developed from the pupil-specific data of the small samples was tested on these data, as it was on the 1,541 observations of the large sample. Since the samples were from the identical population, the variables matched exactly. The results of this replication are shown in Table 1, Column (4).

D. Large Sample With "Corrected" Aggregate Data

The pupils included in this sample are the same as those in the large sample, but correction factors were developed for the non-pupil-specific data. To improve upon the aggregate nature of these data, estimators of pupil-specific data for teacher qualities and attendance were created, using the information available from the small sample.

This approach can be though of as a way to deal with missing data, a problem common to many multivariate analyses. One standard approach is to eliminate missing data using either list-wise deletion or pair-wise deletion (method of moments). Another approach, and the one used here, is to estimate (replace) missing data. Replacing missing values by averages (means) is a well-known way of estimating missing values, and is essentially the method implied in the large and stored samples. It can, however, result in inconsistent estimates.

Instead, the missing variables can be regressed on the available variables. This approach "minimizes the incompatibility of estimates"--the Mahalanobis distance from each case to the mean is near its minimum [Frane, 1976]. Problems are present with this method if the missing observations are nonrandom, in which case the use of conditional means may be required [Heckman, 1976]. These difficulties should not apply here, however, since the method is applied to a random sample.

Another approach, when the data are "unobservable," is to use maximum likelihood methods to estimate coefficients for unobservables, using unobserved components methods. These methods are particularly useful in analyzing sibling data. They are less useful here, however, since (1) the large number of inputs used in this study would create problems of identification, and (2) the methods constrain the parameters to be linear [Taubman, 1977; Chamberlain and Griliches, 1975], and nonlinear specificatio[ns] are extensively explored in this study.

More specifically, the problem in this study was to estimate an equation of the form:

$$A_i = Y_i\phi_i + X_i\beta_i + \varepsilon_i$$

where

i = ith pupil

A = Achieving (eighth grade score minus sixth grade score)

Y = Data in the small sample for which estimators will be created

X = Vector of variables available in both the large and small sample in the same form

ϕ = Coefficients of the Y vector

β = Coefficients of the X vector

ε = Error term

The problem was that there were missing data on Y_i. However, there was a data set (the small sample) which had information on Y_i for pupils in that sample. So, for each of the Y_i variables, we estimated an equation using the small-sample data, where each Y_i was a dependent variable and the independent variables included the following: (1) the X vector--a set of variables describing socioeconomic characteristics (sex of student, for example), school characteristics (enrollment of school, for example), and student body characteristics (proportion of black students, for example); (2) match-up variables to the Y_i variables (the average experience of eighth grade English teachers in a school matches up with the experience of each pupil's English teacher, for example); (3) additional variables available in both the large and small samples. Single-equation ordinary least squares was used to estimate the equations, which all had the same right-hand-side variables. The selection criteria for the set used were the adjusted R^2s and the t-statistics. Estimators and residuals were calculated from these regressions for the small sample.

A test of the correlation between the residuals and the independent variables of the equation $A_i = Y_i\phi_i + X_i\beta_i + \varepsilon_i$, was made to avoid the multicollinearity and associated problems of bias of estimators in the equation of interest. One wanted to find, of course, no correlation.

The $V_{i,j}$s (the error term of each estimator equation for each observation in the small sample; j = estimator, i = observation for each pupil) were multiplied by the corresponding coefficients, ϕ_i, from the small sample equation. This vector multiplication, $V_i\phi_i$, resulted in a scalar (weighted error term) for each observation (pupil) in the small sample. These were regressed with the residual for each pupil from the small sample equation ($\bar{\varepsilon}_i$). The adjusted R^2 for this equation between the error terms was -.0017. It was assumed, therefore, the $E(v\varepsilon) = 0$; that is, there was no covariance between the error terms. Next, the coefficients from the estimator-generating equations in the small sample were used to create estimators for each pupil in the large sample.

In order to permit better evaluation of the estimation procedure, an adjustment was made for the additional variance caused by the estimation procedures for the estimator. The ratio of the variance of the original small-sample equation, S^2_ε, to the variance of the small-sample equation using the estimators, $S^2_{\varepsilon\phi v}$, was computed. This scalar[9] was then used to adjust the variance-covariance matrix and generate new coefficients.

This procedure, then, was the one used to generate this set of estimators (proxy variables) for the large sample: unexcused absences of pupils, rating of the undergraduate institutions attended by the Social Studies and English teachers, experience of the pupil's English and Math teachers, scores of each pupil's Social Studies teacher on the National Teacher Examination, and a match-up of the race of pupil and Math teacher.

A regression was run using these estimators as substitutes for the aggregate data in the large sample, testing the hypothesized relationships developed for the small sample. The results are in Table 1, Column (7). To compare the two sets of results, some modifications in the original small sample had to be made, because there were some data which were not available in the large sample and for which no reasonable match-up existed (third grade score and country of birth). The modified small-sample results are in Table 1, Column (6).

2. SUMMARY OF FINDINGS

The fundamental priors that some school and teacher inputs help learning growth, and do so differentially, were tested in one hypothesis--developing experiment and three replications involving Philadelphia junior high school students. Did the "answers" differ? Interestingly enough, most "answers" were consistent throughout the replications. Some, however, were only similar, and some differed substantially. No matter how one looks at the data, the same conclusion emerges[10]: school resources help learning; organizational and teacher characteristics affect learning growth, and the negative effects of Race and Income can be traced through their impact on the student's ability to absorb particular school inputs.

The "answers" emerging from each experiment are laid out in Table 1 in Columns (2), (3), (4), (6), and (7) in the form of coefficients and t-statistics.[11] Equality of coefficient tests, Columns (5) and (8), were performed to test whether or not answers from the different experiments differed.[12] Comparisons that failed the test are checked. But these significance tests do not, of course, provide anywhere near perfect guidance for interpreting the results. The pupil-specific sample that we used to develop our hypothesis is a "worked-over" sample, so the t-statistics in Column 2 should be read with the recognition that they are overstated, reflecting the understatement of the standard errors. The t-statistics for the remaining samples are more consistent with the underlying assumptions of significance tests, however. The tests of equality (Columns [5] and [8]) need to be interpreted with the recognition that the standard errors of the samples will differ because of differences in sample sizes.

A pragmatic regression strategy calls for using all the results, with the appropriate caveats, in the final interpretations. These are laid out in Table 2, where the findings from the first experiment, with pupil-specific data, are compared with the findings from the three replications.

Most of the estimates of the effects of genetic and socioeconomic inputs appear to be robust. At the junior high ages, only the low achieving males have slower achievement growth than their female counterparts. When an early achievement measure is included (third grade score), the impact of the end-of-elementary-grades score (sixth grade score) shows regression toward the mean; where all earlier achievement is reflected in one score only (sixth grade score), the effects cancel out. The prior that pupils from different economic and racial backgrounds respond differentially to school and teacher inputs is confirmed: sixth grade score (achievement) shows no separate effect on achievement growth, apart from the effects traced through other inputs; being black is associated with less achievement growth, but where the effects of being black are traced through interactions

TABLE 2 - COMPARISON OF FINDINGS OF PUPIL-SPECIFIC SAMPLE WITH THREE OTHER SAMPLES OF PHILADELPHIA EIGHTH GRADE STUDENTS, 1969-71

	Pupil-Specific Sample[a]	Replication[b]
Genetic and Socioeconomic Inputs		
Sex	Low achieving males have lower achievement growth than low achieving females; higher achieving males and females show no difference.	Same findings.
6th Grade Score	Starting score shows regression toward the mean.	Same findings from equations with pupil-specificity. (No evidence of regression toward the mean in equations without pupil-specificity.)
Race	Race of the pupil is not directly associated with achievement growth, though it does interact with school inputs.	Same findings from equations with pupil-specificity. (Being non-white is negatively associated with achievement growth in equation without pupil-specificity.)
Income	Income of the pupil is not directly associated with achievement growth, though it does interact with school inputs.	Same findings from equations with pupil-specificity. (Income is positively associated with achievement growth in equations without pupil-specificity.)
Unexcused Absences	Unexcused absences are negatively related to achievement growth--the higher the achiever, the greater the loss.	Same findings from equations with pupil-specificity. (Different variable-average daily attendance-supports results in equations without pupil-specificity.)
Residential Moves	More residential moves are associated with less achievement growth for middle and high achievers; low achievers were unaffected.	Dissimilar findings: More residential moves are not associated with achievement growth of low, middle or high achievers.
School Inputs: Teacher Quality		
Rating of Teacher's College		
Social Studies Teacher	Being taught by a Social Studies teacher who attended a higher-rated college is associated with greater achievement growth for higher achievers; middle and low achievers were unaffected.	Same findings from equations with pupil-specificity. (No association in equation without pupil-specificity.)

TABLE 2 (CONTINUED)

School Inputs: Teacher Quality (Continued)

English Teacher	Being taught by an English teacher who attended a higher-rated college is associated with greater achievement growth for high achievers, less achievement growth for low achievers. Middle achievers were unaffected.	Same findings for high achievers: Dissimilar findings for middle and low achievers.
Math Teacher	Being taught by a Math teacher who attended a higher-rated college is not associated with achievement growth.	Same findings.
Years of Experience		
English Teacher	More experience is associated with higher achievement growth (the significant association is with teachers with 10 or more years of experience).	Similar findings: Same general finding, but data did not permit exploring impact of 10 or more years of experience.
Math Teacher	Being taught by a Math teacher with more experience is associated (though not strongly) with higher achievement growth.	Similar findings: Same general finding, but stronger findings from equation without pupil-specificity.
Social Studies Teacher	Being taught by a Social Studies teacher with more experience is not associated with achievement growth.	Same findings.
Score on National Teachers Exam		
Social Studies	Being taught by a teacher who scored higher on a NTSS Exam is associated with higher achievement growth.	Same findings from equations with pupil-specificity. (No association in equation without pupil-specificity.)
English	Being taught by a teacher who scored higher on a NTEE is not associated with achievement growth.	Same findings.
Math	Being taught by a Math teacher who scored higher on a NTME is not associated with achievement growth.	Same findings.
Common	Being taught by a teacher who scored higher on a NTE is not associated with achievement growth.	Same findings.
Graduate Education		
Math Teacher	Being taught by a Math teacher with more graduate education is not associated with achievement growth.	Same findings from equations with pupil-specificity. (Negative association in equations without pupil-specificity.)

TABLE 2 (CONTINUED)

School Inputs: Teacher Quality (Continued)		
Social Studies Teacher	Being taught by a Social Studies teacher with more graduate education is not associated with achievement growth.	Same finding.
English Teacher	Being taught by an English teacher with more graduate education is not associated with achievement growth.	Same finding.
Race of Pupil = Race of Math Teacher	Being taught by a Math teacher of the same race is positively associated with a pupil's achievement growth.	Dissimilar findings. No evidence of association in other equations.
Race of Pupil = Race of English Teacher	Being taught by an English teacher of the same race is not associated with achievement growth.	Same finding.
Race of Pupil = Race of Social Studies Teacher	Being taught by a Social Studies of the same race is not associated with achievement growth.	Same finding.
Percent of Black Teachers	A higher % of Black teachers is negatively associated with achievement growth.	Dissimilar findings. No evidence of association in other equations.
School Inputs: Non-Teacher School Quality		
Remedial Expenditures Per Low Achiever	More remedial expenditures were associated with higher growth for middle and high achievers, but were not related to achievement growth for low achievers.	Similar findings. Positive association with learning for middle and high achievers confirmed, but positive association for low achievers found in all other samples.
Class Size ≧ 32	Being in a class with 32 or more pupils is most negatively associated with achievement growth for lower income pupils; higher income pupils were unaffected.	Dissimilar findings. Lower income pupils showed no association between learning and class size, as did higher income pupils.
K-8 Elementary School	Attending 8th grade in a K-8 school is associated with much greater achievement growth.	Same findings.
School Size	Size of school has no association with achievement growth.	Same findings from pupil-specific samples. (Negative association in samples without pupil-specificity.)

TABLE 2 (CONTINUED)

<u>School Inputs: Non-Teacher School Quality (Continued)</u>		
Experience and Education of Principals	Additional experience, degrees, and credits of education of principals are not associated with achievement growth.	<u>Same findings</u>.
Physical Condition of School	Attending a school with fewer pupils per lab, in better physical condition, or newer is not associated with achievement growth.	<u>Same findings</u>.
<u>Peer Group Effects</u>		
% of Black Pupils	An increasing % of Black pupils in a school up to 50% has no association with achievement growth for Blacks or non-Blacks.	<u>Same findings</u>.
	An increasing % of Black pupils in a school from 50 to 100% has no clear association with achievement growth for Blacks, but a negative association for non-Blacks.	<u>Dissimilar findings</u>. No association found for Blacks or non-Blacks.
Change from 1960 to 1970 in Relative Income of Feeder area	Attending a school with a higher relative income change is not associated with achievement growth.	<u>Same findings from pupil-specific samples</u>. (Negative association in samples without pupil-specificity.)

[a]Summary of results in Table 1, Column (2).

[b]Summary of results in Table 1, Columns (3) - (8).

with other inputs, no residual effect remains; similarly, students from higher-income families have more achievement growth, but where the effects of income are traced through interactions with other inputs, no residual effect remains. Motivation (the number of unexcused absences is the surrogate measure) is clearly associated with achievement growth.

Certain qualities of teachers--using a rating of undergraduate institutions, years of experience, score on National Teacher Examinations, and amount of graduate education, as quality measures--have clear bearing on the achievement growth of students. The evidence suggests (1) that high-achieving students benefit most from being taught by English and Social Studies teachers who received their undergraduate degrees from higher-rated institutions, (2) that students benefit from being taught by more experienced English and Math teachers, (3) that the National Teacher Examination scores are not associated with student learning, except for Social Studies teachers, (4) that more graduate education for teachers is not associated with achievement growth in any of the subjects, and (5) that a match of race between pupil and teacher is not reflected in more or less learning.

There are a number of clear associations and nonassociations between aspects of the school other than its teachers and junior high school student achievement growth: a positive one between remedial education expenditures and learning for students with high ability, a positive one between attending a K-8 school and achievement growth, none between additional experience, degrees, and credits of education of principals and student learning, and none between the measurable physical conditions of the school and learning. The class size debate must continue, however:[13] while the sample with the most ideal data shows that classes larger than 32 are negatively associated with the achievement growth of low-income pupils, the other replications do not support this finding. The only consistent finding in these samples is that higher-income students are unaffected by the class size.

Finally, we find, over the four samples, the consistent result that variations between 0 and 50 percent black in the racial mix of the schools were not associated with learning changes for blacks or nonblacks; neither was there a discernible relationship for blacks in the range of 50 to 100 percent black; for nonblacks in the more than 50 percent black schools, no robust finding emerged.

It is clear, in this sorting out of robust findings, that whether or not the sample has pupil-specific data is a major determinant of whether or not a result is replicated. When the school inputs are pupil-specific, either obtained originally or estimated (Columns [2], [6], and [7]) most of the findings are repeated. The exceptions are the ones relating to the impact of the number of residential moves, the percentage of black teachers, class size for low-income pupils, and more than 50 percent black pupils in a school. When pupil-specific data are not used, more findings fall away. Most strikingly, if the school inputs are not specified in terms of the pupil, the SES results are significantly altered--being nonwhite or poor assumes excessive importance.

It is equally clear, as one might expect, that while we gain in certainty as the replication procedure develops, we lose in the number of conclusions we can be confident about.

3. POLICY IMPLICATIONS

What policies are suggested by the priors in relation to the results of the hypothesis-developing pupil-specific data set? What policies remain after the replications? The empirical results of Table 1 and the interpretation of the results of Table 2 are translated into suggested policies in Table 3. Column (1) lists the prereplication advice and Column (2) the postreplication advice for helping to maximize achievement growth between the sixth and eighth grades. An additional type of replication is grafted on to this table; in it we have noted the results from a previous study by the authors (1977) of third to sixth grade student achievement growth in the Philadelphia School District. An asterisk in Table 3 indicates that this elementary school study had results leading to the same policy suggestions; a check indicates that the elementary school study had results leading to similar policy suggestions.

So--it seems appropriate to generate some enthusiasm for action in a policy suggestion which has a Same in Column (2), and asterisks or checks in both columns. Such a finding has weathered a fair number of storms. There are a good number of such findings on the list, relating to pupil motivation, teacher qualities, class size, qualifications of the principal, and physical characteristics of the school.

Any enthusiasm for action, of course, must be confined to the Philadelphia School District. Whether or not such policies would emerge from studies of other school districts is an item in a research agenda, not a known piece of information.

TABLE 3. PRE- AND POSTREPLICATION POLICY SUGGESTIONS FOR IMPROVING SIXTH TO EIGHTH GRADE ACHIEVEMENT GROWTH IN THE PHILADELPHIA SCHOOL DISTRICT

Prereplication	Postreplication
* Use the number of unexcused absences of a pupil as a trouble signal.	* Same.
√ Rearrange assignments or hire teachers of the "softer" subjects so that those from higher-rated colleges are placed with the high-achieving pupils.	√ Same.
√ Higher pay for more experienced English and Math teachers is warranted in terms of productivity.	√ Same.
Use the Social Studies National Teacher Exam score as a predictor of teacher ability to improve overall achievement growth. The common and other subject exam scores are not useful indicators.	Same.
* Differential pay for teachers who have education beyond the B.A. is not warranted.	* Same.
* Policies in teacher placement should not include matching teacher's race to pupil's race.	* Same.
Allocate the Black teachers in a school system evenly among schools, if the objective is to give each student body an equal impact of racial balance of staff.	* Use criteria, other than achievement, as the basis for allocating Black teachers.
* The evidence that the experience and amount of education beyond the B.A. of principals has no payoff for student achievement growth should be borne in mind in determining hiring and salary criteria.	* Same.
Efforts to reach the low-achieving target group should be increased, because remedial expenditures are not helping those for whom they are intended.	Same, but remedial expenditure help low achievers, too.
√ As school systems try to cope with declining enrollments, the finding that having an 8th grade part of the elementary school is the most productive organization, is an important input.	√ Same.
√ Have class sizes less than 32 in schools with many low income students, and large classes in schools with many high income students.	√ Larger classes do not have a negative effect.
Impact on achievement growth is not a useful criteria for determining school size.	√ Smaller schools might have a positive effect on pupil achievement growth.
√ Integration levels of schools do not seem to affect achievement.	Assign students to schools without regard to racial mix.
* Additional expenditures on plant and equipment are not warranted in terms of pupil learning.	* Same.

* Duplicated in a study of 6th grade students by the authors (1977). √ Partially duplicated in a study of 6th grade students by the authors (1977).

4. CONCLUDING REMARKS

Our examination of several types of samples of Philadelphia eighth graders yields some conclusions about what helps them to learn and some conclusions about how to learn what helps them to learn.

In broad terms, the results reaffirm the conclusions of prior econometric work and partially developed educational theory that some school and teacher inputs affect student learning, and that they do so differentially. The strongest results of the replications described suggest that, in Philadelphia at least, sixth to eighth grade learning is significantly affected by the experience of teachers, by the quality of the undergraduate institutions teachers attended, by the grade organization of the school, and by remedial expenditures. The results suggest that the number of unexcused absences is a useful signal of trouble. And the results suggest that there are a number of costly items that do not have a payoff in learning growth: education of teachers beyond the B.A., small classes for most students, experience and education beyond the B.A. for principals, and physical characteristics of the school.

Methodologically, we conclude several things. First, it seems clear that pupil-specific data are essential to illuminating the relationships between inputs and learning---aggregate measures of school inputs exaggerate the role of socioeconomic factors, render most school inputs impotent, and do not allow investigation of differential response to school inputs. Second, using a small sample of pupil-specific data to develop pupil-specific proxies for a large sample of aggregated data provides a cheap and productive way of moving towards more disaggregation, where the pupil-specific data are not easily available. Third, encouragement needs to be given--in education, as in other fields where large quantities of public resources are spent--to developing large, easily available data bases which are disaggregated, longitudinal, and have wide geographic coverage.

APPENDIX

The definitions, $\bar{X}$'s and σ's of each of the variables in Table 1 are listed here with the corresponding row numbers. The $\bar{X}$'s and σ's are presented for the independent variables of the small (SS), large (LS), stored (STS) samples, and for the estimators of the large sample (EST). The bracketed letters following SS and LS are key letters to the sources listed at the end.

The $\bar{X}$'s and σ's for the dependent variable, the difference between the eighth grade and sixth grade scores, are: SS[A]: $\bar{X}$ = 11.72, σ = 8.53; LS[A] $\bar{X}$ = 13.04, σ 8.91; STS: $\bar{X}$ = 12.88, σ = 9.16.

1. Sex: Dummy variable, 0 = female, 1 = male. SS[B]: $\bar{X}$ = .46, σ = .50; LS[A]: $\bar{X}$ = .49, σ = .50; STS: $\bar{X}$ = .47, σ = .50.
2. (1) x Low achievement: Interaction of Sex with dummy variables. 0 = non-low achiever; 1 = low achiever (sixth grade score $\leq$ 5.1).
3. Second Generation Native Born: Dummy variable, 0 = no, 1 = yes. SS[C]: $\bar{X}$ = .98, σ = .15.
4. Third Grade Score: Score on Iowa Test of Basic Skills given at the end of third grade. SS[B]: $\bar{X}$ = 32.26, σ = 8.35.
5. Quadratic formulation of Third Grade Score
6. Sixth Grade Score: Score on Iowa Test of Basic Skills given at the end of the sixth grade. SS[B]: $\bar{X}$ = 53.21; σ = 13.26; LS[A]: $\bar{X}$ = 57.08, σ = 14.72; STS: $\bar{X}$ = 55.95, σ = 14.44.
7. Race: Dummy variable, 0 = nonblack, 1 = black. SS[B]: $\bar{X}$ = .69, σ = .46; LS[A]: $\bar{X}$ = .63, σ = .48; STS: $\bar{X}$ = .66, σ = .47.

8. Income: Estimated family income (in thousands). SS[D,E]: $\bar{X}$ = 8.07, σ = 2.35; LS[D,E]: $\bar{X}$ = 9.22, σ= 3.00; STS: $\bar{X}$ = 9.03, σ = 2.89.

9. Unexcused Absences: Average of annual number of unexcused absences over two years, 1969/70 - 1970/71. SS[F]: X = 6.25, σ = 7.86; EST: $\bar{X}$ = 4.37, σ = 5.62.

   Or Average Daily Attendance: Average for two years, 1969/70 - 1970/71, of the percentage of average daily attendance in pupil's school. LS[G]: $\bar{X}$ = 83.21, σ = 4.58; STS: $\bar{X}$ = 83.11, σ = 4.23.

10. (9) x Sixth Grade Score: Interaction of sixth grade score with unexcused absences.

11. Residential Moves: Total number of residential moves of pupil between 1969 and 1971. SS[B]: $\bar{X}$ = .20, σ = .52; LS[A]: $\bar{X}$ = .36, σ = .70; STS: $\bar{X}$ = .34, σ = .67.

12. (11) x Sixth Grade Score: Interaction of sixth grade score with number of residential moves.

13. Rating of Social Studies Teacher's College: Gourman rating of eighth grade Social Studies teacher's undergraduate college. SS[H,I]: $\bar{X}$ = 415.8, σ = 61.3; EST: $\bar{X}$ = 414.0, σ = 34.7. Or school $\bar{X}$ of ratings of Social Studies teachers' colleges: School average of Gourman ratings of all Social Studies teachers' undergraduate colleges. LS[H,I]: $\bar{X}$ = 433.9, σ =35.0, STS: $\bar{X}$ = 435.2, σ = 36.4.

14. (13) x Sixth Grade Score: Interaction of sixth grade score with rating of eighth grade Social Studies Teacher's college.

    Or interaction of sixth grade score with school $\bar{X}$ of ratings of Social Studies teachers' colleges.

15. Rating of English Teacher's College: Gourman rating of eighth grade English teacher's undergraduate college. SS[H,I]: $\bar{X}$ = 421.7, σ = 82.9; EST: $\bar{X}$ = 406.5, σ = 47.7.

Or school $\bar{X}$ of Ratings of English teachers' colleges: School average of Gourman ratings of all English teachers' undergraduate colleges. LS [H,I]: $\bar{X}$ = 430.2, σ = 33.9; STS: $\bar{X}$ = 432.5, σ = 42.3.

16. (15) x Sixth Grade Score: Interaction of sixth grade reading score with rating of eighth grade English teacher's college.

Or interaction of sixth grade reading score with school $\bar{X}$ of ratings of English teachers' colleges.

17. English Teacher's Experience: eighth grade English teacher's experience, in years up to 11. SS[I]: $\bar{X}$ = 5.51, σ = 3.13; EST: $\bar{X}$ = 6.08, σ = 2.13.

Or school $\bar{X}$ of English teachers' experience: School average of all English teachers' experience (in years up to 11). LS[I]: $\bar{X}$ = 5.78, σ = 1.95; STS: $\bar{X}$ = 5.56, σ =1.84.

18-19. English Teacher's Experience >3>10: Two additional pieces of a three-piece linear function (spline) of eighth grade English teachers' experience with corner points at 3 and 10 years. (18) = maximum (0, years of experience -3). (19) = maximum (0, years of experience -10).

Or school $\bar{X}$ of English teachers' experience -3: Additional piece of a two-piece linear function (spline) of the school average of all English teachers' experience, set to equal maximum (0, average years of experience -3).

20. Math Teacher's Experience: eighth grade Math teachers' experience, in years up to 11. SS[I]: $\bar{X}$ = 5.26, σ = 3.10; EST: $\bar{X}$ = 5.58, σ = 1.98.

    Or school $\bar{X}$ of Math Teachers experience: School average of all Math teachers' experience (in years up to 11). LS[I]: $\bar{X}$ = 5.36, σ = 1.97; STS: $\bar{X}$ = 5.56, σ = 1.84.

21. Social Studies Teacher's National Teacher Exam Score: eighth grade Social Studies teacher's score on the National Teacher Exam in Social Studies. SS[I]: $\bar{X}$ = 64.58, σ = 5.96; EST: $\bar{X}$ = 64.15, σ = 4.18.

    Or school $\bar{X}$ of Social Studies Teachers' Exam score: School average of Social Studies teachers' scores on the National Teacher Exam in Social Studies. LS[I]: $\bar{X}$ = 63.90, σ = 3.17; STS: $\bar{X}$ = 63.98, σ = 3.14.

22. Math Teacher's education beyond B.A.: Number of additional credits beyond the B.A. of the eighth grade Math teacher. SS[I]: $\bar{X}$ = 1.41, σ = .53.

    Or school $\bar{X}$ of eighth grade Math teachers' education beyond B.A.: School average of eighth grade Math teachers extra credits beyond the B.A. LS[I]: $\bar{X}$ = 1.44, σ = .36; STS: $\bar{X}$ = 1.40, σ = .33.

23. Math Teacher's race = Pupil's race: eighth grade Math Teacher's race is same as pupil's race, 0 = no, 1 = yes. SS[I]: $\bar{X}$ = .38, σ = .49; EST: $\bar{X}$ = .45, σ = .40.

    Or difference between Percent Black Students and Percent Black Teachers: Absolute value of the percent black pupils minus percent black teachers in pupil's school. LS[J,K]: $\bar{X}$ = 33.85, σ = 17.54; STS: $\bar{X}$ = 34.45, σ = 18.74.

24. Percent Black Teachers: Average percent black teachers in schools pupils attended, 1969/70 - 1970/71. SS[K]: $\bar{X} = 35.7$, $\sigma = 16.7$; LS[K]: $\bar{X} = 33.33$, $\sigma = 18.10$; STS: $\bar{X} = 34.38$, $\sigma = 18.42$.

25. Remedial Education Expenditures: Expenditure in pupil's school on remedial education per low-achieving pupil. SS[L]: $\bar{X} = 38.58$, $\sigma = 24.06$; LS[L]: $\bar{X} = 47.05$, $\sigma = 38.02$; STS: $\bar{X} = 46.45$, $\sigma = 36.91$.

26. (2) x if Low Achiever: Interaction of remedial expenditure per low achiever with dummy variable; 0 = non-low achiever, 1 = low achiever (6th grade score ≤ 5.1).

27. Class size ≥ 32: Average number of pupils per classroom unit reporting attendance in pupil's seventh and eighth grade. Dummy variable: 0 = class size < 32, 1 = class size ≥ 32. SS[G]: $\bar{X} = .29$, $\sigma = .45$; LS[G]: $\bar{X} = .33$, $\sigma = .47$; STS: $\bar{X} = .35$, $\sigma = .48$.

28. (27) x Income: Interaction of income with class size ≥ 32.

29. Attending a K-8 school: Dummy variable, 0 = attending eighth grade not in an elementary school, 1 = attending an eighth grade in an elementary school. SS[M]: $\bar{X} = .09$, $\sigma = .29$; LS[A]: $\bar{X} = .06$, $\sigma = .24$; STS: $\bar{X} = .04$, $\sigma = .20$.

30. Percent Black Students: Average percent black students in school pupil attended, 1969/70 - 1970/71. SS[J]: $\bar{X} = 65.64$, $\sigma = 36.04$; LS[J]: $\bar{X} = 60.20$, $\sigma = 39.31$; STS: $\bar{X} = 64.37$, $\sigma = 38.03$.

31. (30) x Race: Interaction of pupil's race with percent black students in school.

32. Percent Blacks in school $\geq$ 50: Second piece of a two-piece linear function (spline) of percent black students in school with corner at 50%; set to equal maximum (0, % Black - 50).

33. (32) x Race: Interaction of pupil's race with percent black students in school (-50).

34. Relative Income Change in School Feeder Area, 1960-70: Income of student's school feeder area in 1960 relative to $\bar{X}$ income of Philadelphia ÷ same ratio for 1970. SS[E]: $\bar{X}$ = .03, σ = .06; LS[E]: $\bar{X}$ = .06, σ =.08; STS: $\bar{X}$ = .05, σ = .08.

35. School Enrollment: Number of pupils enrolled in school. SS[G]: $\bar{X}$ = 1632.54, σ = 585.07; LS[G]: $\bar{X}$ = 1676.44, σ = 532.96; STS: $\bar{X}$ = 1704.20; σ = 514.29.

Sources: [A] School District of Philadelphia (SDP) Pupil History Files; [B] Individual Pupil Records, Form EH-7; [C] Individual Admission Application Form, EH-40; [D] SDP 1960-70 Pupil Address File; [E] Authors, 1978; [F] SDP Roll Sheets; [G] SDP October Monthly Reports; [H] Gourman Report; [I] SDP Permis File; [J] Enrollment, Negro and Spanish-Speaking in the Philadelphia Public Schools; [K] Summary of Personnel in the Philadelphia Public Schools; [L] Detail of Proposed General Fund School Operating Budget; [M] Forms E-83 and E-84.

NOTES

[1]Henri Theil [1971] suggested that "a plea can be made to divide the available observations into three parts, the first of which is used for the choice of the specification, the second for the estimation, and the third for conditional prediction based on the estimated equation in order to verify whether the method actually works" [p. 603]. And Carl Christ [1966] observed that "in time-series studies the two main virtues of predictive tests--extending the number of observations with which a model is confronted, and doing so in a way that prevents us from choosing the model in the light of knowledge of the data with which it is to be confronted-can be attained by fitting the model to all the data in question just as well as they can be attained by fitting to some of the data and predicting the rest." But, he continued, "In cross-section studies the situation may be different. Here the sample is typically very much larger than in time-series studies. We can therefore divide an available sample into two parts, each containing hundreds or thousands of observations, one part to be used initially to help suggest the form of the model and the other part to be used later as a test of the predictive ability of the model chosen. It is not difficult to make that division so as to prevent our knowledge of the entire sample from influencing our choice of the model" [p. 548].

[2]Edgar L. Feige [1975], in an article on the incentives created by professional journal editorial policies, commented: "Unfortunately, we have all too often come to associate 'poor' results with the lack of achievement of statistical significance and 'good' results with the achievement of statistical significance. . . .it is undoubtedly encouraged by implicit journal editorial policies which assign a considerably lower probability of acceptance for publication to empiri-

cal studies which report 'negative' or, more correctly 'nonsignificant' results" [p. 1291].

[3]In the landmark Hobson v. Hansen case, Circuit Judge J. Skelly Wright of the Distric Court "held that Superintendent and Board, in operation of public school system in District of Columbia, unconstitutionally deprived Negro and poor public school children of their right to equal educational opportunity with white and more affluent public school children" [269 F. Supp. 401 (1967), p.401, D.D.C. 1967]. In the follow-up suit Judge Wright conveyed his disaffection with the opposing testimonies of expert witnesses: "the unfortunate if inevitable tendency has been to lose sight of the disadvantaged young students on whose behalf this suit was first brought in an over-grown garden of numbers and charts and jargon. . . . The reports by the experts-one noted economist plus assistants for each side-are less helpful than they might have been for the simple reason that they do not begin from a common data base, disagree over crucial statistical assumptions, and reach different conclusions" [327 F. Supp. 844 (1971), p.859, D.D.C. 1971]. Also, see Summers and Wolfe [1976], Berk [1977], Summers and Wolfe [1977] for an exchange on the use of statistics in the courtroom.

[4]Some of the data, such as test scores and sex, are unambiguously specific to the pupil. Other data, such as the class size and teacher background, are treated as uniform for all students in the classroom. If students in the same class are handled differently--receiving more or less of the teacher's time, for example--the data are not then literally specific to the pupil.

[5]We have developed a procedure described in an article by the authors [1978] using 1970 Philadelphia Census data for estimating block income. The estimates were generated from data on block mean housing values, block mean contract rental values, tract distribution of block contract rental values, and

tract distribution of income values. The block income appropriate to each pupil was taken to be his or her family income. This procedure involves: (1) forming the cumulative distributions of data for each tract of owner-occupied housing values, contract rental values, and family income; (2) converting these cumulative distributions into relative distributions (percentiles); (3) determing for each block the percentile in the tract distribution of mean owner-occupied housing value and the percentile for mean contract rental value; (4) determining the corresponding normal deviate arguments; (5) adjusting these by the regression coefficient for the tract between housing and income data for a cross-classified 20 percent sample; (6) assigning percentiles to the adjusted arguments; (7) finding the income values for thes percentiles; (8) adjusting for differences in the income distribution of renters and owners; (9) averaging the two income values for each block. The procedure was carried out for black and nonblack income and housing distributions, and each pupil was assigned an income estimate on the basis of his or her race.

[6]Definitions, $\bar{X}$'s, σ 's, and sources of the data are in the Appendix.

[7]These were the subjects all students were required to take.

[8]The Gourman rating, published in The Gourman Report (Phoenix, Arizona: The Continuing Education Institute, 1967) was used. It is a rating based on the undergraduate programs of nearly all colleges and universities in the United States, with information drawn from professional societies, commercial publications, foundations, etc., as well as the institutions themselves. The areas rated include: (1) individual departments, (2) administration, (3) faculty (including student/staff ratio and research), (4) student services including financial and honor programs, and (5) general areas such as facilities and alumni support. The Gourman rating is a simple average of all of these.

[9]We could proceed this way because we had tested to be sure that there w no covariance between the variance of the estimators and the variance of the small sample equation.

[10]See the authors [1977]. The same general conclusions were reached for elementary school students.

[11]For interaction variables, the t-statistics in the table indicate only whether or not there is a significant difference in the impact of each input among different types of students. Tests of the form:

$$\theta = \frac{\hat{\alpha} + \hat{\beta}y*}{\sqrt{\hat{\sigma}^2_{\hat{\alpha}} + \hat{\sigma}^2_{\hat{\beta}}y*^2 + 2\hat{\sigma}_{\hat{\alpha}\hat{\beta}}y*}}$$

were used to determine the specific values of the interaction terms at which the results were significant. The θ values for a full range of these interaction term values are shown in Table 1.

[12]For the variables without interaction the test took the form:

$$\tilde{\theta} = \frac{\hat{\beta}_1 + \hat{\beta}_2}{\sqrt{\hat{\sigma}^2_1 + \hat{\sigma}^2_1}} ,$$

where $\hat{\beta}_1$ is the coefficient of the experimental sample, $\hat{\beta}_2$ is the matching coefficient of the replication sample, $\hat{\sigma}_1$ is the standard error of the coeffici of the experimental sample, and $\hat{\sigma}_2$ the standard error of the coefficient of the replication sample. For variables with interactions, the test was of the form:

$$\tilde{\theta} = \frac{(\hat{\alpha}_1 + \hat{\beta}_1 y*) - (\hat{\alpha}_2 + \hat{\beta}_2 y*)}{\sqrt{[\hat{\sigma}^2_{\alpha_1} + \hat{\sigma}^2_{\beta_1} y*^2 + 2cov(\hat{\alpha}_1 \beta_1)] + [\hat{\sigma}^2_{\hat{\alpha}_2} + \hat{\sigma}^2_{\hat{\beta}_2} y* + 2cov(\hat{\alpha}_2 \hat{\beta}_2)]}} ,$$

where the terms with 1 as a subscript refer to the experimental sample, and the terms with 2 as a subscript refer to the replication sample.

[13]A review of the results on class size is fully detailed in a recent Educational Research Service Study (1978) which concludes: "There is general consensus that the research findings on the effects of class size on pupil achievement across all grade levels are contradictory and inconclusive." (p.68)

REFERENCES

Berk, R., "A Comment on Summers and Wolfe," Journal of Human Resources, Summer 1977, XII:3, 401-405.

Chamberlain, G. and Griliches, Z., "Unobservables with Variance-Components Structure: Ability, Schooling, and the Economic Success of Brothers," International Economic Review, June 1975, 16:2, 422-450.

Christ, C., Econometric Models and Methods, New York 1966.

Educational Research Service, Inc., "Class Size: A Summary of Research," Research Brief, Virginia 1978.

Feige, E., "The Consequences of Journal Editorial Policies and a Suggestion for Revision," Journal of Political Economy, 1975, 83:6, 1291-1296.

Frane, J., "Some Simple Procedures for Handling Missing Data in Multivariate Analysis," Psychometrika, September 1976, 41:3, 409-415.

Gourman, J., The Gourman Report, Phoenix, 1967.

Heckman, J., "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," mimeo, Univ. of Chicago 1976.

Summers, A. and Wolfe, B., "Intradistrict Distribution of School Inputs to the Disadvantaged: Evidence for the Courts," Journal of Human Resources, Summer 1976, XI, 328-342.

Summers, A. and Wolfe, B., "Do Schools Make a Difference?" American Economic Review, September 1977, 67:4, 639-652.

Summers A. and Wolfe, B., "Reply," Journal of Human Resources, Summer 1977, XII:3, 406-409.

Summers , A. and Wolfe, B., "Estimating Household Income from Location," Journal of the American Statistical Association, 1978, 73:362, 288-292.

Taubman, P., Sources of Inequality of Earnings, Amsterdam 1977.

Theil, H., Principles of Econometrics, New York 1971.