DOCUMENT RESUME

ED 164 659 UD 018 909

AUTHOR Moock, Peter R.; Rhodes, David
TITLE Achievement in the Philadelphia Public School System.
INSTITUTION Columbia Univ., New York, N.Y. ERIC Clearinghouse on the Urban Disadvantaged.
PUB DATE 78
NOTE 7p.
AVAILABLE FROM Institute of Urban and Minority Education, Box 46, Teachers College, Columbia University, New York, New York 10027 ($1.00)
JOURNAL CIT IRCD Bulletin; v13 n3 Sum 1978

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Academic Achievement; *Black Students; Educational Assessment; Educational Research; Elementary Secondary Education; *Performance Factors; *Research Methodology; *Research Reviews (Publications); Student Motivation; Teacher Role
IDENTIFIERS Pennsylvania (Philadelphia); *Summers (A A); *Wolfe (B L)

ABSTRACT

An achievement study by Summers and Wolfe, published in 1975, is analyzed in this paper. The data set used in the study (three samples of students from 103 elementary, 42 junior high schools and 5 senior high schools) is described. Prior achievement research is outlined as it pertains to the Summers and Wolfe study. The main conclusions of the study listed are: (1) a student's motivation affects his or her learning; (2) instructors with 3 to 9 years of experience are particularly effective, but those having more than 10 years of experience reduce the rate of learning in mathematics; (3) "Head Start" participation does not improve a child's achievement growth by the latter half of elementary school; (4) elementary school students who test at grade level or lower perform better when they are with more high achieving students; and (5) black students perform better in smaller elementary schools and in junior high schools with larger black populations. Methodological errors in the study are pointed out, as are inconsistencies between tables and text, in an effort to dispel any misinterpretations of what is generally considered to be an influential piece of educational research. (Author/WI)

Bulletin

A Publication of the Institute for Urban and Minority Education
Teachers College, Columbia University

Volume XIII, No. 3 Summer 1978

Achievement in the Philadelphia Public School System

Peter R. Moock and David Rhodes

The Editor of the **IRCD Bulletin** *wishes to express appreciation to his colleague Professor Moock and to Mr. Rhodes for this critique of the Summers and Wolfe study, one version of which was published in the* **IRCD Bulletin** *Volume XI (Summer 1976). We share their concern that this study, which is likely to have an important influence on public policy relative to education, be carefully and substantively examined and that its conclusions be critically reviewed before this study is used to inform public policy.—E.W.G.*

* * *

In February of 1975, the Federal Reserve Bank of Philadelphia published in its *Business Review* a study by Summers and Wolfe (1975), of achievement in the Philadelphia school system.[1] Although Summers and Wolfe have used the same data set as the basis for other articles,[2] we wish to focus our attention on the *Business Review* study and on a companion technical paper presented at the 1974 winter meetings of the Econometric Society.[3]

Our review of these papers begins with a description of the data set. Following this is a brief discussion of the educational production function literature, which places Summers and Wolfe's research in a broader context. Our review continues with a critique of the authors' methodology and concludes with a discussion of some inconsistencies between the texts of the two papers and the regression results.

(continued on page 8)

Peter R. Moock is Associate Professor of Economics and Education at Teachers College, Columbia University, and coauthor of a forthcoming book about the study of educational production (Bridge, Judd, and Moock, 1978). **David Rhodes** is President of the School of Visual Arts in New York City and is a candidate for the Ph.D. degree in economics and education at Teachers College. The authors wish to thank William Garner, Victor Levine, and students in the Teachers College Seminar in the Economics of Education for their comments on earlier versions of this review.

 continued from page 1

I. The Data Set

The data set consists of three samples of students from 103 elementary schools, 42 junior high schools, and 5 senior high schools in the Philadelphia public school system. The elementary and junior high samples were randomly chosen, whereas the senior high sample was chosen to ensure a sufficiently high proportion of low income and black students. For each student in the study a two- or three-year school history was prepared. In addition to measures of achievement at the beginning and end of this period, data are available on a student's individual and family background and on the teacher and school resources to which the student was exposed during this period.

Unlike many other educational production function studies, this one by Summers and Wolfe does not use data aggregated at the school or district level. Instead, all the variables are measured at the level of the student, the student's own home, own classroom, or own school, whichever is appropriate. Furthermore, in our opinion, Summers and Wolfe have made a significant contribution to the educational production function literature by not assuming that the effects of particular school inputs are the same for students who differ by race, income, or initial achievement. Although some studies have allowed for the possibility of interaction between school inputs and a student's race or socioeconomic status, usually by partitioning the student population by one or both of these factors, the study by Summers and Wolfe provides one of the most interesting discussions of and explicit tests for interactions among these variables. Clearly, this kind of research holds promise for the future.

II. The Relation to Prior Research

Although some isolated examples of empirical studies aimed at identifying the determinants of student achievement appeared prior to 1966,[4] the major impetus for this genre of research was the publication in that year of the U.S. Office of Education study, *Equality of Educational Opportunity*, better known as "The Coleman Report" (Coleman et al., 1966). Thus, most of the input-output research in education has occurred in the past decade. A problem characteristic of most input-output research is the paucity of theory underlying the empirical analysis. Partly because the goals of education are elusive and partly because the process whereby any one of these goals is achieved certainly involves a very complex pattern of individual, family, and school contributions, the understanding of this process by educational researchers is not very advanced. In the absence of clear, theoretical guidelines, current researchers of "educational production" must try many alternative specifications of the relationship between educational inputs and achievement, until they find ones that seem to "fit the data." In no sense are they testing hypotheses; they are generating hypotheses, which could be tested later in replication studies.

Morever, because the data for these studies are obtained from school settings in which the assignment of students to particular schools, and to particular classrooms within schools, reflects economic circumstances (e.g., students normally attend local schools), pedagogical considerations (e.g., in many schools students are "tracked" according to earlier achievement levels), and political decisions (e.g., funds are allocated according to politically determined criteria), it is difficult for researchers to distinguish causal relationships from others in which "inputs" serve as proxies for factors not included in the estimating equations. In experimental research, subjects are assigned randomly to different treatments. In education, however, explicit experiments are seldom tolerated, for political and ethical reasons, and researchers must try to untangle the effects of the many differences that occur "naturally" among students. A variable which does not affect achievement, but which happens to be correlated with the true cause(s) of achievement not in the analysis, will be *correlated* with achievement, and this statistical relationship can easily be misinterpreted as causal. Researchers using nonexperimental data should be very sensitive to the danger of misinterpreting the statistical effects of proxy variables, especially in the absence of ex ante hypotheses.

Although if asked the question directly, Summers and Wolfe would certainly agree that no causal relationships are "proved" in their study, still nowhere in the body of the *Business Review* article, which was prepared for wide distribution,[5] do they offer the appropriate caveats to the reader. Instead, they relate their conclusions, usually without qualification:

> A student's motivation affects his or her learning [1975: 11].

> Instructors with three to nine years of experience are particularly effective, but those having more than ten years actually reduce the rate of learning mathematics [1975: 14].

> Headstart participation does not improve a child's achievement growth by the latter half of elementary school [1975:14].

> Elementary school students in the sample who test at grade level or lower perform distinctly better when they are with more high-achieving students. Students performing above their grade level are not particularly affected [1975: 17].

> Black students perform better in smaller elementary schools and in junior highs with larger Black populations [1975: 19].

In the *Business Review* article, only in an appendix that explains regression coefficients and their standards errors, do the authors state, albeit indirectly, that they have "mined the data," that they discuss only their final, "best fit" results, and that their conclusions should be regarded as highly tentative:

> Strictly speaking, this [declaring a coefficient "statistically significant"] only applies to one testing of a hypothesized relationship. Since we engaged in extensive empirical testing and refining of these relationships, we do not use these results rigidly, but use them only as general indicators of statistical significance [1975: 23].

In the 1974 paper and in a recent article in the *American Economic Review* Summers and Wolfe (1977) admit more clearly the extent to which hypotheses are used to support data and not vice versa:

> The data have been mined, of course. One starts with so few hypotheses convincingly turned up by theory that classical hypothesis testing is in this application sterile. The data are there to be looked at for what they can reveal.
>
> The standard tests of significance provide guidance of only a very crude sort—hence, the usual asterisks are missing from the *t*-statistics in the tables [1977: 642].

Summers and Wolfe are quite correct in qualifying their results in this way, but in restricting these comments to their academic publications and omitting them from the popular version of their study, they have chosen not to warn the audience more vulnerable to statistical deception. Such warnings are clearly appropriate to guard against the unwarranted use of preliminary research findings for policy purposes.

III. Methodological Considerations

In addition to stating their conclusions in terms that we feel are too strong, Summers and Wolfe may have committed certain methodological errors. Two methodological issues are discussed below. We may assume that criticism along these lines was anticipated by the authors themselves, since they discuss both of these issues in one or more of their published papers. We are surprised, however, to find very few modifications of their analysis between the first publication (1974) and the latest (1977).

A. Only Current-Year Teacher Information Included in the Estimating Equations

For each of the three age-defined samples, the authors attempt to explain variance in achievement gains of students over a two- or three-year period. Yet the measures of inputs relating to the teachers of individual students are for the final year only of the two- or three-year period. This point is made explicitly in neither the *Business Review* article nor in its companion paper, although it is acknowledged in a more recent article, in a footnote that contains an ex post justification for the exclusion of earlier teacher data (1977: 644, fn. 13).

Surely, there are no theoretical grounds for asserting that what a student learns, say, between 1976 and 1978 depends on the qualities of his or her 1977-78 teacher(s) but not on the qualities of his or her 1976-77 teacher(s). Moreover, the procedure used by Summers and Wolfe flies in the face of one of their own repeated claims for distinction, vis-à-vis most other studies of educational production:

> A two- or three-year school history was compiled for each student, which was then matched with data on school-wide resources of the school he or she attended, with his or her family income..., and with data on his or her individual teachers. We were able, therefore, to...look at pupils longitudinally...[1974: 3-4].

On the contrary, in the statistical analysis only school enrollment, library facilities, disruptive incidents, student body characteristics, and a student's attendance behavior are averaged over the two- or three-year period. All the remaining independent variables reflect the final year only. Although family income (and most school characteristics) tend not to change much over time, and estimates for different years will be highly correlated, a student's teachers may look very different from one year to the next. Summers and Wolfe are incorrect in implying that such differences are incorporated into their analysis.

B. Use of a Change Score as the Dependent Variable

The second issue to be discussed in this section of our review is Summers and Wolfe's choice of dependent variable. For each of the three samples, Summers and Wolfe's output measure is the student's score on some cognitive achievement test minus his or her score on the same test three (or, in the case of the eighth-grade sample two) years earlier. Educational researchers were warned against using change scores in a well-known [illegible] by Cronbach and Furby (1970), and Summers and Wolfe seem aware of some of the problems that could be involved:

> This formulation, it has been argued, is erroneous because the differences between initial and final scores regress to the mean—that is, because tests have random error, there will tend to be a negative correlation between initial score and change in achievement. The concern is, of course, that if initial achievement is omitted from the right-hand side, the estimates of all the coefficients of variables correlated with initial achievement will be biased [1977: 641, fn. 8]

Summers and Wolfe argue, however, that their results are generally not biased:

> ...when the regressions were run with and without the initial scores, with two exceptions...the variables which were significant did not lose significance, and the changes in the coefficients were, with the two exceptions, not large enough to alter any of the broad conclusions drawn from the study [1974: 5].

When the output of an educational production study is defined as the gain in student achievement, there is no happy solution to the problems that arise. On the one hand, the omission of initial achievement from the right-hand side of the equation can bias the least-squares estimators. On the other hand, while there may be statistical justification for including initial achievement on the right-hand side, and Summers and Wolfe do so to measure the extent of the bias when they omit initial achievement, there is little theoretical justification for treating the initial score as a predicator of the change in scores.

There is, however, a simple and relatively problem-free alternative. This is to make final achievement the dependent variable and to retain initial achievement as one of the independent variables. Then, the estimators are unbiased, and the regression coefficient for any other input (say, teacher experience) has a straightforward interpretation. It is the expected increment in output (final achievement) attributable to an additional unit of the input (one year of teacher experience) controlling for the student's initial achievement level.[6] Although Summers and Wolfe claim in their latest publication that this formulation is "less satisfactory," they offer no substantive justification for this claim (1977: 641).

IV. Inconsistencies between Tables and Text

All of our criticisms to this point may be controversial in that they reflect, to some extent, our idiosyncratic opinion as to what constitutes sound research or adequate discretion. There are, however, more serious problems to be found in Summers and Wolfe's research. In several instances, their discussion fails to reflect the results presented in their own statistical tables.[7] To see these tables, which were not included in the *Busines Review* article, the reader is referred to the paper that Summers and Wolfe presented at the meetings of the Econometric Society (1974: 33-68).[8]

The inconsistencies between tables and text are of two kinds. There are errors of omission. For example, although the number of latenesses appears in the best fit regression tables, the direction of the effect is not as one would reasonably predict. This unhappy result is not mentioned by Summers and Wolfe in either of the texts reviewed here. In addition to errors of omission, we find several errors of commission, i.e., statements by Summers and Wolfe that obscure or quite simply invert other unhappy results. We present three of these by way of example.

A. Disruptive Incidents—Sixth Grade

In Table D-1 of Summers and Wolfe's paper (1974: 40), we find an estimate of the effect of disruptive incidents on the change in a sixth-grade student's achievement. The relevant portion of the regression equation (rows 26 and 27) is reproduced here as follows:[9]

$$X_0 = \underset{(4.04)}{1.93X_1} \quad \underset{(-3.36)}{.05X_1X_2} + \ldots,$$

where X_0 = student's sixth-grade composite score minus third-grade composite score,[10] $\bar{X}_0 = 20.3$,

X_1 = annual number of disruptive incidents in school, $\bar{X}_1 = 2.89$, $SD_1 = 2.31$

X_2 = student's third-grade composite score, $\bar{X}_2 = 32.1$, $SD_2 = 9.7$.

From this equation, we see that the estimated effect of disruptive incidents on a student's achievement *growth* is a function of the student's earlier achievement *level*:

$$\frac{\partial X_0}{\partial X_1} = 1.93 - 0.5X_2.$$

This implies that the effect of disruptive incidents is *positive*, at least for many students.[11] The critical score on the initial achievement test (the score at which the effect of disruptive incidents turns from positive to negative) is 38.6, or six and one-half months above the sample mean:

$$\frac{\partial X_0}{\partial X_1} \quad 1.93 \quad .05X_2 \quad 0,$$

$$X_2 = 38.6.$$

In other words, for students whose initial score is smaller than 38.6, disruptive incidents are associated with *greater*, not smaller, achievement growth. Only the highest achievers (fewer than half the sample) appear to be affected negatively by disruptive incidents. In view of this finding, we are puzzled by Summers and Wolfe's contention:

> In elementary, junior high, and senior high schools, a greater occurrence of harsh incidents lowers the achievement growth of high-achievers significantly, but low achievers are affected much less. School policies which help reduce the number of serious incidents, then, will improve the learning of testable skills of middle- and high-achieving students [1975: 18].

Not only are these statements inconsistent with the results for elementary school students, the variable (number of disruptive incidents) does not even appear in the best fit equation for senior high students (Summers and Wolfe, 1974:48).

Only recently have Summers and Wolfe set the record straight although, in doing so, they do not retract explicitly their earlier, misleading statements:

> Finally what is one to say about the finding that for students who are at or below grade level, more Disruptive Incidents...are associated with greater achievement growth? It is only for those well above grade level that we found the expected negative effect. This may well be a case where very different results would be obtained if the data were available on a classroom-specific basis. If disruptive incidents only broke out in classrooms where achievement growth barely occurred, then spreading the impact among all classes may have produced this anomalous effect. In any case, it would seem a bit premature to engage in a policy of encouraging disruptive incidents to increase learning! [1977: 647]

B. Gourman Rating—Eighth Grade

In Table D-2, rows 11 and 12 (Summers and Wolfe 1974:44), we find a regression estimate of the effect on achievement growth of the Gourman rating of the undergraduate institution attended by a student's social studies teacher:

$$X_0 = \underset{(-1.63)}{-.03X_1} + \underset{(1.73)}{.0006X_1X_2},$$

where X_0 = student's eighth-grade composite score minus sixth-grade composite score,[12] $\bar{X}_0 = 11.7$,

X_1 = eighth-grade social studies teacher's undergraduate college, Gourman rating, $\bar{X}_1 = 415.84$, $SD_1 = 61.27$

X_2 = student's sixth-grade composite score, $\bar{X}_2 = 53.2$, $SD_2 = 13.3$.

The correct interpretation of the regression results is similar to that for disruptive incidents in the sixth grade. Once again, the effect of the particular input differs depending on the student's earlier achievement level. In this case, the statistical effect of the teacher's Gourman rating is positive for students who achieved an initial score of 50 or better (about 52 percent of the sample), but it is negative for those who achieved at lower levels.

Here again, we submit that Summers and Wolfe's published remarks are misleading, at best. "In junior high school, having attended schools with higher Gourman ratings seem to matter very little in the teaching of English or mathematics," They conclude, apparently because the Gourman ratings for these two teachers do not appear in their final, best fit equation, "but it helps in the teaching of social science—especially for high-ability students" (1975:14). On the contrary, if a higher Gourman rating helps at all, according to their results it is *only* for high-ability students.

Moreover, even if we believed this result to be true, we should be reluctant to accept the apparent policy implication without first carefully examining the possible impact on the full range of educational outcomes, not just one aspect of cognitive achievement. Summers and Wolfe seem less cautious. They conclude, "Net output may be increased by targeting . . . higher-rated college background to the appropriate students" (1975:14). Do Summers and Wolfe mean to tell school administrators that teachers from lower-rated colleges should be assigned to the lower-achieving students in a school or district? In making this recommendation, they ignore the affective outcomes of education. In addition, the recommendation is based on results that are statistically significant only for the very dullest and for the very brightest of students in their sample; for the majority, the relationship between achievement and Gourman rating is not significant (see Table D-2, 1974:47). Recommending such a policy on the basis of nonexperimental data would seem rash, even if the relationship observed were statistically significant. To make this recommendation based on nonsignificant results seems foolhardy.

C. Class Size—Twelfth Grade

In Table D-3, rows 12 and 13 (Summers and Wolfe 1974:48), we find a regression estimate of the effect of class size on the change in a student's reading scores between the ninth and twelfth grades:

$$X_0 = -1.13X_1 + .04X_1X_2 + \ldots,$$
$$(-2.23) \qquad (4.11)$$

where X_0 = student's twelth-grade reading score minus ninth-grade reading score,[11] $\bar{X}_0 = 2.26$,

X_1 = number of pupils in English class, $\bar{X}_1 = 24.68$, $SD_1 = 2.52$,

X_2 = student's ninth-grade reading score, $\bar{X}_2 = 22.93$, $SD_2 = 21.42$.

Once again, the effect of an input depends on a student's earlier achievement level. In this case, although the effect of an additional student in class is negative for most of the sample, the equation indicates that the effect is positive for a student whose ninth grade score was greater than 28.25. For the sample on which this equation is based, approximately 30 percent fall into this relatively high-achieving group. Moreover, the test of significance indicates (see 1974:50) that only the positive effect of class size for high achieving students is significantly different from zero.

One must question, therefore, Summers and Wolfe's assertion that students in smaller classes progress at a higher rate regardless of ability. "Senior high English classes that did not exceed 26 had the highest learning rates (for students of any ability described by the sample); low-ability students benefited most from smaller classes" (1975:12). The specification of 26 students as the critical class size seems to be a non sequitur. It does not follow from the regression results. Moreover, in this sentence Summers and Wolfe have changed the unit of analysis from the individual student to the class of students. Finally, as already shown, their regression results do not indicate a negative relationship between class size and student achievement growth except for students with low levels of initial achievement, and for these students the relationship is not statistically significant.

Concluding Remarks

The study by Summers and Wolfe of achievement in the Philadelphia public school system has attracted considerable attention, much of it deserved. We too applaud the detailed data set used in the analysis of educational production and the attention given to possible interactions between school inputs and student background.

Because the study by Summers and Wolfe is widely known and considered seriously by both scholars and educational policy makers, we believe that it is incumbent upon others engaged in this type of research to examine it for flaws and to guard against misinterpretations of the kind that could lead to inappropriate policy. We suspect that Summers and Wolfe, in reporting their research, may have exaggerated some findings and misinterpreted others. We have tried, in this review, to communicate our suspicions.

Notes

1. The article, entitled "Which School Resources Help Learning? Efficiency and Equity in Philadelphia Public Schools," was republished in the *IRCD Bulletin*, Volume XI (Summer 1976). Copies of the *IRCD Bulletin* are available from the Institute for Urban and Minority Education at $1.00 each.
2. See Summers and Wolfe, "Intradistrict Distribution of School Inputs to the Disadvantaged: Evidence for the Courts" (1976) and "Do Schools Make a Difference?" (1977). These articles are not stressed in the discussion that follows, the first because it deals with a somewhat different topic, and the second because it concentrates on only one of Summers and Wolfe's three Philadelphia samples (see below).
3. The technical paper, entitled "Equality of Educational Opportunity Quantified: A Production Function Approach," is available on ERIC microfiche. ED 107 736.
4. See, for example, the study by Mollenkopf and Melville (1956).
5. Copies of the *Business Review* article are available at no charge from the Public Information Department of the Federal Reserve Bank, Philadelphia, Pa. 19105.
6. For examples of researchers who use this procedure, see Hanushek (1972:11-52), Murnane (1975), and Winkler (1975). Another procedure, essentially equivalent to this one, involves two regression

runs. In the first, final achievement is regressed on initial achievement, and the residual of this regression (the difference between a student's observed final achievement and predicted final achievement) is regressed on other inputs in the second. Again, the regression coefficient for, say, teacher experience can be interpreted as the contribution of an additional year of teacher experience to a student's final achievement, net of that portion of final achievement that has been explained by initial achievement.

It is, of course, possible that the interpretations are correct and the tables are wrong, i.e., contain typographical errors. Although this might account for some of the inconsistencies, it seems implausible that it could account for all of them.

8 To repeat, this paper is available on ERIC microfiche ED 107 736.

9 Beneath each regression coefficient, the t statistic is reported, in parentheses. In defining the variables, we also report the means and standard deviations, to the extent that these descriptive statistics have been supplied by Summers and Wolfe.

10 The scores are expressed in grade equivalent units, and the test is the Iowa Basic Skills Test.

11 Clearly, we are using the term "effect" in a purely statistical sense. Given its positive association with achievement growth, the number of disruptive incidents reported for a school would seem to be serving as a proxy for some omitted input into the educational production function, perhaps for the thoroughness of the school's administrative staff.

12 See note 10.

13 The scores are national percentile rankings, and the test is either the California Aptitude Test or the Comprehensive Test of Basic Skills.

REFERENCES

Bridge, R.G., C. M. Judd, and P.R. Moock. *The Determinants of Educational Outcomes: The Impact of Families, Peers, Teachers and Schools.* Cambridge, Mass.: Ballinger Publishing Co., 1978.

Coleman, J.S., et al. *Equality of Educational Opportunity.* Washington, D.C.: U.S. Government Printing Office, 1966.

Cronbach, L.J., and L. Furby. "How Should We Measure 'Change'—Or Should We?" *Psychological Bulletin* 74 (1970):68-80.

Hanushek, E. *Education and Race: An Analysis of the Educational Production Process.* Lexington, Mass.: Lexington Books, 1972.

Mollenkopf, W.G., and S.D. Melville. "A Study of Secondary School Characteristics as Related to Test Scores." Princeton, N.J.: Educational Testing Service, 1956.

Murnane, R.J. *The Impact of School Resources on the Learning of Inner City Children.* Cambridge, Mass.: Ballinger Publishing Co., 1975.

Summers, A.A., and B.L. Wolfe. "Equality of Educational Opportunity Quantified: A Production Function Approach." Paper presented at the Econometric Society Winter Meeting, December 27-30, 1974. Available on ERIC microfiche, ED 107 736.

———. "Which School Resources Help Learning: Efficiency and Equity in Philadelphia Public Schools." *Business Review* of the Federal Reserve Bank, February, 1975. Reprinted in *IRCD Bulletin* 11 (Summer 1976).

———. "Intradistrict Distribution of School Inputs to the Disadvantaged: Evidence for the Courts." *Journal for Human Resources* 11 (1976): 328-42.

———. "Do Schools Make a Difference?" *American Economic Review* 67 (1977):639-52.

Winkler, D.R. "Educational Achievement and School Peer Group Composition." *Journal of Human Resources* 10 (1975): 189-204.

IRCD BULLETIN

Edmund W. Gordon, Editor

Teachers College, Columbia University
Box 40
525 West 120th Street
New York, New York 10027

Volume XIII, Number 3