*An Exchange: Part I**

Does Money Matter? A Meta-Analysis of Studies of the Effects of Differential School Inputs on Student Outcomes

LARRY V. HEDGES RICHARD D. LAINE ROB GREENWALD

Research on educational production functions attempts to model the relation between resource inputs and school outcomes such as educational achievement. Over the last decade a series of influential reviews of this literature have suggested that there is no systematic relation between resource inputs and school outcomes when controlling for student characteristics such as socioeconomic status. The inference procedure used in these reviews, vote counting, is known to be problematic. This study is a reanalysis of data from these earlier reviews, using more sophisticated synthesis methods. It shows systematic positive relations between resource inputs and school outcomes. Moreover, analyses of the magnitude of these relations suggest that the median relation (regression coefficient) is large enough to be of practical importance. While this reanalysis suggests that previous data do not support the conclusions that Hanushek and others derived from it, limitations of their data set warrant caution in using it for policy formation.

Educational Researcher, Vol. 23, No. 3, pp. 5-14.

Although public dialogue about the need for improvement in the American educational system has vigorously continued over the past decade, no consensus exists on what course should be followed or even on the need for additional expenditures to produce the desired improvements. In fact, some scholars have questioned whether there is a relation between the amount of resources and level of accomplishment of students in schools (Hanushek, 1981, 1986, 1989, 1991). Given limited state budgets and questions about the efficacy of public schools, evidence that school expenditures are unrelated to student performance has deflected attention from the question of revenue sufficiency in discussions about how to improve education. This work has been the pillar upon which the counterintuitive notion that money does not matter in schools has been constructed. Yet the data upon which this conclusion is based support exactly the opposite conclusion and demonstrate that expenditures are positively related to school outcomes.

This article is a reanalysis of the evidence examined by Hanushek in his influential synthesis of the literature on the relations between school inputs and student performance. Although there are weaknesses in some of the studies Hanushek utilized, for the purposes of this article, we accept the validity of Hanushek's formulation of the problem to be investigated (e.g., his definition of the characteristics of schools, families, and student output; the general model specification that he required;[1] and his sample of studies[2]). Setting aside the weaknesses of Hanushek's sample of studies, we focus on the data analysis and interpretation stages of Hanushek's research synthesis. The key question we examine is whether Hanushek's conclusions about the lack of relation between school inputs and student outcomes, and particularly his general conclusion that per pupil expenditure (PPE) and achievement are unrelated, are supported by a synthesis of the results of his sample of studies using more sophisticated statistical methods.

Education Production Functions as Models

The dominant paradigm utilized in analyzing the effects of educational resources on student outcomes over the last few decades has been the education production function. Production function studies attempt to derive a model of the relation between educational inputs and outcomes.[3] The goal of such studies is to develop quantitative models that allow one to predict the effect on student outcomes of a given change in resources. For example, a production function could predict how much the mean achievement on a standardized test would change if per pupil expenditures were increased by $100. Similarly, it might enable the identification of the most cost effective means[4] to improve public education.

Attention was focused on the production function approach by the landmark study *Equality of Educational Opportunity* (often referred to as the Coleman Report) (Coleman et al., 1966). Although the Coleman Report is the best known study of this type, there have been a large number

LARRY V. HEDGES *is Stella W. Rowley Professor, Department of Education, University of Chicago, 5835 South Kimbark Ave., Chicago, IL 60637. He specializes in statistics and research synthesis.* RICHARD D. LAINE *is executive director, Coalition for Educational Rights, 200 North Michigan Ave., Suite 501, Chicago, IL 60601. He is also a PhD student at the University of Chicago specializing in educational finance policy.* ROBERT GREENWALD *is a Searle Fellow, specializing in school finance litigation at the University of Chicago.*

*Part II, a reply by Eric A. Hanushek and a rejoinder by Hedges, Laine, and Greenwald, will appear in the May issue.

of other studies and synopses (see Averch, Carroll, Donaldson, Kiesling, & Pintero, 1972; Glassman & Biniaminov, 1981; Heim & Perl, 1974; Murnane, 1981). Although broad guidelines regarding the specification of models have evolved, no consensus has emerged on the exact specification of the educational production function and, more specifically, on how to measure school resources, student characteristic variables, or the outcome measures used in these production functions.

Hanushek's Sample of Studies

Hanushek has completed the most comprehensive and important synthesis of the existing research on educational production functions to date. In a series of articles over the past decade, Hanushek utilized data from 38 different articles and books and focused on seven school inputs: teacher/pupil ratio, teacher education, teacher experience, teacher salary, PPE, administrative inputs, and facilities. Each publication may have contained several regression equations involving different samples of students, input variables, or outcomes. In all, the 38 publications included a total of 187 equations. Each equation may have contained one or more partial regression coefficients that were included in the vote count.

Hanushek classified the relation between each input variable and an output variable into one of five categories according to the direction of the regression coefficient's sign (positive or negative) and its statistical significance (significant or nonsignificant). The fifth category contained coefficients that were nonsignificant but for which it was impossible to determine the direction of the coefficient's sign from the reported results. Due to the limited information provided by this fifth category, it is deleted in our reanalysis of Hanushek's research.

Hanushek's Conclusions

Through his tally method, Hanushek showed (see Table 1) that as as few as 7% (8 of 113)[5] of the relations between an input and an outcome were positive and statistically significant (in the case of teacher education), and at most 29% (40 of 140) were positive and statistically significant (in the case of teacher experience). Hanushek (1989) concluded that relatively few of the studies yielded results that were both positive and statistically significant and that "the results are startlingly consistent in finding no strong evidence that teacher-student ratios, teacher education, or teacher experience have the expected positive effects on student achievement.... Administration and facilities also show no systematic relationships with performance" (p. 47). When he examined the relation between the most general measure of resource inputs, PPE, and student outcome, he found 20% (13 of 65) of the relations to be positive and statistically significant. From this percentage, Hanushek came to his now widely cited conclusion that "there is no strong or systematic relationship between school expenditures and student performance" (Hanushek, 1989, p. 47).

Problems With the Hanushek Review

Before we explore the statistical problems of combining evidence across studies, we analyze Hanushek's conclusions on a more basic level. The pattern of results given in Hanushek's vote count is not consistent with the null hypothesis of no effect in every study.

For example, if PPE and educational outcome were truly unrelated in every study, 50% of the studies would be expected to obtain positive relations and 50% negative relations by chance. In addition, only 5% of the studies would obtain results that were statistically significant, with that 5% evenly divided between positive and negative relations. For the portion of Hanushek's data in which the direction could be determined, far greater percentages of the results are positive (as much as 70%, in the case of PPE) than one would expect from chance alone.[6] Moreover, a far greater percentage of all of his coefficients are statistically significant. Studies with either positive or negative significant relations between a resource input and educational outcome were found in as few as 12%[7] of the studies (for teacher education), and as many as 35% (for teacher experience), including those studies with unknown signs. Thus, the percentage of studies having significant coefficients ranged from 2.3 to 7 times the 5% that would be expected due to chance alone, if resource inputs and outcomes were truly unrelated. This basic statistical analysis undermines Hanushek's conclusion that these data prove that there is no systematic relationship between expenditures and performance.

The analytic method that Hanushek used is often called vote counting (Light & Smith, 1971). For a given resource input, the "result" of each study is the estimated partial regression coefficient of the resource input on student output, holding constant family background and other inputs. The vote-counting methodology tabulates these results according to the sign (positive, negative, or unknown) and the statistical significance (significant or nonsignificant, usually at the $\alpha = .05$ level) of the results. The category with the most results (the most "votes") is generally taken to represent the true state of the relation in question.

Despite the intuitive appeal of vote counting as a summary procedure, it has serious failings as an inference procedure. The most obvious shortcoming is that regardless of whether vote counting can identify if a relation exists, it cannot provide an indication of its magnitude. More serious weaknesses of the vote count method were identified by mathematical studies of the properties of vote counting as a decision procedure. Even when an effect is present in every study, vote counting typically has very low power to detect effects (it is prone to Type II errors) (Hedges & Olkin, 1980). Moreover when effects are relatively small, which is quite common in social science research, and the individual study sample sizes are small to moderate, vote counting has a paradoxical property: The probability that a vote count will correctly detect an effect that is present in every study tends to zero as the number of studies increases (Hedges & Olkin, 1980).

The structure of Hanushek's argument is essentially one of accepting (at least approximately) a null hypothesis after attempts to reject it have failed.[8] The credibility of such arguments depend on the strength of the methods used in the attempt to falsify the null hypothesis. The probability of making an error in reaching the conclusion to "accept" the null hypothesis depends on the power of the tests used in the attempt to reject it. Because vote counting has such low power as an inference procedure, the failure to rejec a null hypothesis using this procedure is not persuasive evidence that the null hypothesis is even approximate true.

The problem of combining evidence across different em pirical research studies to draw general conclusions, often called meta-analysis or research synthesis, has received considerable attention over the past 15 years (see, e.g., Cooper, 1984; Glass, McGaw, & Smith, 1981; Hedges & Olkin, 1985; Light & Pillemer, 1984). Much of this work has concentrated on developing methods for conducting research syntheses that will reduce the threats to their validity. Some, but certainly not all, of this work has focused on the analytic methods used to combine empirical evidence across studies.

Reanalysis Methodology

We attempted to replicate Hanushek's selection of coefficients to be counted in each input category and then used these coefficients or their p values in our analyses. Table 1 presents the tabulation of results given by Hanushek and those obtained in our reanalysis. The close agreement between our sample tally (specifically in the PPE category) and that of Hanushek, after deleting those tallies that were classified as nonsignificant with an unknown sign, suggests that any differences in overall results should be primarily consequences of the analytic methods used. We also examined a subsample of the studies in which achievement was the outcome measure; this subsample will be discussed later.

There are two general classes of statistical methods in meta-analysis: combined significance tests and combined estimation methods. Both were used in our reanalysis.

Table 1

Summary of the Production-Function Coefficients Utilized in the Analyses

	Significant		Nonsignificant			
Input variable	Positive	Negative	Positive	Negative	Unknown[a]	Total
PPE						
Hanushek	13 (24%)	3 (6%)	25 (46%)	13 (24%)	11 (NC)	65
Reanalysis	13 (24%)	3 (5%)	25 (45%)	14 (25%)		55
Combined significance	12 (34%)	3 (9%)	13 (37%)	7 (20%)		35
Effect size estimation	10 (27%)	1 (3%)	20 (53%)	7 (18%)		38
Teacher experience						
Hanushek	40 (32%)	10 (8%)	44 (35%)	31 (25%)	15 (NC)	140
Reanalysis	39 (30%)	7 (5%)	52 (40%)	33 (25%)		131
Combined significance	34 (32%)	7 (7%)	39 (36%)	27 (25%)		107
Effect size estimation	15 (26%)	2 (4%)	26 (46%)	14 (25%)		57
Teacher education						
Hanushek	8 (11%)	5 (7%)	31 (41%)	32 (42%)	37 (NC)	113
Reanalysis	10 (11%)	6 (7%)	39 (44%)	33 (38%)		88
Combined significance	8 (12%)	5 (7%)	35 (51%)	20 (29%)		68
Effect size estimation	4 (10%)	3 (7%)	13 (32%)	21 (51%)		41
Teacher salary						
Hanushek	11 (24%)	4 (9%)	16 (36%)	14 (31%)	24 (NC)	69
Reanalysis	9 (21%)	4 (9%)	16 (37%)	14 (33%)		43
Combined significance	6 (23%)	3 (12%)	11 (42%)	6 (23%)		26
Effect size estimation	4 (15%)	3 (11%)	10 (37%)	10 (37%)		27
Teacher-pupil ratio						
Hanushek	14 (13%)	13 (12%)	34 (32%)	46 (43%)	45 (NC)	152
Reanalysis	12 (10%)	15 (13%)	44 (38%)	45 (38%)		116
Combined significance	10 (11%)	12 (13%)	39 (42%)	31 (34%)		92
Effect size estimation	6 (9%)	7 (10%)	21 (30%)	35 (51%)		69
Administrative inputs						
Hanushek	7 (19%)	1 (3%)	14 (38%)	15 (41%)	24 (NC)	61
Reanalysis	5 (14%)	2 (6%)	14 (40%)	14 (40%)		35
Combined significance	5 (19%)	1 (4%)	11 (41%)	10 (37%)		27
Effect size estimation	5 (14%)	2 (6%)	14 (40%)	14 (40%)		35
Facilities						
Hanushek	7 (16%)	5 (12%)	17 (40%)	14 (33%)	31 (NC)	74
Reanalysis	7 (9%)	8 (10%)	30 (39%)	32 (42%)		77
Combined significance	6 (11%)	7 (13%)	20 (36%)	22 (40%)		55
Effect size estimation	3 (6%)	4 (9%)	15 (32%)	25 (53%)		47

Note. Percentages may not sum to 100 due to rounding.
[a](NC) used to indicate that figures for nonsignificant unknown are not counted in the percentages.

Combined Significance Tests

A combined significance test is a way of combining statistical significance values (p values) from studies that test the same conceptual hypothesis but that use different designs (e.g., levels of aggregation or specification) or that may measure the outcome variables differently (e.g., standardized tests that differ in subject) to obtain an overall level of significance for the collection of studies under consideration. Such tests allow one to determine if the data are consistent with the null hypothesis in all studies or if the data suggest that the null hypothesis is false in at least some of the studies.

Although there are many combined significance tests, we chose the inverse chi-square (Fisher) method because it has sound technical properties and is widely used in statistics (see Hedges & Olkin, 1985, chap. 3). We tested two null hypotheses for each of Hanushek's resource input variables: (a) the positive case, where the null hypothesis is that there is no positive relation between the resource and output, and (b) the negative case, where the null hypothesis is that there is no negative relation between the resource and output.

Note that combined significance tests are not designed to support inferences about the size of the average effect or about the consistency of effects across studies. It is entirely possible to reject the null hypothesis in both the positive and negative cases, which would mean that there is evidence of both some positive and some negative relations. However, in order to reach the conclusion that "no strong or systematic relationship" exists between the major educational inputs to schooling and student outcome, the data would have to be consistent with the null hypothesis in both the positive and the negative case.

Effect Magnitude Analyses

Because output variables were not measured on the same scale in all studies, the partial regression coefficients for the resource input variables could not be combined directly. Consequently, all of our indices of relations predicted standardized output, that is, output in standard deviation units.

Two of the resource input variables (PPE and teacher salary) were initially measured in dollars and hence were directly comparable, or could be made so after a correction for inflation.[9] For these resource inputs, the measure of relation or effect magnitude was a "half-standardized" partial regression coefficient.[10] The half-standardized regression coefficient measures the number of standard deviations of change in output associated with a one-unit ($1 in this case) change in input.

The other resource inputs were measured in ways that could not readily be placed on the same scale. For these inputs, the index of effect magnitude was the fully standardized regression coefficient. This coefficient measures the number of standard deviations of change in output that would be associated with a one standard deviation change in input.

Robustness Testing

Statistical Dependence

One characteristic of these data that was not explicitly examined in previous reviews is that some of the coefficients counted separately are probably stochastically dependent. One source of this dependence is that some publications estimated essentially the same model for several different measures of outcome (e.g., reading achievement and mathematics achievement) using data collected from the same individuals. Since different measures of outcome from the same individuals are likely to be correlated, the coefficients obtained from them are likely to be correlated as well. A second and much rarer source of dependence in this sample of studies is that some seemingly independent publications appear to use the same data (see Hanushek, 1971, 1972; Levin, 1970b, 1976). To evaluate the possible consequences of dependence, we created a "conservatively independent" sample of estimates by grouping each set of coefficients that were identified as possibly correlated, and then using the median value of the p value or effect size for all further analyses. For example, the study by Cohn and Millman (1975) yielded 12 coefficients for the effect of several of the input variables on outcome in the full sample, one for each of 12 measures of output. In the conservatively independent sample, Cohn and Millman yielded one p value for the effects of PPE (the median of the 12 individual p values) and one effect size for PPE (the median of the 12 individual effect sizes).[11]

Influence of Outliers

Another way of examining the influence of deviant effects is to trim the effects examined so that the very largest and smallest effects are excluded from the analysis. We developed a trimmed sample of p values and effect sizes by deleting, in each analysis, the largest 5% and the smallest 5%, and retaining the middle 90% of the values. This process deleted p values or effect sizes from different publications and thus differs slightly from the "delete 1" robustness sample described below. It also has the property that it deleted symmetrically from both sides of the distribution results.

Influence of Single Publications

Another characteristic of the data that was not previously examined was the degree of influence of any individual publication on overall conclusions. Since different publications contributed different numbers of coefficients, p values, or effect size estimates, it is possible that a single study migh have a disproportionate influence on overall conclusions. While the fact of a disproportionate influence of a single publication does not imply that the results are incorrect, i does raise questions of robustness of conclusions acros studies. To evaluate the dependence of overall conclusions on the influence of a single publication, we developed robustness sample for each input variable by deleting th p values or effect size estimates associated with the singl publication whose deletion would most change the overall results. In most cases, the publication with the larges influence on overall results has several coefficients, this robustness sample has considerably fewer coefficients or p values than does the corresponding complete sample.

Results

The analyses below clearly show systematic positive pat terns in the relations between educational resource inpu and student outcomes. It must be remembered that our co clusions are dependent on the use of the sample pop tion of studies initially identified by Hanushek. Using research synthesis methodology detailed in this article a more recent and more stringently defined sample of production function studies would overcome the weaknesses of Hanushek's universe of studies and would provide a more definitive understanding of the impact of educational inputs on student outcomes.

Combined Significance Tests

The results of the combined significance tests for the coefficients in all studies, and the subsample of studies limited to equations with achievement outcomes, are reported in Table 2. In order to reach conclusions that increasing one

Table 2

Results of Combined Significance Tests

Input variable (outcomes)	Equations (studies)	Sample: Full analysis (df)	Independent (df)	90% (df)	Delete 1 (df)
Positive case (Ho: $\beta \leq 0$)					
PPE					
All studies	55 (11)	187.29 (70)	84.25 (38)	157.80 (62)	83.96 (50)
Achievement	37 (10)	154.09 (54)	84.07 (38)	139.80 (50)	50.75 (34)
Teacher experience					
All studies	131 (25)	822.61 (214)	478.85 (120)	596.92 (194)	696.67 (200)
Achievement	97 (23)	673.30 (168)	437.44 (120)	498.28 (152)	547.37 (154)
Teacher education					
All studies	88 (18)	282.05 (136)	124.67 (60)	203.04 (124)	235.35 (132)
Achievement	61 (16)	196.88 (100)	108.58 (58)	163.93 (92)	149.84 (86)
Teacher salary					
All studies	43 (10)	131.42 (52)	33.27 (18)	99.61 (48)	89.96 (40)
Achievement	24 (9)	93.87 (34)	32.43 (18)	62.05 (30)	50.41 (22)
Pupil-teacher ratio[a]					
All studies	116 (23)	265.06 (184)	94.15 (82)	185.12 (164)	182.22 (168)
Achievement	74 (22)	178.20 (140)	84.93 (82)	122.48 (124)	95.35 (124)
Administrative inputs					
All studies	35 (6)	97.76 (54)	38.28 (18)	84.76 (50)	69.90 (42)
Achievement	18 (6)	53.54 (26)·	40.00 (18)	40.54 (22)	27.80 (14)
Facilities					
All studies	77 (17)	147.92 (110)	83.13 (44)	116.23 (98)	122.73 (104)
Achievement	52 (17)	129.96 (74)	82.17 (44)	108.23 (66)	104.77 (68)
Negative case (Ho: $\beta \geq 0$)					
PPE					
All studies	55 (11)	79.57 (70)	40.60 (38)	44.11 (62)	31.26 (56)
Achievement	37 (10)	44.68 (54)	41.29 (38)	36.23 (50)	28.52 (44)
Teacher experience					
All studies	131 (25)	230.61 (214)	76.79 (120)	130.24 (194)	158.95 (202)
Achievement	97 (23)	181.62 (168)	112.51 (120)	102.99 (152)	133.71 (154)
Teacher education					
All studies	88 (18)	174.53 (136)	79.37 (60)	92.35 (124)	131.17 (122)
Achievement	61 (16)	146.70 (100)	79.21 (58)	88.78 (92)	103.34 (86)
Teacher salary					
All studies	43 (10)	177.61 (52)	77.00 (18)	75.93 (48)	32.21 (48)
Achievement	24 (9)	170.43 (34)	77.21 (18)	68.76 (30)	25.04 (30)
Pupil-teacher ratio[a]					
All studies	116 (23)	289.24 (184)	137.47 (82)	218.34 (164)	232.27 (170)
Achievement	74 (22)	262.22 (140)	148.14 (82)	191.33 (124)	205.25 (126)
Administrative inputs					
All studies	35 (6)	84.97 (54)	6.63 (18)	42.50 (50)	33.70 (42)
Achievement	18 (6)	63.02 (26)	47.37 (18)	20.55 (22)	20.56 (24)
Facilities					
All studies	77 (17)	174.82 (110)	54.89 (44)	114.27 (98)	117.63 (94)
Achievement	52 (17)	118.17 (74)	56.95 (44)	69.45 (66)	61.53 (58)

Note. 90% = middle 90%; delete 1 = delete most influential study.

[a] The signs have been reversed in those studies that utilize the variable class size (pupil/teacher ratio) to be consistent with teacher/pupil ratio so that $\beta > 0$ means that smaller classes have greater outcomes.

of the input variables is associated with improvement in the student outcome measure, we would have to reject the null hypothesis in the positive direction.[12] The data from Table 2 are summarized in Table 3 as either (a) rejecting the null hypothesis (R) or (b) failing to reject the null hypothesis (–).

The panel of results on the left side of Table 3 reports tests of the null hypothesis that there is no positive relation between resource inputs and student outcomes. In other words, these results will tell us if at least one study contains a positive relation between an educational input and student outcome.

Reviewing the combined significance tests for the positive case, we see that almost all of the combined p values are significant at the $\alpha = .05$ level. This result holds in the complete sample ("full") and the conservatively independent sample for both all outcomes and the subsample of outcomes limited to achievement. It holds for every one of the samples containing the middle 90% of the p values except for teacher/pupil ratio and facilities, and for every one of the "delete 1" robustness samples except for facilities, where the results are not significant. Thus, these data imply that over all the studies, with the few exceptions noted above, there are at least some positive relations between each of the types of educational resource inputs studied and student outcome.

The panel of results on the right side of Table 3 concerns tests of the null hypothesis that there is no negative relation between resource inputs and student outcomes. In other words, these results will tell us if at least one study contains a negative relation between an educational input and student outcome. The pattern of the combined significance test results for the negative case is slightly more complex.

For the input variables PPE, teacher experience, and teacher/pupil ratio, none of the combined (full, independent, middle 90%, deletion of the most influential study) p values are significant. This suggests that there is no statistically reliable evidence of negative relations between these resource inputs and student outcome. Taken together with the results of the combined significance tests in the positive case above, these results suggest that since there are positive relations between outcome and PPE, and teacher experience, and teacher/pupil ratio but no negative relations between outcome and these resource inputs, the typical relation is positive. This would suggest that Hanushek's sample of studies supports the conclusion that increasing ex-

Table 3

Summary of Results of Combined Significance Tests

Input variable (outcomes)	Positive case (Ho: $\beta \leq 0$) Sample				Negative case (Ho: $\beta \geq 0$) Sample			
	Full	Indep.	90%	Delete 1	Full	Indep.	90%	Delete 1
PPE								
All studies	R	R	R	R	—	—	—	—
Achievement	R	R	R	R	—	—	—	—
Teacher experience								
All studies	R	R	R	R	—	—	—	—
Achievement	R	R	R	R	—	—	—	—
Teacher education								
All studies	R	R	R	R	R	—	—	—
Achievement	R	R	R	R	R	R	—	—
Teacher salary								
All studies	R	R	R	R	R	R	R	—
Achievement	R	R	R	R	R	R	R	—
Pupil/teacher ratio[a]								
All studies	R	R	R	R	R	—	—	—
Achievement	R	R	R	R	R	—	—	—
Teacher/pupil ratio								
All studies	R	R	—	R	—	—	—	—
Achievement	R	R	R	R	—	—	—	—
Administrative inputs								
All studies	R	R	R	R	R	—	—	—
Achievement	R	R	R	R	R	R	—	—
Facilities								
All studies	R	R	—	—	R	—	—	R
Achievement	R	R	R	R	R	—	—	—

Note. R indicates that the null hypothesis is rejected at the $\alpha = 0.05$ level. A dash indicates failure to reject the null hypothesis at $\alpha = 0.05$
Indep. = independent; 90% = middle 90%; delete 1 = delete most influential study.
[a]The signs have been reversed to be consistent with teacher/pupil ratio so that $\beta > 0$ means that smaller classes have greater outcomes.

341

penditures and teacher experience will increase student outcome.

The resource inputs of teacher education, teacher salary, pupil/teacher ratio, administrative inputs, and facilities have significant combined p values in the negative case for both the sample containing all outcomes and the sample restricted to achievement outcomes. The analysis of the robustness sample shows that data from a single publication are responsible for the negative statistically significant combined effects for all of these variables except for facilities.

These analyses are persuasive in showing that, with the possible exception of facilities, there is evidence of statistically reliable relations between educational resource inputs and school outcomes, and that there is much more evidence of positive relations than of negative relations between resource inputs and outcomes.

Effect Size Analyses

The median effect sizes for each of the independent variables are given in Table 4. The median half-standardized regression coefficient for PPE computed over all studies is .0014. This coefficient is large enough to be of considerable practical importance. It suggests that an increase of PPE by $500 (approximately 10% of the national average) would be associated with a 0.7 standard deviation increase in student outcome. By the standards of educational treatment interventions, this would be considered a large effect. The median effect obtained in studies that used academic achievement as the outcome is even larger and the median effects in the various robustness tests are of comparable magnitude; these are large enough to be educationally important.

The median effects for teacher experience are also positive in each case, although the measures used for teacher experience make the interpretation of the magnitude more difficult. The median effects for teacher salary are generally positive overall, but their magnitude appears to be too small to be of practical importance, except in the case of the conservatively independent sample and the sample with the most influential publication deleted for those studies that use achievement as the outcome variable. This might be interpreted to imply that except for the consequences of nonindependence and of a single influential publication, these results point to a positive effect of teacher salary. A confounding factor that must be considered is the use of both starting and average salaries in the studies that included this variable.

In contrast, the median effects for teacher education are negative in each of the samples of studies. The median regression coefficients for teacher/pupil ratio, pupil/teacher ratio, administrative inputs, and facilities show a mixed pattern of median regression coefficients, sometimes being positive and sometimes being negative. Thus, the effect size analysis for these coefficients does not show a persuasively consistent pattern, except for teacher education.

Taken together, the effect size analyses suggest a pattern of substantially positive effects for global resource inputs (PPE) and for teacher experience. The effects of certain resource inputs (teacher salary, administrative inputs, and facilities) are typically positive, but not always. The typical effects of class size (expressed either as pupil/teacher ratio or teacher/pupil ratio) are decidedly mixed.

Table 4

Median Regression Coefficients

	Equations	Sample			
Input variable	(studies)	Full	Indep.	90%	Delete 1
PPE[a]					
All studies	38 (8)	.0014	.0004	.0014	.0014
Achievement	26 (7)	.0020	.0004	.0020	.0013
Teacher experience					
All studies	57 (10)	.0700	.0400	.0700	.0590
Achievement	28 (9)	.0415	.0400	.0455	.0414
Teacher education					
All studies	41 (7)	−.0200	−.0200	−.0200	−.0200
Achievement	19 (6)	−.0300	−.0200	−.0300	−.0300
Teacher salary[a,b]					
All studies	27 (6)	.0008	.0005	.0008	.0008
Achievement	12 (5)	−.0013	.0366	−.0013	.0390
Pupil/teacher ratio					
All studies	45 (8)	.0600	.0073	.0563	.0619
Achievement	22 (7)	.0150	−.0097	.0150	−.0153
Teacher/pupil ratio					
All studies	24 (6)	−.0010	.0094	−.0010	−.0048
Achievement	16 (6)	.0176	.0210	.0176	.0114
Administrative inputs					
All studies	35 (6)	.0099	.0042	.0099	.0178
Achievement	18 (6)	−.0169	−.0068	−.0169	.0162
Facilities					
All studies	47 (7)	−.0100	.0110	−.0100	−.0100
Achievement	22 (7)	.0150	.0111	.0150	.0200

Note. These results are summaries of fully standardized regression coefficients unless otherwise noted. Indep. = independent; 90% = middle 90%; delete 1 = delete most influential study.
[a]Half-standardized regression coefficients. [b]The values of the coefficients in this category have been multiplied by 100.

It might seem odd that the effects of global resource inputs (PPE) are so clearly positive while the effects for the components are less consistently positive. However, this is not at all contradictory. This pattern of results is consistent with the idea that resources matter, but allocation of resources to a specific area (such as reducing class size or improving facilities) may not be helpful in all situations. That is, local circumstances may determine which resource inputs are most effective, and local authorities utilize discretion in wisely allocating global resources among the areas most in need.

Are the Effect Size Analyses and the Significance Test Analyses Contradictory?

It might strike some as peculiar that the analyses based on combined significance tests suggested a more profoundly positive pattern of results than was obtained in the effect size analysis (at least for all variables other than PPE and teacher experience). This is primarily because the combined significance tests were based on more data than the effect size analyses. Most studies reported enough data (p values) to be included in the combined significance test, but fewer studies reported enough data (standardized regression coefficients) to compute effect sizes. For reasons that are not altogether clear, there was a tendency for studies that yielded positive effect sizes to fail to report enough information to compute their exact magnitude. For example, Table 1 shows that while 63% of the studies of teacher education that appeared in the combined significance test analysis yielded positive results, only 42% of the studies containing effect sizes had positive results.

Potential Weaknesses of This Data Set

Failure to Include Studies With Statistically Nonsignificant Effects of Unknown Sign

The more sophisticated analyses that we used posed greater data requirements than the vote count used by Hanushek. Consequently, some studies did not provide enough data to be included in our analyses. The category of studies that produced statistically nonsignificant results with unknown sign is the most problematic. *P* values associated with these studies, if large and included, could potentially reduce the overall significance of our combined *p* value analyses. Similarly, the effect sizes associated with these studies would generally (though not necessarily) tend to be small in absolute magnitude, and, thus, if they could be included, might reduce the absolute value of the median effect size.

Even assuming that the studies with nonsignificant results of unknown sign produced the *least* favorable results possible, they would not generally overturn the positive findings of the combined significance tests. For example, consider the case of PPE: Even if all of the 11 studies with nonsignificant results of unknown sign produced *p* values that contributed nothing to the combined chi-square statistic, the value calculated in the analysis of the full sample (including the additional 11 studies) would be 187.3 and would still greatly exceed 115.4, the 95% critical value of the chi-square with 92 degrees of freedom. Hence, we would reach the same conclusion: that there is strong evidence of at least some positive effects of PPE on outcome. Similarly, assuming that all 11 studies with nonsignificant effects and unknown sign produced effect sizes of zero (or had a negative value), the median effect size in the full sample would be .0004. This is smaller than the value we obtained, but still large enough to be educationally important.

Weaknesses in the SES Measures Used

As made prominent by the Coleman Report, socioeconomic status (SES) measures play an important role in determining the relationship between inputs and outcomes in education. Yet some of the studies used in Hanushek's analysis employed measures of student background that were weak, at least by today's standards. For example, some used only mother's education as an indicator of SES. In some studies, while outcomes were measured at the student level, SES was measured at the aggregate level. We believe that better measurement of student background is critical to the creation of valid education production functions, and to allowing comparisons of models that differ in the variables employed.

Age of the Data

Many of the data sets used in these studies were collected in the early 1960s. American education has changed a great deal during the last 30 years. It is therefore reasonable to believe that the factors affecting production of educational outcomes, and their relative weights, have changed during the last 30 years.

Most of the Studies Are Cross-Sectional

Most of the studies in Hanushek's data set are cross sectional rather than longitudinal. There is an emerging consensus among methodologists that longitudinal designs are more satisfactory for examining school effects. For example, there is evidence that cross-sectional studies are frequently less sensitive for detecting school effects than their longitudinal competitors. Thus, the collection of studies that we examined may actually underestimate the effects of resource inputs on school outcomes (see Bryk & Raudenbush, 1988). Moreover, because effects in education may be cumulative, the long-term effects of resources may be magnified when analyzed over many years.

Selection (Publication) Bias

Selective reporting of data can compromise their interpretability unless the details of the selection process are well understood and can be adjusted for in a statistical selection model. Unfortunately, selection processes are rarely understood in analyses of real data. The problem of bias due to selection as a part of the publication process has received considerable attention in research synthesis (see Begg & Berlin, 1988; Hedges, 1984; Rosenthal, 1978). A great deal of attention has been directed to publication bias that arises when studies that obtain statistically significant results are more likely to be published than otherwise identical studies that do not obtain statistically significant results. Methods for adjusting for the effects of selection under various models of bias toward statistically significant results h been developed (see, e.g., Hedges, 1984, 1992). Genera these adjustments reduce the magnitude of the estimated effects, but do not affect the sign. Moreover, if the true valu of effects is actually zero (i.e., if the null hypothesis is true) they produce no net adjustment because selection via two-tailed *p* values causes no bias when effects are zero.

Unfortunately, it is easy to posit other plausible selectio mechanisms that are more difficult to model explicitly. Fo example, studies that confirm prevailing beliefs might be viewed as providing less information and therefore havin lower priority for publication because they "merely confir the obvious." If this selection process was operating, selection might have biased the published effects so as to make them too small. Alternatively, one might argue that studie that contradict prevailing views (e.g., that "money does n matter" in the education process) might be subject to greate scrutiny and therefore be less likely to be published. If this selection process was operating, selection might have bias the published effects so as to make them too big, becau studies finding effects with contrary signs would be less likely to be published.

The reality is that selection bias might have effects in eit direction, which makes it difficult or impossible to comple ly rule out selection effects as threats to the validity published research findings. Note, however, that selecti biases the findings of a single published research study profoundly as it does a collection of research studi (Hedges, 1984). Consequently, if one regards selectio as fatally compromising the interpretation of res reviews, one must also accept that it fatally compromi

343

the interpretation of single published research studies, a position that few empirical researchers are willing to accept.

Conclusions

The production function studies of the relation between resource inputs and school outcomes examined by Hanushek do not support his conclusion that resource inputs are unrelated to outcomes. The analytic method he used to synthesize results across studies has low statistical power, and hence his conclusion (accepting the null hypothesis) would seem particularly suspect.

Reanalysis with more powerful analytic methods suggests strong support for at least some positive effects of resource inputs and little support for the existence of negative effects. Effect magnitude analyses suggest that these effects (at least for PPE) are large enough to be of real importance. Moreover, these findings seem to be robust against obvious threats to their validity. While the pattern of effect sizes is most persuasive for global resource variables (PPE and teacher experience), the median effects are positive for most resource variables, with the clear exception of teacher education.

We are not arguing that the studies used by Hanushek are an adequate basis for resolving the question about the magnitude of the relation between resource inputs and school outcomes. We have serious reservations about the age of some of the data and the measurement and design of some of the studies. We are currently conducting a synthesis similar to the one described in this article using a more adequate sample of studies. The conclusions from this research should be able to better answer detailed scientific and policy questions about the likely effects of altering resource inputs.

Even if the conclusions drawn from the studies analyzed in this paper are correct, we would not argue that "throwing money at schools" is the most efficient method of increasing educational achievement. It almost surely is not. However, the question of whether more resources are needed to produce real improvement in our nation's schools can no longer be ignored. Relying on the data most often used to deny that resources are related to achievement, we find that money *does* matter after all.

Notes

Preparation of this article was supported in part by a grant from the Russell Sage Foundation. This article was developed out of a project initiated by Richard Laine. We thank Rochelle Gutiérrez and William S. McKersie for their collaboration in the initial phases of the research. We would also like to thank Eric Hanushek, participants of the MESA Seminar, the Workshops on Education and Social Policy and on Public Sector Economics at the University of Chicago, and three anonymous reviewers for their feedback on earlier versions of this work.

[1] Hanushek defines a qualified study as a "production-function estimate that is: (a) published in a book or refereed journal; (b) relates some objective measure of student output to characteristics of the family and the schools attended; and (c) provides information about the statistical significance of estimated relationships." (Hanushek, 1989, p. 50, footnote 5).

[2] See the Appendix for a list of the studies Hanushek synthesized.

[3] Inputs include school resources such as expenditures, teacher characteristics, facilities, and student characteristics such as socioeconomic status or ability. Outcomes include achievement as measured by standardized tests, future educational patterns, and adult earnings.

[4] Greatest gain in output for a change in input of a specified cost (see Levin, 1970a).

[5] The percentages in this paragraph were calculated by including Hanushek's category of nonsignificant coefficients with an unknown direction of sign, and therefore differ from the percentages listed in Table 1.

[6] Baker (1991) previously noted that there were more positive and statistically significant results for PPE than would be expected from chance alone. Spencer and Wiley (1981) questioned Hanushek's interpretation of the statistical significance of the regression coefficients.

[7] The percentages in this paragraph were calculated by including Hanushek's category of nonsignificant coefficients with an unknown direction of sign, and therefore differ from the percentages listed in Table 1.

[8] While orthodox treatments of the logic of statistical inference do not condone accepting the null hypothesis, the practice is routine and, we would argue, defensible in many circumstances. Our criticism is not with the practice itself, but with the failure to recognize that powerful tests of the hypothesis are needed to make acceptance of the null hypothesis credible.

[9] In this analysis, the dollar value of PPE and teacher salary in each study was adjusted to 1991 dollars using the Elementary and Secondary Price Index for the year in which the data were collected (see U.S. Center for Education Statistics, 1992).

[10] The half-standardized regression coefficient b_H is defined as $b_H = b/S_O$ where b is the unstandardized regression coefficient and S_O is the standard deviation of the output variable.

[11] Although Cohn and Millman report results for 12 outcome measures in their appendix, the last equation appears to have been a duplication of equation 11 and therefore was not included in our analyses (see Cohn & Millman, 1975, pp. 110-120).

[12] Because the studies combined the input variables of teacher/pupil ratio and class size (pupil/teacher ratio) in the teacher/pupil ratio category, we performed both a combined and a distinct analysis of each variable. Since these two variables are inversely related, a positive relation for the variable teacher/pupil ratio and a negative relation for class size (pupil-teacher ratio) correspond to essentially the same conclusion.

Appendix

Studies Providing Data Utilized in the Analyses

Armor, D., Conry-Oseguera, P., Cox, M., King, N., McDonnell, L., Pascal, A., Pauly, E., & Zellman, G. (1976). *Analysis of the school preferred reading program in selected Los Angeles minority schools* (Report No. R-2007-LAUSD). Santa Monica, CA: Rand.

Behrendt, A., Eisenach, J., & Johnson, W. R. (1986). Selectivity bias and the determinants of SAT scores. *Economics of Education Review, 5,* 363–371.

Bieker, R. F., & Anschel, K. R. (1973). Estimating educational production functions for rural high schools: Some findings. *American Journal of Agricultural Economics, 55,* 515–519.

Boardman, A., Davis, O., & Sanday, P. (1977). A simultaneous equations model of the educational process. *Journal of Public Economics, 7,* 23–50.

Bowles, S. (1970). Toward an educational production function. In W. L. Hansen (Ed.), *Education, income and human capital* (pp. 11–60). New York: National Bureau Economic Research.

Brown, B. W., & Saks, D. H. (1975). The production and distribution of cognitive skills within schools. *Journal of Political Economy, 83,* 571–594.

Burkhead, J. (1967). *Input-output in large city high schools.* Syracuse, NY: Syracuse University Press.

Cohn, E. (1968). Economies of scale in Iowa high school operations. *Journal of Human Resources, 3,* 422–434.

Cohn, E., & Millman, S. D. (1975). *Input-output analysis in public education.* Cambridge, MA: Ballinger.

Dolan, R. C., & Schmidt, R. M. (1987). Assessing the impact of expenditure on achievement: Some methodological and policy considerations. *Economics of Education Review, 6,* 285–299.

Dynarski, M. (1987). The Scholastic Aptitude Test: Participation and performance. *Economics of Education Review, 6*, 263–274.

Eberts, R. W., & Stone, J. A. (1984). *Unions and public schools: The effect of collective bargaining on American education.* Lexington, MA: Lexington Books.

Hanushek, E. A. (1971). Teacher characteristics and gains in student achievement: Estimation using micro-data. *American Economic Review, 6*, 280–288.

Hanushek, E. A. (1972). *Education and race: An analysis of the educational production process.* Cambridge, MA: Heath-Lexington.

Heim, J., & Perl, L. (1974). *The educational production function: Implications for educational manpower policy* (Monograph No. 4). Ithaca, NY: Cornell University, Institute of Public Employment.

Henderson, V., Mieszkowski, P., & Sauvageau, Y. (1976). *Peer group effects and educational production functions.* Ottawa, Canada: Economic Council of Canada.

Jencks, C. S., & Brown, M. (1975). Effects of high schools on their students. *Harvard Education Review, 45*, 273–324.

Katzman, M. (1971). *Political economy of urban schools.* Cambridge, MA: Harvard University Press.

Kenny, L. W. (1982). Economies of scale in schooling. *Economics of Education Review, 2*, 1–24.

Kiesling, H. J. (1967). Measuring a local school government: A study of school districts in New York state. *Review of Economics and Statistics, 49*, 356–367.

Levin, H. M. (1970). A new model for school effectiveness, In U.S. Office of Education, *Do teachers make a difference?* (pp. 55–78). Washington, DC: U.S. Government Printing Office.

Levin, H. M. (1976). Economic efficiency and educational production. In T. Joseph, J. T. Froomkin, D. Jamison, & R. Radner (Eds.), *Education as an industry* (pp. 149–190). Cambridge, MA: National Bureau of Economic Research.

Link, C. R., & Mulligan, J. G. (1986). The merits of a longer school day. *Economics of Education Review, 5*, 373–381.

Link, C. R., & Ratledge, E. C. (1979). Student perceptions, IQ, and achievement. *Journal of Human Resources, 14*, 98–111.

Maynard, R., & Crawford, D. (1976). School performance. In D. L. Bawden & W. S. Harrar (Eds.), *Rural income maintenance experiment: Final report* (Vol. 6, Pt. 2, pp. 1–104). Madison: University of Wisconsin, Institute for Research on Poverty.

Michelson, S. (1970). The association of teacher resources with children's characteristics. In U.S. Office of Education, *Do teachers make a difference?* (pp. 120–168). Washington, DC: U.S. Government Printing Office.

Michelson, S. (1972). For the plaintiffs—equal school resource allocation. *Journal of Human Resources, 7*, 283–306.

Murnane, R. J. (1975). *Impact of school resources on the learning of inner city children.* Cambridge, MA: Ballinger.

Murnane, R. J., & Phillips, B. (1981). What do effective teachers of inner-city children have in common? *Social Science Research, 10*, 83–100.

Perl, L. J. (1973). Family background, secondary school expenditure, and student ability. *Journal of Human Resources, 8*, 156–180.

Raymond, R. (1968). Determinants of the quality of primary and secondary public education in West Virginia. *Journal of Human Resources, 3*, 450–470.

Ribich, T. I., & Murphy, J. L. (1975). The economic returns to increased educational spending. *Journal of Human Resources, 10*, 56–77.

Sebold, F. D., & Dato, W. (1981), School funding and student achievement: An empirical analysis. *Public Finance Quarterly, 9*, 91–105.

Smith, M. (1972). Equality of educational opportunity: The basic findings reconsidered. In F. Mosteller, & D. P. Moynihan (Eds.), *On equality of educational opportunity* (pp. 230–342). New York: Random House.

Strauss, R. P., & Sawyer, E. A. (1986), Some new evidence on teacher and student competencies. *Economics of Education Review, 5*, 41–48.

Summers, A., & Wolfe, B. (1977), Do schools make a difference? *American Economic Review, 67*, 639–652.

Tuckman, H. P. (1971). High school inputs and their contributions to school performance. *Journal of Human Resources, 6*, 490–509.

Winkler, D. (1975). Educational achievement and school peer group composition. *Journal of Human Resources, 10*, 189–204.

References

Averch, H. A., Carroll, S. J., Donaldson, T. S., Kiesling, H. J., & Pintero, J. (1972). *How effective is schooling? A critical review and synthesis of research findings* (Report R-956-PCSF-RC). Santa Monica, CA: Rand.

Baker, K. (1991). Yes, throw money at schools. *Phi Delta Kappan, 72*, 628–631.

Begg, C. B., & Berlin, J. A. (1988). Publication bias: A problem in interpreting medical data. *Journal of the Royal Statistical Society, Ser A, 151*, 1–27.

Bryk, A. S., & Raudenbush, S. W. (1988). Toward more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education, 97*, 65–108.

Cohn, E., & Millman, S. D. (1975). *Input-output analysis in public education.* Cambridge, MA: Ballinger.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity.* Washington, DC: U.S. Government Printing Office.

Cooper, H. M. (1984). *The integrative research review.* Beverly Hills, CA: Sage.

Cooper, H. M., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis.* New York: Russell Sage Foundation.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

Glassman, N. S., & Biniaminov, I. (1981). Input-output analyses i schools. *Review of Educational Research, 51*, 509–539.

Hanushek, E. A. (1971). Teacher characteristics and gains in student achievement: Estimation using micro-data. *American Economic Review, 61*, 280–288.

Hanushek, E. A. (1972). *Education and race: An analysis of the educational production process.* Cambridge, MA: Heath-Lexington.

Hanushek, E. A. (1981). Throwing money at schools. *Journal of Policy Analysis and Management, 1*, 19–41.

Hanushek, E. A. (1986). The economics of schooling: Production anc efficiency in public schools. *Journal of Economic Literature, 24* 1141–1177.

Hanushek, E. A. (1989). The impact of differential expenditures on school performance. *Educational Researcher, 18*(4), 45–65.

Hanushek, E. A. (1991). When school finance "reform" may not b a good policy. *Harvard Journal on Legislation, 28*, 423–456.

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics, 9*, 61–85

Hedges, L. V. (1992). Modeling publication selection effects in m analysis. *Statistical Science, 7*, 246–255.

Hedges, L. V., & Olkin, I. (1980). Vote counting methods in research synthesis. *Psychological Bulletin, 88*, 359–369.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysi* New York: Academic Press.

Heim, J., & Perl, L. (1974). *The educational production function: Implications for educational manpower policy* (Monograph No. 4). Ithaca, NY: Cornell University, Institute of Public Employment.

Levin, H. M. (1970a). A cost effective analysis of teacher selectio *Journal of Human Resources, 5*, 24–33.

Levin, H. M. (1970b). A new model for school effectiveness, In U.S. Office of Education, *Do teachers make a difference?* (pp. 55–78). Was ington, DC: U.S. Government Printing Office.

Levin, H. M. (1976). Economic efficiency and educational productic In T. Joseph, J. T. Froomkin, D. Jamison, & R. Radner (Eds.), *Education as an industry* (pp. 149–190). Cambridge, MA: National Bureau of Economic Research.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of revie ing research.* Cambridge, MA: Harvard University Press.

Light, R. J., & Smith, P. V. (1971). Accumulating evidence: Procedures for resolving contradictions among different studies. *Harvard Edu tional Review, 41*, 429–471.

Murnane, R. J. (1981). Interpreting the evidence on school effectivene *Teachers College Record, 83*, 19–35.

Rosenthal, R. (1978). The "file drawer problem" and tolerance for n results. *Psychological Bulletin, 86*, 638–641.

Spencer, B. D., & Wiley, D. E. (1981). The sense and the nonsense school effectiveness. *Journal of Policy Analysis and Management, 1*, 43–52.

U.S. Center for Education Statistics. (1992). *Digest of education statistics, 1992.* Washington, DC: U.S. Government Printing Office.

Received May 5,
Revision received November 15,
Accepted December 2,

345