Money Might Matter Somewhere: A Response to Hedges, Laine, and Greenwald

ERIC A. HANUSHEK

Educational Researcher, Vol. 23, No. 4, pp. 5–8.

Over the past quarter century, there has been an attempt to use systematic statistical analysis to inform educational policy. This analysis has not been greeted enthusiastically by educators, particularly because it has clearly indicated that their current operations are inefficient and broadly ineffective (Hanushek, 1989). The recent analysis of Larry Hedges, Richard Laine, and Rob Greenwald (1994) addresses the overall conclusions in my article about the lack of relationship between resources and performance. In this response, I endeavor to make clear that, when the dust settles, the previous conclusions and perspective on educational policy remain intact and are, perhaps, even strengthened by their reanalysis.

Hedges, Laine, and Greenwald's simple summary is that "using more sophisticated synthesis methods . . . shows a systematic positive relation between resource inputs and school outcomes." They then suggest that this should change the direction of the educational policy debate. The two main points I wish to make are (a) "more sophisticated" is not synonymous with correct, and (b) their interpretation is potentially very misleading when it comes to policy matters.

Hedges, Laine, and Greenwald pose the fundamental issue as one of using the right statistical methods. Unfortunately, I believe that Hedges, Laine, and Greenwald commit the larger error of asking the wrong question. This problem tends to get lost in their statistical manipulations and their zeal to overturn prevailing conclusions about the effectiveness of pure resource policies in promoting student achievement. Although I address some of the important shortcomings of their statistical approach, that discussion must at times turn technical. It is important that the policy significance not be lost in the technical details, and I return to that at the end of this response. Most importantly, the policy interpretations do not depend really on the statistical issues.

Background Context

My summary of existing research reviewed all of the estimated input-output relationships for schools that could be found in the literature (Hanushek, 1989). These input-output relationships, also called educational production functions, simply relate various school resources to student performance. The statistical procedures, generally some form of regression analysis, are designed to separate the influences of school inputs from those of family and other community factors.

While much of the general discussion is framed in terms of expenditures, the majority of the studies actually investigate the effects of specific resources. A primary reason for this analytical approach is that increasing expenditure on schools occurs through changes in basic resources, or inputs, of schools. In general terms, added expenditure involves some combination of raising teacher salaries, increasing teacher/pupil ratios, adding to materials and supplies, or expanding administrative expenditure. Teacher salaries themselves are determined primarily by levels of teacher experience and teacher education. A majority of production function studies investigate a combination of these specific resources to schooling.

My summary provided a tabulation of the findings about how major items of resource use—teacher/pupil ratios, levels of teacher education, and the like—are related to student performance. That analysis concentrates on the simplest questions that could be asked: Is provision of added resources closely related to improved student performance, and are we confident of the answers we find given the statistical and analytical uncertainties in the underlying work? While the prior belief driving much of U.S. educational policy was that the resources identified should lead to improved student performance, the results of the summary of studies indicate that such a presumption is totally unwarranted.

Hedges, Laine, and Greenwald set out to show that my original statement, "there is no strong or systematic relationship between school expenditures and student performance" (Hanushek, 1989), is incorrect. A key element of their critique actually hangs on the interpretation of the words "strong" and "systematic." Strong and systematic are, of course, not technical statistical terms. They were meant by me to summarize a situation in which the vast majority of studies on the relationship between specific resources and student performance give no real confidence that there is any relationship (i.e., that the estimated relationships are statistically insignificant for the majority of studies). Moreover, in many instances, one finds the "wrong" sign (i.e., the estimate suggests higher resource usage is associated with lower student performance). Hedges, Laine, and Greenwald, on the other hand, implicitly use a very different definition of these summary terms.

ERIC A. HANUSHEK *is director, W. Allen Wallis Institute of Political Economy, Harkness Hall, University of Rochester, Rochester, NY 14627-0158. He specializes in the economics and finance of education.*

They interpret this statement as meaning that there is no evidence that the estimated relationship in *any* study has the expected sign and is statistically significant. They then go on to produce a series of statistical tests that provide evidence that at least one of the hundred or more estimated parameters for each of the major resource factors could be positive and statistically significant. This is simply an absurd interpretation of what is the important and inescapable conclusion of the data: There is no strong or systematic relationship between key school resources and student performance.

It would surprise no one to find that some of the educational production function estimates confidently point to effective use of resources. Such a finding would be entirely consistent with the standard interpretation of the results: Some districts use resources well while others make very poor use of resources. To the extent that Hedges, Laine, and Greenwald simply confirm this with their manipulations, nothing is new or different.

Technical Issues

It is important to understand the nature of the evidence from which they draw their conclusions. Hedges, Laine, and Greenwald's analysis is based on the same underlying studies found in Hanushek (1989), but their approach is very different.

Hedges, Laine, and Greenwald correctly note that "vote counting," the simple tabulation of sign and significance levels of estimated parameters from existing studies, is a very simple procedure that does not provide as much information as one would ideally like. Specifically, vote-count methods do not indicate magnitude of effects. Further, Hedges and Ingram Olkin (1985) show that in certain circumstances, ones bearing little relationship to the analyses of educational production functions, paradoxical results of a statistical variety can arise with this procedure. With this motivation, Hedges, Laine, and Greenwald adapt a specific form of existing meta-analysis methodology to circumstances other than that with which it was designed to deal.

They wish to combine statistical information from the various underlying studies to provide overall tests of basic hypotheses about whether or not resources affect performance. Unfortunately, the application of their statistical approach requires a series of key assumptions and analytical choices that each work to invalidate their technical analysis.

First, they eliminate consideration of all studies that report statistically insignificant relationships but that do not provide information on the estimated direction of effect (e.g., some studies will report that "class size had no significant effect on achievement" without reporting whether the insignificant estimate suggested a positive or negative effect of class size). This has the effect of completely ignoring 30% to 40% of the estimates for the effects of teacher education, teacher salary, teacher/pupil ratio, administrative inputs, and facilities. Remember that statistically insignificant means that there is not strong evidence that a relationship exists. Hedges, Laine, and Greenwald eliminate this large proportion of the evidence that supports "no effect" simply because it does not fit into the constraints of their methodology.

Second, the approach they employ for hypothesis testing is based on combining statistical information (p values) from a series of individual parameter estimates. But to apply their methodology, one must use underlying estimates that are statistically independent. When the estimates are not independent, poor estimates in one instance will tend to turn up again in other estimates. Yet the underlying estimates of resource effects are clearly not independent because various estimates are produced from the same data set, refer to the same schools, and relate to the same sample of students. Hedges, Laine, and Greenwald select specific but overlapping subsets of the underlying regression results including analyzing only one estimate from any published article and arbitrarily trimming large and small estimates in an effort to circumvent the underlying data problems. Nevertheless, by their own description, none of their subsetting of studies yields a sample of independent estimates appropriate for the statistical methodology, because each subsample retains estimates that are necessarily not independent of each other.

The creation of specific subsets leads to further reductions in the underlying studies that enter into Hedges, Laine, and Greenwald's analysis. For example, instead of the 152 estimates of the effects of teacher/pupil ratios in my article, they use between 41 and 92 estimates, depending on their sample selection criteria; for teacher's education, instead of 113 underlying estimates, they employ between 29 and 67. Finally, having created separate samples of results, Hedges, Laine, and Greenwald go even further. They implicitly want the reader to accept their own vote-counting methodology: If results in a majority of their specific subsets appear consistent with their conclusion that resources matter, the case is made. The subsamples and the resulting separate tests by Hedges are no more independent than the underlying data that they employ.

Third, they have to deal with the fact that there are statistically significant positive and statistically significant negative results for each of the factors. Conventionally, an estimate is labeled "statistically significant" if there is less than a 5% possibility of getting the estimate by chance when no true relationship exists. A motivating observation of Hedges, Laine, and Greenwald is that more than 5% of the underlying estimated parameters are statistically significant, that is, more are statistically significant than one would expect from simply a chance occurrence. Hedges, Laine, and Greenwald are out to demonstrate that resources have a positive effect on performance. Thus, they divide the estimates according to the sign of the estimated coefficients so as to conduct a series of one-tailed tests of positive effects and a separate series of one-tailed tests of negative effects of resources. This extraordinarily unusual, and technically suspect, approach is forced upon them because the raw data in the total sample of estimated coefficients simply do not generally support their basic proposition.

The fact that there are both too many negative and too many positive coefficients that are labeled statistically significant raises real problems of interpretation. This could happen if the parameters were not independently drawn from the same distribution, for example, if there was not a common underlying relationship between resources and achievement that was being estimated in the different studies. Said differently, this could arise if resources were effective in some circumstances but not in others. It could also happen if the underlying distributional assumptions for the statistical tests in the original studies are wrong, for example, if publications contain a bias toward presenting statistically significant results or if the required statistical assumptions are not met in the underlying studies. I suspect that both basic explanations are true, but Hedges, Laine, and Greenwald appear to assume that just parameter variation is relevant. How does this concern enter? The Fisher method used by Hedges, Laine, and Greenwald for combining of significance tests is based on the underlying distributional assumptions being correct, that is, that the underlying *p* values are accurate. This would not be the case if the distributional assumptions for the underlying studies do not hold, as the data suggest.

Each of these problems is serious enough to undermine Hedges, Laine, and Greenwald's preferred interpretation of the data. In simplest terms, the couching of their analysis in technical phrases like "combined significance tests," "robustness testing," and "median half-standardized regression coefficients" gives the misleading impression that sophisticated statistical methodology has led to conclusive results, where the previous analysis did not. Such a conclusion is clearly wrong.

The application of the methodology also deserves attention. Table 3 in Hedges, Laine, and Greenwald's article is the key summary table that they relied upon in reaching their conclusions. It represents outcomes of the hypothesis testing (reject or not) for the various resource measures and for each of their restricted samples. This table is clearly meant to provide a graphic indication of general rejection of the hypothesis that *all* coefficients are simultaneously zero or negative but less common rejection across their restricted subsamples of the positive variant of this hypothesis. This summary table is presumably based on the test statistics in Table 2, which are the underlying ingredients that go into their hypothesis testing. But even if one were to accept the general methodology, which I believe is inappropriate in terms of the underlying statistical assumptions, the data and the conclusions Hedges, Laine, and Greenwald draw from Table 3 are truly difficult to accept. Take, for example, the important case of teacher/pupil ratios. The existing literature provides 152 separate estimates of the effect of teacher/pupil ratios on student performance: 14 are positive and statistically different from zero; 13 are negative and statistically different from zero; 34 are positive and statistically insignificant; 46 are negative and statistically insignificant; and 45 are statistically insignificant but with no reported sign. On the face of it, it would seem difficult to make the case that these results support the conventional wisdom that smaller classes are systematically important to student performance. Indeed, by Hedges, Laine, and Greenwald's statistics in Table 2, no matter which of Hedges, Laine, and Greenwald's subset of these results is used, the hypothesis of no positive effects and the hypothesis of no negative effects are both rejected. Moreover, across their subsamples, the test statistics for the negative (unexpected) sign suggest that this specific hypothesis is the more strongly rejected. Table 2 does indicate in a footnote that the alternative variable definitions of teacher/pupil ratio and pupil/teacher ratio, which are rather arbitrarily used in the underlying input-output estimation, are put on a consistent basis of teacher/pupil ratios for the Hedges, Laine, and Greenwald analysis (as they were in my own analysis). Thus, it comes as some surprise to turn to Table 3, which purportedly is a simple summary of the hypothesis testing results from Table 2, and to find two separate results: Results for pupil/teacher ratio have somehow been separated from results for teacher/pupil ratio, even though these resource measures are conceptually indistinguishable. When this curious division is made, conventional wisdom appears to be strongly supported: The positive case shows almost uniform rejections, while the negative case is rejected only in one subsample. Hedges, Laine, and Greenwald provide no explanation for the inexplicable testing procedure, although with this division the results now conform to the story they wish to tell.

Finally, Hedges, Laine, and Greenwald devote most of their attention to estimates of the effects of per pupil expenditure. This in effect concentrates their attention on many of the weaker studies in the collection, because per pupil expenditure is generally calculated only at the school district level. Thus, they implicitly concentrate on studies of aggregate district performance, or they include studies that significantly mismeasure the resources going to individual students (because of within-district variations in expenditures). Quite clearly, the better evidence comes from studies of individual students and individual classroom teachers, studies that investigate the components of instructional expenditure such as teacher salary and teacher/pupil ratio. These studies typically rely on more accurate descriptions of the relevant resources for each student and avoid various aggregation problems. Moreover, a majority of the individual level analyses employ longitudinal performance data, an analytical characteristic whose importance is underscored by Hedges, Laine, and Greenwald, while this is not true for the studies of per pupil expenditure. Analyses of the better underlying studies into specific resource effects provide little support for any systematic relationship between resources and student performance.

If their implied predictor of a seven-standard-deviation improvement in student performance over the past quarter century had been realized, Albert Einstein would currently rate below the national average.

Policy Issues

The biggest concern clearly relates to policy uses and interpretations of these data. The interest in the entire topic derives directly from the policy perspective, held by some, that simply increasing funding to all schools, to low wealth schools, or to low spending schools would be sufficient to improve the performance of their students. On the other hand, if simply increasing expenditure does not ensure improved performance, more thoughtful and complex policies that deal with the organization and decision making in districts would be required. The clear and overwhelming conclusion from available evidence is that more complex policies are required.

The past work demonstrates that simply adding resources to districts will not ensure improvement in student performance. Even if some districts can employ resources effectively, as undoubtedly is the case, there is no assurance that overall increases in resources will lead to overall improvements. Indeed, the evidence, as it has been correctly interpreted by most interested in the policy issues, is consistent with some districts finding effective ways to use resources and others following very ineffective policies. Simply knowing that some districts might use added resources effectively does not provide any guide to effective policy, unless many more details can be supplied. Most importantly, it is necessary to have a description of the decision-making process or the details of efficacious uses of resources, but such information is not forthcoming from existing research. In other words, even if the statistical results of Hedges, Laine, and Greenwald could be trusted to indicate reliably that *some* systems have used resources effectively, this result by itself would not provide any real policy guidance. Recognizing that many districts do not use resources effectively, others extend the argument to say something such as, "We would of course advocate more spending only on effective programs." But this is clearly tautological, having no policy relevance unless the method of ensuring identification of effective uses can also be specified.

What about the estimated magnitude of resource effects? Hedges, Laine, and Greenwald combine the actual coefficient values across the disparate underlying studies to assess the effectiveness of pure resource policies. They estimate that a 10% increase in real resources will improve measured achievement by 0.7 standard deviations, a "large effect" in Hedges, Laine, and Greenwald's words. But, nationally over the past 25 years, real expenditure per pupil has risen more than *100%*, while all available evidence suggests that performance has at best remained constant but has more likely declined. A variety of explanations can be used to rationalize these actual data with the Hedges, Laine, and Greenwald prediction, for example, the costs of special education, the increased immigrant population, and the difficulties of measuring the full range of school outputs. None offers an explanation that would convincingly offset Hedges, Laine, and Greenwald's implicit prediction of a seven-standard deviation increase in student performance (Hanushek, Rivkin, & Jamison, 1992).

If their implied predictor of a seven-standard-deviation improvement in student performance over the past quarter century had been realized, Albert Einstein would currently rate below the national average. Moreover, the pattern of Hedges, Laine, and Greenwald's results makes them particularly unconvincing, because they suggest little beneficial effect of teacher salaries, teacher/pupil ratios, and other specific resources while suggesting important effects of aggregate expenditure. Their explanation is that sophisticated school decision makers find the best use of funds even though the specific factors do not count for much. There is, however, a serious problem with this explanation. The combination of teacher salary and teacher/pupil ratio describes instructional expenditure per pupil quite well. If variations in these underlying factors, and implicitly in instructional expenditure per pupil, are unimportant but variations in total expenditure per pupil are important, the only candidate for effective resource use is the factors making up the difference between total and instructional expenditure. For the most part, what is left out is simply administrative expenditure per pupil. Do they really want us to believe that the best way to improve schools today is to increase administrative expenditures?

My summary of the existing empirical analyses (Hanushek, 1989) was meant to provoke a discussion of educational policy issues that went beyond pursuit of the overly simplistic question, "Should we or should we not spend more money on schools?" Hedges, Laine, and Greenwald appear to argue that "throwing money at schools" is not the first-best policy, but it may be a second-best solution and even be a necessary factor. This conclusion is, I believe, misleading and potentially damaging. In my view, throwing money at schools is not a second- best approach but may be a 20th-best approach. More importantly, pursuing it initially may preclude pursuing some of the more productive approaches. For example, substantial evidence points to dramatic and consistent differences in performance among teachers; it is just that these differences are unrelated to teachers' salaries (see, e.g., Hanushek, 1992). If we made a blanket increase in spending, say the 10% increase Hedges, Laine, and Greeenwald use as an example, we are likely to see a substantial portion of this go into uniform salary increases for teachers. Such uniform increases would not increase the correlation between salaries and student performance, but they would slow down turnover of teachers, so that policies designed to attract better people into teaching would be thwarted. The specific policy of this example, of course, does not have to be the way money is spent. The available evidence simply indicates that the natural proclivities of school systems do not systematically lead to effective use of resources.

Improvement of our schools is, I believe, a very important national policy goal to pursue. It would be very unfortunate if policymakers were confused into believing that throwing money at schools is effective. More serious reform is required if we are to realize the full benefits of our schools.

Note

I am indebted to Bruce Chapman, Stanley Engerman, John Jackson, and Michael Wolkoff for helpful comments.

References

Hanushek, E. A. (1989). The impact of differential expenditures on school performance. *Educational Researcher, 18*(4), 45–51.

Hanushek, E. A. (1992). The trade-off between child quantity and quality. *Journal of Political Economy, 100*(1) 84–117.

Hanushek, E. A., Rivkin, S. G., & Jamison, D. T. (1992). Improving educational outcomes while controlling costs. *Carnegie-Rochester Conference Series on Public Policy, 37*, 205–238.

Hedges, L. V., Laine, R. D., & Greenwald, R. (1994). Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes. *Educational Researcher, 23*(3), 5–14.

Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis. San Diego: Academic Press.

Received January 26, 1994
Revision received February 22, 1994
Accepted March 1, 1994