Outcomes, Incentives, and Beliefs: Reflections on Analysis of the Economics of Schools
Author(s): Eric A. Hanushek
Reviewed work(s):
Source: *Educational Evaluation and Policy Analysis*, Vol. 19, No. 4 (Winter, 1997), pp. 301-308
Published by: American Educational Research Association
Stable URL: http://www.jstor.org/stable/1164444
Accessed: 14/01/2012 13:59

Educational Evaluation and Policy Analysis
Winter 1997, Vol. 19, No. 4, pp. 301–308

Outcomes, Incentives, and Beliefs: Reflections on Analysis of the Economics of Schools

Eric A. Hanushek
University of Rochester

Economic analysis of education and schooling has progressed considerably over the past few decades, and this essay attempts to put a few key issues into perspective. I look at the field from the particular vantage point of an economist with an interest in how school resources are used and how student performance can be improved. This perspective—at least as applied—may be a bit narrow, although I think it is central to much of the policy discussion in education. Overall, I believe that this research line is poised to improve educational policymaking, but much of its potential depends on developments yet to come.

As I reflect on the development of policy analysis in education, four aspects stand out. The first involves a transformation of approach that is now deeply embedded in the fabric of analysis and unlikely to go away. This transformation centers on the now-obligatory concern with observed performance and student outcomes. The second is the potential for experimentation to add to our policy-relevant knowledge. While educational research and evaluation have emphasized statistical analyses, this is prone to misleading results unless stringent conditions are met. Like medicine, education could benefit from more extensive use of random-assignment experiments. The third is a revolution yet to occur but one that I believe will likely stand as the centerpiece of the next major development. This potential line of development involves direct attention to the incentive structures in schools. Finally, I have one continuing concern that I do not believe will go away in the near future. This concern is the tendency for analysis to be intertwined with hopes, dreams, or normative views. This tendency lessens the scientific content of analysis and heightens concern that advice might lead us in the wrong direction.

The Completed Transformation

In my opinion, the modern era of evaluation and analysis of schools was ushered in 30 years ago by the "Coleman Report" (Coleman et al., 1966). The Coleman Report, a congressionally mandated study of public schools, remains the largest and probably still most influential study of education to date. The U.S. Office of Education was charged with documenting the degree of inequality of educational opportunity that existed in the country. Instead of merely cataloging the racial distribution of students in schools, the characteristics of facilities, and the characteristics of faculty—a natural approach of the day—the research team led by James S. Coleman collected parallel information about student performance. Their idea, which has been reinforced by subsequent analysis, was that we should care most about things that directly affect student performance.

Much of the discussion of the Coleman Report has revolved around the conclusions that are popularly attributed to that document. The interpretation at the time, which has persisted until today, is that the analysis showed that "schools do not make a difference." The analysis itself was subjected to intense scrutiny and severe criticism, and the correct interpretation of the analysis is now clearer.[1] The Coleman Report found, consistent with much of the subsequent research, that the measured resources of school are not closely related to student performance (see Hanushek, 1997). While more resources sometimes appear to be important in determining student achievement, most of the time they are not important, and sometimes they even appear to harm achievement. This finding of a lack of any general resource relationship is, however, very different from finding that schools have no differential impact. A number of subsequent studies document rather conclusively that schools have significantly different affects on student achievement, even if the good schools are necessarily not those rich in traditionally measured inputs (Hanushek, 1997).

Most people concentrate on the analytical approach and policy conclusions of the Coleman Report, but this research, more than any other piece of work, changed the way that people do or should think about the analysis of schools. The lasting impact of the study is that now serious policy analysis needs to make a direct link to outcomes of interest. It is no longer acceptable to ignore outcomes and to talk simply about input differences. Even if it was once thought that things like average class sizes or degree levels of teachers are good proxies for things we care about, the majority opinion now tends to hold that they are not central to student performance.[2] Therefore, serious policy analyses must be clearer about the outcomes of interest and about how any discussion relates to those outcomes.

The interest in actual performance instead of just inputs to schools seems so natural now that many people probably do not realize that this is a relatively new phenomenon. The importance of the Coleman Report in bringing about this revolution is moreover frequently missed.

The correct way to measure performance does remain the subject of research and controversy. A portion of the controversy centers on the use of standardized tests to judge student performance. The Coleman Report and a majority (75%) of the subsequent studies rely on some sort of standardized test score to measure student performance. The underlying notion is not that test scores per se are what are valued. Instead, the Coleman Report relies on an underlying two-stage model where (1) test scores measure qualitative differences in students at the time of schooling, and (2) test scores are directly related to the true outcomes of interest such as subsequent success in the labor market. Pursuing this approach permits direct investigation of school programs and resources without having to wait until a student has gone through an entire labor market career. Although the validity of this linkage between test scores and subsequent success has been questioned (Card & Krueger, 1996), there is substantial evidence that this is a reasonable characterization (Hanushek, 1996).[3]

A different perspective has also entered into the outcome measurement and testing debate. In the discussions of standards and testing, a variety of people have eschewed the use of standardized tests, arguing that they do not adequately measure higher-order skills and that they could well lead to pressures to lessen the content of instruction as people respond to test measurement of simple skills. Part of this discussion confuses the various purposes of testing. In order to design individual-specific educational programs, more sophisticated "portfolio" testing or open-ended measurement might be useful. Yet, these tests do not typically permit direct comparison of an individual's performance over time, let alone comparison across school and programs. Thus, they do not form the basis for evaluation of schools. Moreover, while open to improvement, the fact that standardized tests have been shown to correlate with subsequent success indicates that they do provide a basis for explicit comparisons of student performance.

These debates about existing and proposed outcome measurement, however, go far beyond the point here. Simply put, if we wish to understand student performance and what determines it, we have to look at student performance. We cannot base policy prescriptions only on assumptions about which inputs determine student performance because many plausible sounding arguments and existing assumptions have not held up well when confronted by data.

The Coleman Report began the revolution of requiring attention to actual student performance. This revolution could now be called complete—akin to the eradication of smallpox—if it were not for a few pockets of resistance. Probably the most virulent of these is the discussion of school finance, where little attention is given to student outcomes. Ignoring student outcomes is reasonable if school finance discussions are viewed entirely as discussions of taxpayer and fiscal equity, but unreasonable if thought of in terms of education policy. Outside of this area, policy debates quite consistently relate to student outcomes, even if some of the linkages with their determinants are not completely understood.

The Role of Experimentation

Much of the history of inquiry into the determinants of student performance and achievement has taken a straightforward input-output perspective that follows (and improves on) the Coleman Report. A variety of inputs including school resources, school process information, and family background are related to different measures of student performance. One motivation for this approach is to assess how well schools are using their resources. Another motivation, either explicit or implicit, is finding out "what works" with an underlying idea that this will then form the basis of subsequent policy development. We could, for example, regulate against bad things or legislate toward good things. Or we could simply provide better information to those in the field so that they can improve on their performance, assuming that a lack of information is the key problem in existing schools.

This motivation seems quite naive. It seems unrealistic to believe that good policy would follow very directly from implementing the outcomes of any given statistical analysis.

From a perspective of the relationship of research and policy, two underlying interpretative issues arise. Most research articles, after finding a set of things that is correlated with student performance, immediately go to a section on policy conclusions. The steps between the statistical analysis and the section on policy conclusions are seldom discussed. I am most concerned about two intervening steps that are needed. First, one must be convinced that the identified aspect of schools is causally related to student performance, as opposed to simply correlated. Second, one must believe that the positive relationship can be duplicated elsewhere, out of sample. I am currently very skeptical of both of these.

The causality issue seems overwhelmingly important. When specific measures of teachers or schools are regressed on student performance, some samples find one significant, while others find another significant. That is, the same set of resources and school inputs are not found to influence student performance in a consistent and predictable way across studies, but instead individual studies identify and stress a rather idiosyncratic collection of factors that appear by conventional statistical tests to relate reliably to student achievement.

One can find statistically significant effects of given resources under three distinct circumstances:

- There is a strong (causal) relationship between the input and student achievement that yields considerable confidence that there is a relationship between inputs and outcomes when subjected to statistical analysis.

- There are other factors that have a strong causal effect on student performance but are omitted from the analysis. If correlated with the included inputs, these omitted factors can make the included inputs appear to be significantly related to student performance even if there is no such relationship. (This case simply relates to the statistical tests within the estimation of misspecified models. The tests, along with the estimated coefficients, are biased by omitting the important factors.)

- The particular sample may lead by chance to an estimated relationship that appears to be statistically significant when, in fact, no relationship exists. This situation is exactly what the statistical tests are designed to deal with, and the magnitude of the Type I error is simply the probability of this event.

Only the first circumstance warrants jumping to a causal interpretation and the resultant policy conclusions. But the results across studies, and not just each study taken in isolation, are relevant. The general pattern of few systematic results identified across studies seems consistent with simply picking up sample-specific correlations that do not represent any true causal effects.

The issue of ability to successfully transport any positive program or resource effects to other districts can be thought of as a special case of the causality problem. It appears that many education programs that are found to be successful really have not fully identified the key factors leading to success. For example, some innovative programs appear to succeed because of special local conditions, a few unique people, or the initial enthusiasm of participants. These things are not transported to other venues when the program is taken on the road, especially if the program is mandated by central decisionmakers.

A portion of this confusion results simply from a lack of replication. Unlike studies in the laboratory sciences, individual statistical or evaluation studies in education are seldom replicated. Because most of the evidence that accumulates comes from isolated statistical studies, most of the attention is focused on the reported statistical tests. Yet, there is reason to suspect, given both publication biases and tendencies for specification searches in exploratory analyses, that these statistical tests will understate the true size of Type I errors. In turn, this permits a variety of estimates to be taken more seriously than they should be.

Part of it is also a reliance on purely statistical approaches instead of more frequently turning to random-assignment experiments. Statistical modeling requires both knowledge of the underlying educational process and considerable care in designing the collection of data from schools. If stringent conditions are met, the results can provide data about causal effects that can then be translated into explicit policies. When not met, there is considerable potential for biased and misleading results. An alternative approach, pursued vigorously in the medical sciences, is to rely more on random-assignment of subjects to different treatments. The advantage of this approach is that randomization can substitute for knowing and measuring all of the factors affecting outcomes. Thus, under many circumstances, significant effects from a random-assignment experiment are more likely to arise because of an underlying causal effect than because of misleading analytical results.

We have had a handful of significant educational experiments over the past quarter century. The performance contracting experiment of the Office of Economic Opportunity of the 1970s was designed to investigate whether or not private firms under performance incentive contracts could outperform the public schools in educating disadvantaged students (Gramlich & Koshel, 1975). The class-size experiment in Tennessee of the 1980s (Project STAR) concentrated on differences in achievement between students in small classes (13–17 pupils) and large classes (22–25 pupils) in kindergarten through grade three (Mosteller, 1995; Word et al., 1990).[4] And the voucher experiment in Milwaukee of the 1990s permitted some disadvantaged youth to attend a school of their choice from a set of authorized private schools (Peterson, Greene, & Noyes, 1996; Witte & Thorn, 1996).

These experiments do illustrate the difficulty of conducting conclusive experiments (part of the reason why medical experiments frequently involve a series of different trials). The results of each have been the subject of continuing controversy. The contracts employed in the performance contracting experiment were seriously flawed, limiting what could be learned from the experiment. The STAR program raised questions about when and how effects from small classrooms arise that cannot be answered within the framework of the initial experiment. And the Milwaukee experiment, lacking some key features of random pupil assignment, left doubt about the magnitude of any achievement effects.

But instead of commissioning additional experiments, the answer generally is either to reanalyze the old data or to institute a full-fledged program without further analysis (such as the 1996 reductions of class sizes in California). The expense of the experiments, frequently given as an explanation for underutilization, truly pales relative to the expense of a full-fledged program that fails. For example, the STAR experiment has frequently been identified as "expensive" at an annual cost of $3 million in 1987. Yet, statewide reductions in class size to 15 students per class would cost Tennessee some $200–$300 million, depending on the grade levels for the reductions (Hanushek, 1994, pp. 144–145).

Incentives: The Missing Line of Inquiry and Policy

Ultimately, however, I do not think that replication of the existing studies is likely to take us very far toward designing better policies. I also do not believe that experiments of the traditional kind will necessarily handle many of the largest questions. The underlying idea behind current policy approaches is that we will be able to find a simple set of policy instruments that we can legislate or regulate into existence centrally. I doubt that we will get to that point. This is the centralized command-and-control model of policymaking. While it may seem strange to think of a system with 50 separate states and close to 15,000 separate school districts as having centralized decisionmaking, decisionmaking is still largely conceived as the central authority—first the state, then the district—mandating *how* schools are organized and staffed and *how* instruction should proceed. The emphasis (and rewards) are not closely related to what is to be accomplished but instead come out as a tangle of procedural regulations and mandates governing detailed aspects of hiring, staffing, and teaching. The knowledge, innovative potential, and energies of the teachers, principals, and others on the front line of instruction are not directly tapped but instead are constrained.

An implication of this is that, even if a common set of resources is applied to all schools, outcomes will not be the same. Why? Because resources are employed in very different ways across teachers and schools. As such, what are ostensibly the same curricula, salary schedules, and operating procedures will be implemented to very different ends, depending on the skills, perceptions, and inclinations of an array of different local decisionmakers. Moreover, these decisionmakers seldom see much direct feedback from the consequences of their decisions. Particularly when we think in terms of student performance, little rides on success or failure. Yes, there is the intrinsic motivation of teachers, the satisfaction they receive from doing a good job, and the potential approval or disapproval of parents and principals. Yet, many other implicit and explicit pressures and incentives impinge on teachers' lives, leaving student performance frequently as something to be dealt with after managing the more immediate concerns.

Thinking about the incentive structure of schools and the absence of direct incentives related to student performance actually leads to a much more benign view of teachers and school personnel than that implicitly included in much current discussion. If we know what works, and if that is not being put into place, it implies that current school personnel are either completely unknowledgeable or are unaffected by the interests of students. The alternative view taken here is that the observed outcomes in schools represent more the fact that teachers and school personnel are simply reacting to the incentive structure that they currently face, an incentive structure that does not emphasize student performance.

Two common puzzles—the general ineffectiveness of reduced class size and the lack of general improvements from new technology—help to put the issue of incentives and performance into perspective. While the research into student performance suggests that smaller classes do not generally lead to improved student outcomes (Hanushek, 1997), this finding is difficult for many to understand or believe. After all, smaller classes permit a teacher to design more individualized instruction and to deal with each child's needs. But the finding is not that smaller classes never work. Indeed, I personally believe that there are some teachers, some groups of students, and some subject matters that lend themselves to improving performance through reduced class size. I also believe that there are other classroom circumstances where smaller classes have no affect. If the system were geared to maximizing student performance, we might expect an effort to search out the favorable circumstances while "paying" for them with larger classes where the effects on outcomes would be small or nonexistent. This is not what happens, however, because the overriding objective is not maximizing student outcomes (for a given budget). Instead, with few clear incentives to improve student achievement, discussions of class size become more ones of working conditions, and school objectives become framed in terms of fairness, which is viewed as dictating that all classes be uniformly reduced. Thus, reductions in class size are seldom done differentially with an objective of boosting achievement, and the results indicate that any such reductions seldom generally boost achievement. The outcome of changes in class size might, however, be very different if schools faced a different set of incentives and were more selective in when and how they reduced class sizes.

Unfulfilled projections of the potential for improved performance from use of new technologies offer the second example. At one time, introduction of televised instruction was touted as an upcoming revolution in the school that would lead to clear increases in student learning. That vision has now been replaced by ideas of how computerized instruction or the Internet will revolutionize the classroom. With the dramatic decline in price of computers over the past decade and with the penetration of school and home computers that has already occurred, one might expect to see the results now in changes in the classroom and in student performance. But it has yet to be seen. Many current teachers are less familiar with computers than their students. Learning how to use a computer can be frustrating and time-consuming in general. Learning how to use one effectively in instruction is more difficult. But there is little help and even less incentive for a teacher to develop a computerized component of instruction because improved student performance is not generally rewarded and the use of computers might even operate to lessen the demand for teachers. Even though much of policy about technology in the classroom is discussed as if the primary shortage is one of hardware, unopened and unused computers in schools around the country suggest that the incentive structure (and related training) of teachers could be much more important. If computers are truly advantageous in learning and if teachers had direct incentives to improve student achievement, we might see more productive use of computer technology than we do now or can expect in the near future.

Research and evaluation have not done a good job at defining incentives and understanding the ramifications of different incentive structures. Many of the most intense policy battles revolve around alternative incentive schemes—merit pay, choice, private contracting, and the like. Yet from a research and evaluation standpoint, we have not made much progress in even developing a language to describe incentive structures, let alone understanding how to model them. We do not currently know how to specify the incentive effects of different contracts or rules, unless perhaps we can translate different components into monetary terms. Concerns about the mixture of intrinsic and extrinsic incentives or group aspects of incentive arrangements, however, immediately show the limitations of thinking only in monetary terms. Moreover, even when placed in simple monetary terms, we do not fully understand the richness of different contracting arrangements—witness the performance contracting experiment (Gramlich & Koshel, 1975).

Given the complexity of incentive arrangements and our current rudimentary knowledge, the consideration of incentive structures appears to be an ideal candidate for an aggressive series of experiments. Nonetheless, we currently have insufficient experience in the design of experiments themselves. Specifically, random-assignment experimentation appears better suited to some incentive schemes than to others. Random-assignment experimentation works best when there is a relatively well-defined treatment that can be effectively applied to one group and not another. Thus, something like observing the effects of reduced class size is ideal. On the other hand, alterations in the tenure rules of a state are difficult to experiment with because the incentive structure would presumably apply to an entire state, making the development of a control group difficult. Similarly, if one thought the primary effects of a merit pay scheme work through its effect on the supply of teachers to a district (and not on current teachers working harder), the design of an experiment with adequate controls is difficult. The recent expansion of charter schools provides another example of where cleverness will be required to obtain useful evaluations. Many such projects have the character of "demonstration" programs, where there is no obvious control group. While use of random selection rules, if there is oversubscription for charter schools, is sometimes possible, the evaluation potential depends crucially on individual circumstances. These arguments suggest that investigations of incentive schemes are likely to involve a combination of some random-assignment experimentation with some classical statistical research.

I personally am persuaded that the ultimate policy approach that will prove most successful in improving student achievement involves much more pervasive use of performance incentives (Hanushek, 1994). It is insufficient, however, just to say "use performance incentives." Generic performance incentives almost certainly have very different effects depending on the details of their structure, and we need to build up knowledge about what elements are most important and what the effects of alternative approaches are likely to be.

I also think it likely that the reactions of school personnel and students to incentives in education represent a more fundamental component of educational production relationships than the simple input-output structures that have received so much attention. For the reasons sketched above, I am skeptical that we will ever be able to describe in any detail what the full production relationship looks like. (A similar argument is developed in Murnane & Nelson, 1984). I am more optimistic, however, that we might eventually be able to describe how individuals will react to various incentive structures and how these reactions will translate into student performance. Nonetheless, we are very far from that today.

Progress on understanding incentives in schools is where research and evaluation are likely, in my opinion, to have their largest impact on policy and performance of schools. This analysis is at the top of what I believe to be the unfinished agenda.

The Power of Individual Beliefs: A Continuing Concern

Analysis and evaluation in education differs from a variety of other areas of scientific evaluation. Conclusions from school research often pop up in the development of policies toward education. And these policies frequently have direct feedback loops to participants in the research efforts—either school personnel, the researchers, or both. One effect of this feedback relationship is that positive and normative statements—what "is" and what "should be"—can be blurred and confused.

Education receives the attention it does because people believe (rightfully) that it has a powerful impact on individual incomes and well-being. An important component of public and governmental support for schooling derives from its role in promoting economic opportunities for the next generation. Thus, for example, the distribution of quality schooling across the population intersects directly with notions of social justice.

The situation is even more complicated than the issues of general preferences and values. Many of those who provide policy-relevant opinions and evaluations are far from indifferent to the results. Policies that increase the demand for workers in the education industry (who are trained by people in other parts of the education industry) have direct impact on workers in the education industry, whether or not they have impact on student performance.

It is not necessary to belabor this point. Educational policy analysis will face some continuing problems whenever people doing the analysis feel they have a direct stake—intellectual, philosophical, or financial—in the answer that comes out of the analysis. It does not, for example, take much thought about the nature of school finance policy within states to see the potential for conflict with scientific investigation. There is an element of this conflict in a wide range of scientific research endeavors—if for no other reason than people develop intellectual property rights for certain positions. But the potential problems appear much more important in education where the workers in the industry are the prime source of research and evaluation. This fact, coupled with the limited incentives in schools for improved student performance discussed previously, will require continuing attention.

Notes

[1]For example, immediately after the Coleman Report's publication, a massive, year-long faculty seminar at Harvard University undertook the task of understanding the conclusions and of re-doing a variety of the analyses (Mosteller & Moynihan, 1972). The primary criticisms involved the statistical methodology and the quality of the data for addressing some of the key questions; see, for example, Bowles and Levin (1968) or Hanushek and Kain (1972).

[2]This statement is based on the observed outcomes, given the current organization and incentives found in the typical school. For a full discussion, along with the caveats and controversies, see Hanushek (1997).

[3]Moreover, the 25% of the studies looking at nontest outcomes—including college attendance, labor market earnings, and school dropouts—yield essentially the same results regarding the inconsistency of any positive effects of resources on outcomes.

[4]Interestingly, the STAR experiment was directly related to the provocative early use of meta-analysis by Gene Glass and Mary Lee Smith in the first issue of *Education Evaluation and Policy Analysis*. That article, which combined the results of a variety of class-size experiments, suggested that reduced class size would not affect achievement much until class sizes get noticeably below 20 (Glass & Smith, 1979).

References

Bowles, S., & Levin, H. M. (1968). The determinants of scholastic achievement—an appraisal of some recent evidence. *Journal of Human Resources, 3*(1), 3–24.

Card, D., & Krueger, A. B. (1996). Labor market effects of school quality: Theory and evidence. In G. Burtless (Ed.), *Does money matter? The effect of school resources on student achievement and adult success* (pp. 97–140). Washington, DC: Brookings.

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.

Glass, G. V, & Smith, M. L. (1979). Meta-analysis of research on class size and achievement. *Educational Evaluation and Policy Analysis, 1*(1), 2–16.

Gramlich, E. M., & Koshel, P. P. (1975). *Educational performance contracting.* Washington, DC: The Brookings Institution.

Hanushek, E. A. (with others). (1994). *Making schools work: Improving performance and controlling costs.* Washington, DC: Brookings Institution.

Hanushek, E. A. (1996). Measuring investment in education. *Journal of Economic Perspectives, 10*(4), 9–30.

Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis, 19*(2), 141–164.

Hanushek, E. A., & Kain, J. F. (1972). On the value of "equality of educational opportunity" as a guide to public policy. In F. Mosteller & D. P. Moynihan (Eds.), *On equality of educational opportunity* (pp. 116–145). New York: Random House.

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The Future of Children, 5*(2), 113–127.

Mosteller, F., & Moynihan, D. P. (Eds.). (1972). *On equality of educational opportunity.* New York: Random House.

Murnane, R. J., & Nelson, R. R. (1984). Production and innovation when techniques are tacit. *Journal of Economic Behavior and Organization, 5,* 353–373.

Peterson, P. E., Greene, J. P., & Noyes, C. (1996). School choice in Milwaukee. *Public Interest, 125,* 38–56.

Witte, J. F., & Thorn, C. A. (1996). Who chooses? Voucher and interdistrict choice programs in Milwaukee. *American Journal of Education, 104*(3), 186–217.

Word, E., Johnston, J., Bain, H. P., Fulton, B. D., Zaharies, J. B., Lintz, M. N., Achilles, C. M., Folger, J., & Breda, C. (1990). *Student/teacher achievement ratio (STAR), Tennessee's K–3 class size study: Final summary re-*

port, 1985–1990. Nashville, TN: Tennessee State Department of Education.

Author

ERIC A. HANUSHEK is a professor of economics and public policy at the University of Rochester, Rochester, NY 14627-0158. His specialties are education policy and finance.

Manuscript received December 13, 1996
Revision received July 14, 1997
Accepted July 25, 1997