Scottish Journal of Political Economy, Vol. 51, No. 2, May 2004

WHAT IF THERE ARE NO 'BEST PRACTICES'?

*Eric A. Hanushek**

Abstract

Developing persuasive and consistent models of educational performance has proved elusive. Existing modelling suggests no clear relationship between resources and student performance. This mirrors observed policy outcomes. A possible explanation is that the achievement process is a complicated interactive one such that simple linear additive formulations break down. This analysis presents a stylized model of achievement where unmeasured teacher quality interacts with both resources and specific educational programs. Standard econometric analyses then replicate the aggregate findings in the literature. A policy implication is that finding 'best practices' may fail without recognition of the fundamental interactions.

I Introduction

Educational policy has simultaneously gone in quite different directions. The majority of educational policy concentrates on prescriptive and generally regulatory approaches. The central administrative authority declares a specific approach or sets a minimal set of requirements. Yet, regardless of the appeal of these alternative approaches, they do not show universal efficacy. Quite the opposite. Programs that appear efficacious in one setting do not generalize to other settings. The underlying logic of these approaches, however, is an assumption of universal and additive effects. This basic modeling idea may simply be incorrect. This paper explores some aspects of modeling, evidence, and policy related to educational programs.

II Motivation

Considerable debate has surrounded the development of educational policy and especially the role of resources and programs. The starting point of this debate has been a simple empirical finding that resources for schools do not appear closely related to student performance.

*Stanford University, National Bureau of Economic Research and University of Texas at Dallas.

Table 1
Public school resources in the United States, 1960–2000

	1960	1980	2000
Pupil–teacher ratio	25.8	18.7	16.0
% teachers with master's degree or more	23.5	49.6	56.2[a]
Median years teacher experience	11	12	15[a]
Real expenditure/ADA	$2,235	$5,124	$7,591

Notes:
[a]Data pertain to 1995. The statistical data of the National Education Association on characteristics of teachers was discontinued.
Source: US Department of Education (2002).

The conundrum of school resources is simplest to see in aggregate data. For the United States, data on school resources and student performance have been available for a substantial period of time. Table 1 describes the simple history of real resources for schools and spending between 1960 and 2000. What is clear is that there have been dramatic increases in the resources for schools. Pupil-teacher ratios have fallen by a third; the proportion of teachers with a master's degree has more than doubled; average teacher experience has grown to new highs; and real spending per pupil has more than tripled over the 1960–2000 period.

These resource patterns, which generally match the kinds of policy prescriptions often discussed, would be expected to lead to improvements in student performance. But, Figure 1 provides evidence that such is not the case. In this figure, performance of 17-year-old students on the National Assessment of Educational Progress (NAEP) is plotted from the tests inception to 1999.[1] For these representative students, mathematics and reading performance is slightly higher by the end of the period than the beginning while science performance is significantly down.[2] The rough summary of the figure is that student performance has been flat for three decades – a time when resources for schools increased consistently and dramatically.

Of course such aggregate data are seldom convincing. Over time a variety of other factors could change to distort the relationship with spending. In fact it is frequently asserted that 'kids have gotten worse' as seen by negative changes in family circumstances. The data do suggest that some aspects of family have changed in ways that would generally be thought to be detrimental to student learning: an increase in family poverty (until the early 1990s), a larger proportion of children from single parent families, and more children from non-English speaking families. On the other hand, over the same period average parental education increased, family size went down, and a larger proportion of

[1] NAEP is a governmentally sponsored test that is given to a random sample of students in each year. This test is generally regarded as the 'gold standard' for measuring student performance, relying on both careful test construction and solid administration to ensure a representative sample of students.

[2] While not shown, writing performance was also assessed between 1984–1996, but was suspended. Student performance over that period significantly declined.

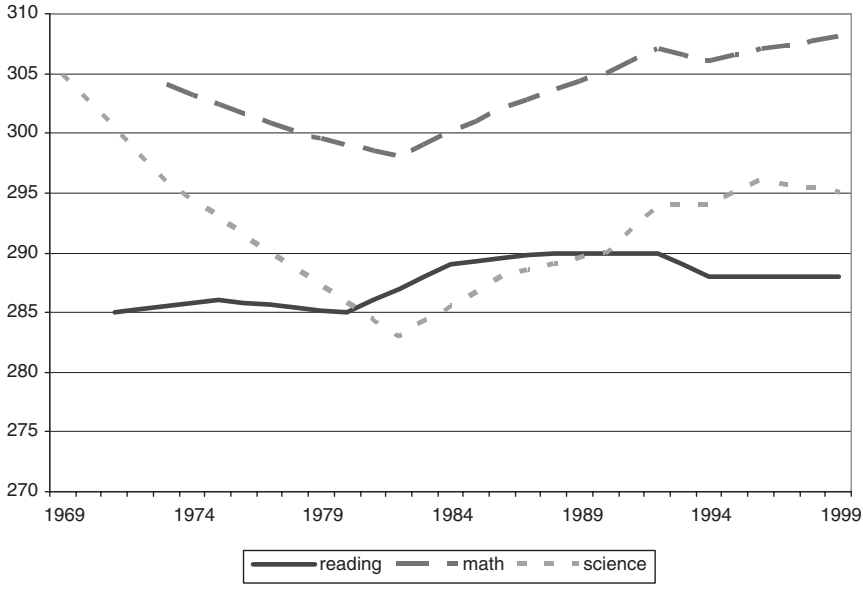

Figure 1. Performance of US 17-year-olds on the National Assessment of Educational Progress (NAEP), 1969–1999.

children attended formal pre-school programs – factors that would be generally regarded as favourable. Exactly how these varying influences net out is difficult to ascertain, but the available analysis does not suggest a large switch in family backgrounds one way or the other.

Similar international evidence is more difficult to obtain, because time series data on student performance outside the United States is very limited. Nonetheless, it is possible to provide some simple comparative results. A series of international tests in mathematics and science provide information about relative performance of students. These tests, with voluntary national participation, yield a simple index of outcomes.

Figure 2 provides a picture of comparative primary school spending for OECD countries where the countries are ordered on the basis of average student performance on the Third International Mathematics and Science Test (TIMSS).[3] In this simple illustration, it is clear that spending is not very systematically related to variations in student performance.

Again, a wide range of nonschool factors could influence the international pattern of achievement. Nonetheless, the simple story remains one of inconsistency between spending and achievement.

Although the aggregate evidence makes a clear *prima facie* case about resources, better evidence is found in econometric studies of student

[3] Summary data may be found in Organisation for Economic Co-operation and Development (2001). The spending data are provided in US dollars per pupil, based on purchasing power parity comparisons. Use of secondary school spending instead of primary school spending does not alter the qualitative picture.

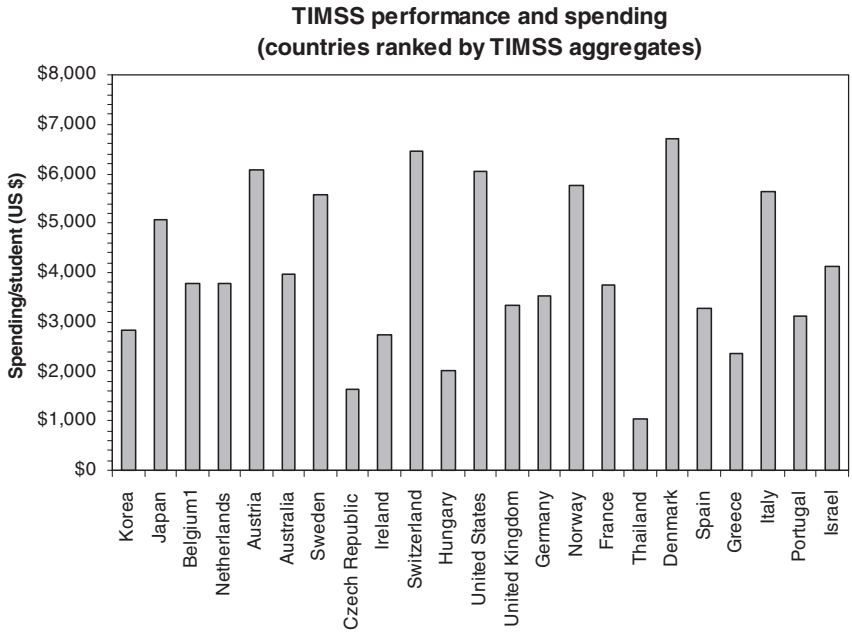

Figure 2. International spending on secondary school students with countries ranked in order of TIMMS performance.

performance. Investigations of the determinants of student performance have been carried out over a substantial period of time, and this supports the picture of the aggregate data.[4] The existing literature from the United States provides over 400 estimates of key school resource parameters. (While evidence for other countries of the world is beginning to develop rapidly, it is not currently possible to summarize the range of results in any convenient manner).[5]

This analysis is reviewed and analyzed elsewhere (Hanushek, 2002, 2003). The results are easily summarized. Key resource measures including pupil-teacher ratios, teacher education, teacher experience, and expenditures per pupil show little consistent relationship with student performance. For example, a vast majority of estimates of the effects of pupil-teacher ratios on achievement are statistically insignificant, and the point estimates (both considering and not considering insignificant estimates) are evenly split between positive and negative effects. Moreover, when just high quality studies are included, the conclusions are unaffected.[6]

[4] While not the first such study, the Coleman Report (Coleman *et al.*, 1966) is usually identified as the beginning of analyses of 'educational production functions.' This major governmental study motivated a range of follow-on studies – in large part because it (incorrectly) suggested that schools have little influence over student performance.

[5] Hanushek (2003) provides an overview of the international evidence.

[6] Hanushek (2002) discusses the evaluation of study quality based on potential problems of missing historical data and of aggregation biases. The highest quality studies consider value-added models for single states, where the policy environment is constant.

The previous evidence on school resources reported that teacher education and teacher experience showed little relationship with achievement. The investigations of teacher differences based on measured characteristics have gone farther to include teacher salaries, level of formal credentials, and teacher scores on tests. With some variation, these analyses have also found little systematic relationship with student performance (Hanushek and Rivkin, 2004).[7]

This evidence has, however, been frequently misinterpreted. Some have concluded that the analysis demonstrates that schools do not matter. This interpretation is quite inappropriate.

The clearest evidence on the importance of schools comes from direct investigations of teacher quality. An alternative approach to the examination of teacher quality reveals something very different from the standard estimation based on measured characteristics. This line of research concentrates on pure outcome-based measures of teacher effectiveness. The general idea is to investigate 'total teacher effects' by looking at differences in growth rates of student achievement across teachers. A good teacher would be one who consistently obtained high learning growth from students, while a poor teacher would be one who consistently produced low learning growth. The estimates come from fixed effect estimation for individual teacher differences.

This fixed effect approach is appealing for several reasons. First, it does not require the choice of specific teacher characteristics, which is often constrained in empirical work by data limitations. Second, and an important part of the development below, it does not require knowledge of how different characteristics might interact in producing achievement. (Most prior work on specific characteristics assumes that the different observed characteristics enter linearly and additively in determining classroom effectiveness).

A variety of studies have pursued this general approach over the past three decades; (Hanushek, 1971, 1992; Armor *et al.*, 1976; Murnane, 1975; Murnane and Phillips, 1981; Rivkin, Hanushek and Kain, 2001).[8] Each finds substantial variation in achievement growth across classrooms.

Careful consideration of such work nonetheless reveals some difficulties that must be overcome in order to estimate the variation of overall teacher effects. For example, teacher effects, school effects and classroom peer effects are generally not separately identified if the estimates come from a single cross section of teachers. Hanushek (1992), however, demonstrates the consistency of individual teacher effects across grades and school years, thus indicating that the estimated differences relate directly to teacher quality and not the specific mix of students and the interaction of teacher and students. Rivkin, Hanushek and Kain (2001) go even further and remove separate school and grade fixed effects and observe the consistency of teacher effects across different cohorts – thus isolating the impact of teachers.

[7] The investigations of teacher test scores have come closest to finding systematic effects, but it still remains rather weakly consistent across studies. See the summary in Hanushek (2003).

[8] A similar study for developing countries (specifically Brazil) finds very consistent findings; Harbison and Hanushek (1992).

The magnitude of estimated differences in teacher quality is impressive. Hanushek (1992) shows that teachers near the top of the quality distribution can get an entire year's worth of additional learning out of their students compared to those near the bottom.[9] That is, a good teacher will get an average student gain of 1.5 grade level equivalents on standardized tests while a bad teacher will get 0.5 grade level equivalents over a single academic year.

A second set of estimates comes from recent work on students in Texas (Rivkin, Hanushek and Kain, 2001). The analysis follows several entire cohorts of students and permits multiple observations of different classes with a given teacher. This work looks at just the variations in performance from differences in teacher quality *within* a typical school. (This analytical strategy is pursued because of the difficulties involved in separating differences in teacher quality from other factors that differ among schools.) The variation in teacher quality is large: Moving from an average teacher to one at the 85th percentile of teacher quality (i.e., moving up one standard deviation in teacher quality) implies that the teacher's students would move up more than 4 percentile rankings in the given year.[10]

The summary of the basic evidence is straightforward. Significant differences in the performance of schools exist. These differences, however, are not easily captured by simple measures of resources. Simple measures of teacher characteristics or of other school resources are not systematically related to these more general estimates of teacher performance, even though important qualitative differences do exist across teachers.

Intuitively, however, the evidence is perhaps ever sharper in terms of analyses of simple spending measures. A surprisingly large proportion of these studies even suggests that added funds actually lower achievement.[11] It seems hard to imagine that adding additional resources actually depresses student performance very frequently.

III An Alternative Interpretation

One of the obvious interpretations of this econometric evidence is that the underlying models are misspecified – and thus that the estimated relationships are biased. Such an explanation is clearly always possible, but in this case there is added justification for it. Most importantly, the finding that teacher quality differences *as identified by the output measures of quality* has such a strong

[9] These estimates consider value-added models with family and parental models. The sample includes only low income minority students, whose average achievement in primary school is below the national average. The comparisons given compare teachers at the 5th percentile with those at the 95th percentile.

[10] For a variety of reasons, these are lower bounds estimates of variations in teacher quality. Any variations in quality across schools would add to this. Moreover, the estimates rely on a series of conservative assumptions that all tend to lead to understatement of the systematic teacher differences.

[11] Part of the problem with the evidence from the pure expenditure studies, as compared to the investigations of real resources, is that these tend to be the lowest quality studies. They are generally highly aggregated and suffer from clear specification problems (Hanushek, 2003).

impact while the measured resources do not is the leading indicator that specification issues may be important.

Consider a simple model of student achievement:

$$O_i = F_i T_i E_i \tag{1}$$

where O is student achievement, F is family input, T is teacher input, and E is school expenditure.

The key notion in this model is that the effectiveness of any expenditure depends on the quality of teacher and the family inputs. Such a relationship would in fact recognize the dual role played by teachers. They lead the instruction in a classroom. But they also make important decisions about the educational process itself. These decisions have a distinct managerial element to them, leading to implications for the effectiveness of any expenditure.

Most analyses of schools concentrate on the role of the principal or headmaster when considering managerial aspects of schools. While having obvious merit, it does not negate the fact that teachers generally have considerable autonomy and can make a number of crucial decisions about the translation of resources into student performance.

The form chosen in equation (1) is clearly quite rigid, but it is sufficient to illustrate the general point about interaction of resources. Such interaction will, as described below, be particularly important when the quality of teachers (or management more generally) is not accurately observed.

In the simplest example, which looks close to many studies of student performance, consider taking observations of (O, F, E) and performing a simple linear regression relating them. Teacher quality, T, is treated as an omitted variable. Omitting teacher quality, as is well known, would not cause serious problems if it is uncorrelated with the other factors. In such cases, standard statistical models for the estimation of resource and expenditure effects would not be biased.

Is it reasonable to believe that teacher quality is uncorrelated with expenditure? At least for the United States, most evidence, including that described previously, supports a presumption of a small correlation. First, the key parameters determining teacher salaries – amount of teacher experience and teacher graduate education – are not closely related to student performance. Teacher graduate education is totally unrelated to student outcomes, while experience is only loosely connected. Second, direct investigations of teacher salaries show limited relationship to student performance. And the best of these studies provide virtually no evidence of salary effects, although the investigation is generally limited to movements along the supply function and not shifts in it. (Hanushek, 2003; Hanushek and Rivkin, 2004).

But it is difficult to go all the way to concluding that there is no correlation of expenditure with quality. Standard statistical theory suggests that the amount of bias is a direct function of the correlation, so it is possible that small correlations do not affect the results too much. The concern, however, is that in the multivariate situations that are typical it is difficult to characterize any bias

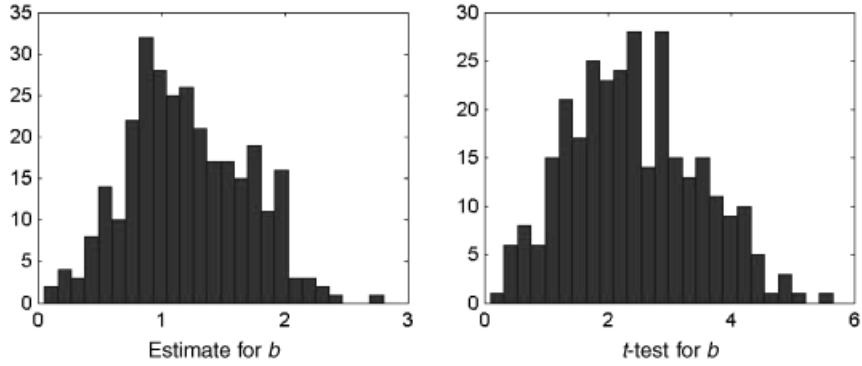

Figure 3. Estimates of expenditure effects in specification that ignores teacher quality:

$$\log O = \text{int} + a^* \log F + b^* \log E.$$

($r_{FE} = 0.85$; $r_{ET} = 0.1$; $n = 100$; 300 replications.)

a priori. The bias depends not just on the correlation of E and T, but also on the correlations with F.

In order to put this into perspective, consider generating achievement data in accordance to equation (1). In this we generate 100 observations by drawing random values for E, F and T, and then calculating O according to equation (1). Expenditure and family inputs (E and F) are drawn to be highly correlated (0.85), but expenditure and teacher quality (T) are correlated just 0.1. We then estimate a misspecified model that leaves out teacher quality. That is, we estimate:

$$\log O_i = \text{int} + a^* \log F_i + b^* \log E_i + e_i. \tag{2}$$

The issue we concentrate on is how the estimates of b correspond to the true value of 1.0 and what the tests of significance indicate. Figure 3 summarizes the estimates across 300 replications. First, the estimates of the expenditure impacts are noticeably biased above one. Second, the standard estimates of t-statistics indicate that over a third (113 estimates out of 300) of expenditure effects are not statistically significant ($t < 2$).

IV Best Practice

An alternative way of viewing the achievement process corresponds more directly to much of the policy discussion. Educational policies are frequently made by central authorities or, at least, administrators outside of each school such as local educational authorities. For these decisions, it is commonplace to decide on the introduction of specific programs. These might be pre-packaged ways of providing a course of instruction; they might be combinations of approaches – e.g., phonics instruction for reading – along with textbooks and workbooks; or they might be 'whole school' reforms that detail curriculum, books, time allocations, reward structures, testing programs, and the like.

These programs are essentially the detailed processes that enter into the production process for achievement. Most economic discussions of production

functions would not concentrate on these. Instead, the general notion is that the underlying process may change *mutatis mutandis*, depending on the specific input configuration. The process in the conceptual language of the production function is simply what is needed to produce the maximum attainable output from a given set of inputs. Thus, consideration of production functions in an abstract sense would simply presume that the correct program was applied and would not attempt to describe it.[12]

The nature of educational decision making, however, changes the emphasis from what might be implied by simple economic analysis. The frequent model that is applied assumes that programs can be characterized by a linear addition to the production process. Thus, a common specification involves just adding an indicator for program to the standard production function model that includes family and school inputs. (Note that this is a somewhat ambiguous formulation, because the lack of the given program is not 'no program' but generally the alternative, or standard, program. This alternative may be very different across sampled schools.)

In the spirit of the previous representation, think of a program having a multiplicative impact on outcomes as in:

$$O_i = F_i T_i P_i E_i. \tag{3}$$

For illustrative purposes we define two alternative programs. Program R reinforces teacher quality, so it would be one that, say, encouraged innovation where the quality of innovation is generally better for better teachers. The alternative, Program C, is compensatory in nature and helps to improve teachers at the bottom of the teaching distribution.

We again generate F, E and T as before with randomly set values between 0 and 8 and correlations of 0.85 and 0.1 for r_{FE} and r_{ET} respectively. The programs, R and C, are randomly assigned with a given probability. R and C take on values between 0 and 1. R is positively correlated with T, while C, is negatively correlated with T. Note that both programs have a positive impact on performance, but that they operate differentially across teachers. Finally, we estimate a model that again is misspecified by leaving out teacher quality, and we concentrate on the estimated program effects. (We again generate 300 replications of samples with 100 observations.) For this, we add an indicator variable that takes on the values $P = 1$ if Program R and $P = 0$ if Program C.

Figure 4 shows the estimates of the impact of program on achievement. Two things are clear. First, the estimated effects are both negative and positive. Second, conventional statistical tests imply that a majority of estimates are not statistically significant (199 out of 300).

But there are particular samples that yield significant positive effects (and also samples with significant negative effects). If policy making relies on a single study – which unfortunately appears to happen frequently – the potential for mistakes is large. Again, individual studies can easily in this illustration give very misleading estimates.

[12] A variety of analyses do in fact try to describe more about the process. See for example Angrist and Lavy (2001).

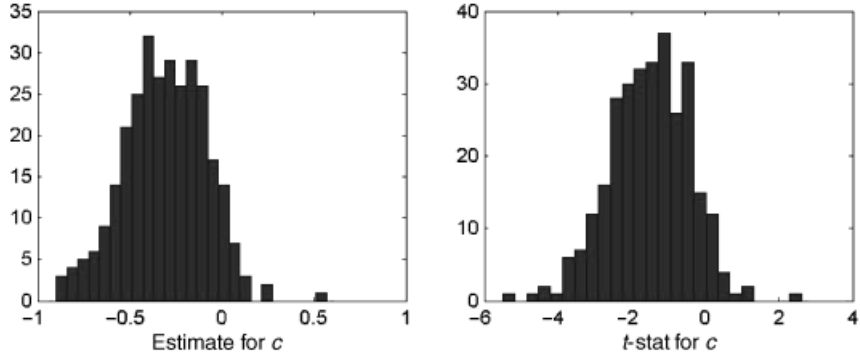

Figure 4. Estimates of program effects for 'quality reinforcing' program:

$$\log O = \text{int} + a^* \log F + b^* \log E + c^* P$$

(Prob[R] = 0.5; $r_{TR} = 0.4$; $r_{TC} = -0.4$; $n = 100$; 300 replications.)

V Implications for Econometrics/Analysis

In some respects there is an almost trivial response to these notions. Introductory econometrics courses routinely consider the implications of model misspecification, including that arising from employing an incorrect functional form and that arising from omitted variables. Thus, these cases are simple examples of the classic textbook situation.

There are, however, several problems with this viewpoint. First, in a practical sense very little experimentation with functional form is done (or reported). Distinguishing among alternative functional forms is often difficult, and leads to relatively minor experimentation – particularly when there are not strong reasons to choose a particular form.

Second, in terms of the model, analyses are often driven by simple data availability. With relatively rare exception, the data used for analysis come from administrative data. Such data, collected to facilitate the normal operations of schools, are generally not designed to support an underlying research program. While the dataset collected for the Coleman Report (Coleman *et al.*, 1966) was tailored to the specific analysis, most subsequent data sets have been constructed from existing data.

Third, even if datasets are developed solely for the purpose of analyzing school factors, the existing literature has not given much guidance about the specific factors of interest or how they might interact. The uncertainty about standard components of inputs means that developing surveys or collecting basic data on schools is not aided much by prior work. The large longitudinal surveys of schooling have been developed to track students and to ascertain how schools affect performance. They have not been especially effective in identifying the elements of school inputs that are important.[13]

[13] Among the important surveys and data collections in the United States are the National Longitudinal Study of the High School Class of 1972, High School and Beyond, and the

Thus, the possibility of misspecification is a very real possibility in analysis of school outcomes, and particularly when the focus is on specific programs. This possibility can provide a ready explanation for some of the results of estimation. It also offers some general perspectives on how to proceed.

One direct implication is that continually pushing on the same basic specifications is unlikely to provide persuasive evidence that any specific inputs are systematically related to student outcomes. There is little reason to believe that the accumulated evidence of several decades will be overturned by new studies that merely reproduce the prior analyses.

This issue is particularly important when considering the analysis of specific programs. Many programs, as discussed previously, appear to be aimed at specific kinds of teachers, so that their impact interacts directly with other teacher characteristics. For example, some programs in teaching reading to children attempt to encourage the teacher to innovate in a variety of ways, while other programs are highly scripted to remove variations across teachers. It seems from these very different perspectives on reading programs that the observed outcomes could depend very directly on the specific set of teachers in a given school.

A variant of this that appears particularly important is that implementation of programs varies dramatically across school settings. It is, for example, frequently mentioned anecdotally that school staff might revolt against programs that are imposed upon them centrally. On the other hand, programs where the staff participates in choice and implementation may receive a much more positive reception. Yet, very seldom is any information available about the implementation of specific programs.

One can simply analyze estimated program effects as average treatment effects, neglecting issues of implementation or the mix of teachers. In other words, can we say anything about the typical effect on student outcomes across a set of sites observed to be employing a particular program? There is nonetheless likely to be ambiguity in interpretation. If the existing programs are largely ones where programs were voluntarily introduced into schools, it may be difficult to infer what would happen with central decision making on the introduction of the program. In fact the 'average treatment effect' will depend directly on the mix of different implementation features.

In terms of the analysis of school resources, alternative approaches are available. One of the key parts of the analytical difficulties is separating the effects of teachers and resources into specific components such as amount of experience or character of educational background. To the extent that the different components interact with each other, it is very difficult to isolate the separate effects. The alternative perspective is avoiding this separation and estimating the 'total teacher effects.' In other words, one can, as described previously, estimate the impact of different teachers on student performance directly without concentrating on the specific components.

National Education Longitudinal Study of 1988. A new addition, the Educational Longitudinal Study of 2002, has yet to be released.

Importantly, measured characteristics explain little of the variation in teacher effectiveness. For example, while some studies have shown that measures of the teacher's own test score correlate with student performance, the correlation is weak, and this measure explains a small portion of the overall variation in teacher skill.[14] Similarly, other standard measures of teacher differences explain little or none of the overall variations in teacher quality that are observed. This kind of analysis has implications for the kinds of policy that are informed by the analysis, discussed below, but it generally circumvents some of the larger problems of estimating the effects of inputs.[15]

One other aspect of analysis and methodology is important. Recent concern about understanding causal effects, a warranted development in these empirical areas, does frequently run into some of the same issues. For example, in the debate on class size policy, a concern has been that the class sizes that are observed have been set purposefully by administrators. In particular, if administrators put the students most in need of help in smaller classes, class size will be positively related to student achievement – leading, as the argument goes – to obtaining estimates of class sizes with an unexpected positive sign. To deal with this, some researchers have looked for factors that drive class size but do not directly affect achievement. These instrumental variable approaches have, for example, looked at institutional rules (Angrist and Lavy, 1999) and at demographic factors (Hoxby, 2000). At the same time, these approaches assume that class size has a linear effect on achievement that is independent of other factors such as teacher quality.[16]

The instrumental variables approach represented by these and other studies are attempts to introduce randomness into the determination of class size. A direct way to do this, of course, is to run a random assignment experiment, where class size is varied randomly for a group of students. This is exactly what was done in Project STAR, the class size experiment in Tennessee during the mid-1980s (see Word *et al.*, 1990). Project STAR was designed to begin with kindergarten students and to follow them for four years. Three treatments were initially included: small classes (13–17 students); regular classes (22–25 students); and regular classes (22–25 students) with a teacher's aide. They found that students in small classes had higher achievement greater than those in large classes.

While there are questions about the quality of the randomization, larger questions exist about how to generalize any findings. The study reports no explicit consideration of teacher quality (although it supposedly randomized the

[14] Hanushek and Rivkin (2004).

[15] This approach would not work if there are significant interactions between teachers and students, that is, if the ability of a teacher to get learning gains depends on the specific group of teachers. Past work, however, suggests that this is not a major problem (Hanushek, 1992).

[16] The various instrumental variables studies reach different conclusions about class size: Angrist ad Lavy (1999) conclude that class size is important, at least when classes come close to the institutional limit of 40 students that exists in Israel; Hoxby (2000) concludes that class size variation is not an important factor in achievement. Another estimation involving instrumental strategies does not clear up the ambiguity (Hanushek, 2003).

assignment of both students and teachers). Yet the available evidence suggests that variations in teacher quality dominate any effects of class size. The average difference in performance of students in small kindergartens has been the focus of all attention to Project STAR, but the results actually differed widely by classroom. In only 40 out of 79 schools did the kindergarten performance in the small classroom exceed that in the regular classrooms (with and without aides). The most straightforward interpretation of this heterogeneity is that variations in teacher quality are extraordinarily important.

What is not considered is whether there is any interaction of teacher quality and class size. The underlying randomized assignment model assumes that the effects of class size are simply additive and uncorrelated with other determinants of achievement through the randomization of the treatments. But if the effect of class size is dependent on teacher quality, for example, the results of the experiment are a function of the specific distribution of teachers in the experiment. Generalizing to other situations would then require knowing more about this distribution and about how it fits into other settings where one might wish to employ class size reduction policies.

A very similar situation exists in terms of programs and larger reform plans. For example, many people have argued that a comprehensive reform strategy is needed where many aspects of a school are simultaneously altered. In such a situation, the evaluation approach is not entirely clear. One common suggestion is to match each treated school with a control school chosen to have similar characteristics. Then, typically a difference-in-differences approach is used to evaluate the program by comparing achievement growth with and without the program: the change in student achievement for schools using a given program is compared to the change over the same time period for a set of comparison schools. This approach, however, clearly depends first on matching schools on the 'right' characteristics. After that, the analysis faces the issues of generalization, which again involve possible interactions with unmeasured characteristics such as degree of implementation or quality of the teachers.

In sum, the general linear estimation of measured teacher, school, and program factors does not appear to be a productive way to proceed. It is also not simply an issue of concentrating on exogenous variation in the various factors but instead appears to involve more fundamental specification issues.

VI Implications for Policy

The policy story that follows this train of logic is very clear. Most importantly, the prevailing policy perspective that filters through to the supporting research is very narrow in the range of policies considered. The common view is that we must be able to identify a set of fixed characteristics of teachers or specific programs that can be instituted through central decision making. This viewpoint arises from a perspective that decision making is best made centrally – and central decision making requires being able to specify relevant inputs or effective programs accurately.

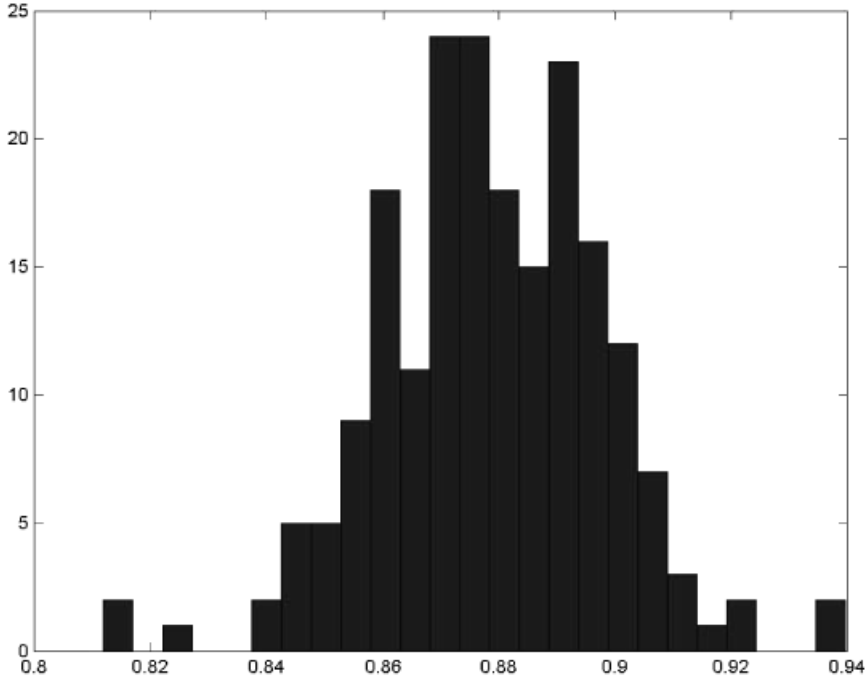

Figure 5. Estimated efficiency of central versus local decision making on programs.

Yet that is neither the only way nor the most common way to think of decision making. The more natural way to think about this is for local personnel to make decisions that incorporate information about the relevant teachers and circumstances.

Achieving this in an efficient manner would almost certainly require a different set of incentives in schools. Currently in schools there are not strong incentives to improve student performance, so decentralization of decisions may not be appropriate. If on the other hand there were stronger incentives at the school – say by having teacher pay or retention depend in some manner on performance in the classroom, the case would be very different.

This conclusion does reflect a presumption that things like teacher quality or the implementation of programs is more observable and controllable at the local level than at the central level. Some evidence about this does exist. Both Armor *et al*. (1976) and Murnane (1975) estimate the total effects of teachers (in the fixed effect way described above) and find that these differences are highly correlated with the subjective assessments of teachers by the principals in the schools. In other words, teacher quality is observable at the local school. This may be short of knowing fully about the interactions of programs and quality or the expected course of implementation. Nonetheless, it may be reasonable to infer that matters like the interaction of teachers with specific programs can be observed locally but not centrally.

The implications of centralized versus decentralized decision making on programs can be demonstrated from our previous simulations. Consider installing a compensatory program when the specific teacher is below average and a reinforcing program when the teacher is above average. Figure 5 shows the distribution of efficiency that corresponds to the situation depicted in Figure 4 where programs were set without regard to teacher quality.[17] Efficiency is the ratio of average outcomes within a sample that are obtained with a centrally imposed policy as opposed to a locally set policy. On average in our illustration, central programs achieve 88 percent of the achievement that is possible with decision making that takes teacher skills into account.

The best form of incentives is currently not well understood, because we have relatively little experience with applying incentive systems to schools (Hanushek and others, 1994). Nonetheless, the alternatives imply considerable loss, given our current ability to measure and specify the precise form of policies.

VII Some Conclusions

This paper began with some simple observations: prevailing estimates do not suggest a consistent effect of resources on student outcomes. Yet, it is hard to imagine that all of these resources would truly have a zero or negative impact.

One possibility is that the impact of resources is complicated – involving interactions with various inputs that are not observed or are not understood. The simplest notion is that teacher quality interacts with resources to determine outcomes. In the illustrative calculations, teacher quality essentially determines the efficiency with which resources are converted into student achievement. In this, we see that resource estimates are biased and also tend to be statistically insignificant.

Perhaps more important, one can introduce the idea of an educational program that has differential effects on teachers of different quality. Such programs would not be considered within standard formulations of production theory, because they represent the process by which different inputs are combined. Nonetheless, in standard evaluation problems in schools, estimates of 'program effects' are common.

We consider a case where there is not a single 'best program' in the sense of one program that lifts everybody's outcomes proportionally. Instead, we consider deciding among two programs – one that reinforces existing teacher quality and one that compensates for lower quality. Within this framework, standard empirical approaches including simple regressions or various difference-in-difference estimators can give very misleading estimates.

Of course these simulations do not prove that anything like this lies behind the estimates that have been developed in the past. They are merely designed to illustrate the fragility of most standard approaches to understanding student achievement when there are complicated interactions between school resources

[17] Note again that the program has varying impacts. Compensatory is simply negatively correlated with teacher quality and reinforcing is positively correlated, but each is imprecisely attuned to the specific teacher.

and characteristics of family and teachers. This possibility does bear directly on the analytical approaches that are likely to pay off. Specifically, even random assignment experiments run into trouble when important information about inputs relevant to the student are missing.

This perspective does give meaning to the idea that certain resources may be necessary but are not by themselves sufficient. When the effectiveness of the resources is a function of local managerial factors, including those of teachers, simply making the resources available does not solve the problems of improving student performance. Note also that this is a problem of both analysis and policy. Analysis that ignores the underlying effectiveness of inputs is likely to provide very misleading results. Policy that ignores the full constellation of inputs may also yield disappointing results.

Considerable attention has recently been devoted to develop a catalog of 'best practices,' programs commonly assumed to be ones that are broadly effective. Without being able to specify the circumstances under which a given program works, this search may prove fruitless.

Acknowledgement

Lei Zhang provided valuable assistance in developing the simulation estimates.

References

ANGRIST, J. D. and LAVY, V. (1999). 'Using Maimondides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics*, **114**, 2, May, pp. 533–75.

ANGRIST, J. D. and LAVY, V. (2001). Does teacher training affect pupil learning? Evidence from matched comparisons in Jerusalem public schools. *Journal of Labor Economics*, **19**, 2, April, pp. 343–69.

ARMOR, D. J., CONRY-OSEGUERA, P., COX, M., KING, N., MCDONNELL, L., PASCAL, A., PAULY, E. and ZELLMAN, G. (1976). *Analysis of the School Preferred Reading Program in Selected Los Angeles Minority Schools*. Santa Monica, CA: Rand Corp.

COLEMAN, J. S., CAMPBELL, E. Q., HOBSON, C. J., MCPARTLAND, J., MOOD, A. M., WEINFELD, F. D. and YORK, R. L. (1966). *Equality of Educational Opportunity*. Washington, DC: US Government Printing Office.

HANUSHEK, E. A. (1971). Teacher characteristics and gains in student achievement: estimation using micro data. *American Economic Review*, **60**, 2, May, pp. 280–88.

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

HANUSHEK, E. A. (2002). Publicly provided education. In A. J. Auerbach and M. Feldstein (eds.), *Handbook of Public Economics*. Amsterdam: Elsevier, pp. 2045–141.

HANUSHEK, E. A. (2003). The failure of input-based schooling policies. *Economic Journal*, **113**, 485, February, pp. F64–98.

HANUSHEK, E. A. and OTHERS. (1994). *Making schools work: improving performance and controlling costs*. Washington, DC: Brookings Institution.

HANUSHEK, E. A. and RIVKIN, S. G. (2004). How to improve the supply of high quality teachers. In D. Ravitch (Ed.), *Brookings Papers on Education Policy 2004*. Washington, DC: Brookings Institution Press.

HARBISON, R. W. and HANUSHEK, E. A. (1992). *Educational performance of the poor: lessons from rural northeast Brazil*. New York: Oxford University Press.

HOXBY, C. M. (2000). The effects of class size on student achievement: new evidence from population variation. *Quarterly Journal of Economics*, **115**, 3, November, pp. 1239–85.

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

MURNANE, R. J. and PHILLIPS, B. (1981). What do effective teachers of inner-city children have in common? *Social Science Research*, **10**, 1, March, pp. 83–100.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (2001). *Education at a glance*. Paris: Organization for Economic Co-operation and Development.

RIVKIN, S. G., HANUSHEK, E. A. and KAIN, J. F. (2001). Teachers, schools, and academic achievement, Working Paper No. 6691, National Bureau of Economic Research (revised).

US DEPARTMENT OF EDUCATION (2002). *Digest of Education Statistics, 2001*. Washington, DC: National Center for Education Statistics.

WORD, E., JOHNSTON, J., BAIN, H. P., FULTON, B. D., ZAHARIES, J. B., LINTZ, M. N., ACHILLES, C. M., FOLGER, J. and BREDA, C. (1990). *Student/teacher achievement ratio (STAR), Tennessee's K-3 class size study: final summary report, 1985–1990*. Nashville, TN: Tennessee State Department of Education.