The Economic Journal, **113** (*February*), F64–F98. © Royal Economic Society 2003. Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

THE FAILURE OF INPUT-BASED SCHOOLING POLICIES*

Eric A. Hanushek

In an effort to improve the quality of schools, governments around the world have dramatically increased the resources devoted to them. By concentrating on inputs and ignoring the incentives within schools, the resources have yielded little in the way of general improvement in student achievement. This paper provides a review of the US and international evidence on the effectiveness of such input policies. It then contrasts the impact of resources with that of variations in teacher quality that are not systematically related to school resources. Finally, alternative performance incentive policies are described.

Academic and policy interest in improving schools has followed directly from recognition of the importance of human capital formation to both individuals and society. Much of the motivation comes from theoretical and empirical analyses of the relationship between income, productivity, and economic growth and the quantity of schooling of individuals – the most common proxy for human capital levels. For the most part, however, policy initiatives do not focus on the quantity of schooling but instead on the quality of schooling. It is here that controversy about research into the determinants of quality has led to ambiguities about policy. This discussion reviews basic evidence on student performance and puts it into the context of contemporary policy debates. The central conclusion is that the commonly used input policies – such as lowering class sizes or tightening the requirements for teaching credentials – are almost certainly inferior to altered incentives within the schools.

The general arguments about schooling in the US and elsewhere in the world have a simple structure. First, the high returns to additional schooling are noted. In the US these returns have grown dramatically over the past 20 years, particularly for a college education. During the 1990s, for example, an average college graduate earned in excess of a 70% premium above the average high school graduate, e.g., Pierce and Welch (1996). Schooling returns in other countries, while varying, have also been high (Psacharopoulos, 1989, 1994; OECD, 2001). Second, having noted the individual returns to schooling, the policy discussion quickly shifts to the necessity to invest further in human capital, which is translated directly into an argument for providing more public funding for schools. While again there is some variation depending on each country's school attainment rates, the arguments for increased funding generally do not revolve around supporting more years of schooling for individuals but instead concentrate on improving the quality of the existing years of schooling. Embedded in this shift are a number of presumptions that are widely held. One is that quality has the same payoffs as quantity of schooling. Another is that greater funding will lead to improved quality. This paper considers the latter presumption – that spending and quality are closely related – in detail.

* This research has been supported by a grant from the Packard Humanities Institute.

Before entering into the central discussion, however, it is useful to establish some facts about the value of 'quality'. There is mounting evidence that quality – generally measured by test scores – is related to individual earnings,[1] productivity, and economic growth. While focusing on the estimated returns to years of schooling, early studies of wage determination tended to indicate relatively modest impacts of variations in cognitive ability after holding constant quantity of schooling. More recent direct investigations of cognitive achievement, however, have suggested generally larger labour market returns to measured individual differences in cognitive achievement. For example, Bishop (1989, 1992), O'Neill (1990), Grogger and Eide (1993), Blackburn and Neumark (1993, 1995), Murnane *et al.* (1995), Neal and Johnson (1996), Murnane *et al.* (2000), and Murnane *et al.* (2001) each find that the earnings advantages to higher achievement on standardised tests are quite substantial.[2] International evidence is less plentiful but also demonstrates a labour market return to cognitive skills (Currie and Thomas, 2000; Boissiere *et al.*, 1985). The difficulty of separating cognitive skills from pure schooling has nonetheless made this estimation very difficult (Cawley *et al.*, 2000; Heckman and Vytlacil, 2001) and thus leaves ambiguity about the exact magnitude of effects.

Similarly, society appears to gain in terms of productivity. Hanushek and Kimko (2000) demonstrate that quality differences in schools have a dramatic impact on productivity and national growth rates. This study of growth rates incorporates information on international mathematics and science examinations into standard cross-country growth regressions. It finds a very strong relationship between test performance and national growth and a smaller relationship between quantity of schooling and growth. A series of investigations of the structure suggests a causal relationship.

An additional part of the return to school quality comes through continuation in school. There is substantial US evidence that students who do better in school, either through grades or scores on standardised achievement tests, tend to go

[1] While human capital has been central to much of labour economics for some four decades, its measurement has been more problematic. The most commonly employed measure is simply the years of school completed, but this measure neglects any quality differences arising from both school and nonschool differences across individuals. The most commonly employed measure of quality involves cognitive test scores, although the adequacy of this measure has not been fully investigated; see, for example, Murnane *et al.* (2001). Early analyses of earnings employing test scores generally treated them as fixed measures of ability difference, e.g., Griliches (1974). Considerable evidence, however, including some presented below indicates that the typical cognitive tests are very much dependent on both families and schools.

[2] These results are derived from quite different approaches. Bishop (1989) considers the measurement errors inherent in most testing situation and demonstrates that careful treatment of that problem has a dramatic effect on the estimated importance of test differences. O'Neill (1990), Grogger and Eide (1993), Bishop (1991), and Neal and Johnson (1996) on the other hand simply rely upon more recent labour market data along with more representative sampling and suggest that the earnings advantage to measured skill differences is larger than that found in earlier time periods and in earlier studies (even without correcting for test reliability). Murnane *et al.* (1995) considering a comparison over time, demonstrate that the results of increased returns to measured skills hold regardless across simple analysis and error-corrected estimation. Murnane *et al.* (2000) and Murnane *et al.* (2001) employ representative samples but introduce other measures of individual skill. Blackburn and Neumark (1993, 1995), like much of the early literature, concentrate mainly on any bias in the estimated rates of return to schooling when ability measures are omitted.

farther in school; see, for example, Dugan (1976); Manski and Wise (1983). Rivkin (1995) finds that variations in test scores capture a considerable proportion of the systematic variation in high school completion and in college continuation, so that test score differences can fully explain black-white differences in schooling. Bishop (1991) and Hanushek *et al.* (1996) find that individual achievement scores are highly correlated with school attendance. Behrman *et al.* (1998) find strong achievement effects on both continuation into college and quality of college; moreover, the effects are larger when proper account is taken of the endogeneity of achievement. Hanushek and Pace (1995), using the High School and Beyond data, find that college completion is significantly related to higher test scores at the end of high school.

Quality is nonetheless virtually impossible to dictate through policy. The quest for improved quality has undoubtedly contributed to recent expansions in the resources devoted to schools in the US and other countries. Eager to improve quality and unable to do it directly, government policy typically moves to what is thought to be the next best thing – providing added resources to schools. Broad evidence from the experience in the US and the rest of the world suggests that this is an ineffective way to improve quality.

This Feature both points to interest in the topic and highlights some of the interpretive issues that arise. The discussion of school policy frequently involves an intensity not common to many other academic debates, because the results of analyses of schools at times have direct influence on policy. Thus, for example, the arguments for reduced class size in Krueger (2002), largely reproduced in this Feature as Krueger (2003), have already provided fuel for advocates of lowering class sizes.[3] And the Dustmann *et al.* (2003) article will similarly find a waiting policy audience as teacher employment policies are debated around the globe. What makes the issues more complicated is the difficulty of interpreting results, based as they are on imperfect data and incomplete description of the underlying structure (Todd and Wolpin, 2003).

Because class size policies are currently being broadly discussed, attention to other significant dimensions of input policy decisions tends to be neglected. The following article presents the available evidence on a broad set of resource policies in schools. While some of the evidence is less reliable than others, the overall picture is remarkably consistent. Even discounting significant portions of the available evidence, one is left with the clear picture that input policies of the type typically pursued have little chance of being effective.

1. School Inputs and Outcomes

Much of the policy discussion throughout the world concentrates on schooling inputs, a seemingly natural focus. And, with the longstanding importance that has been attached to schooling, considerable change has occurred in the levels of

[3] Hanushek (2002) provides a critique of Krueger (2002), which differs inconsequentially from the Feature version. Specifically, in searching for alternative weightings of the results, the Feature version reweights estimates by the 'impact index' of journals instead of by citations. As Krueger notes, this has minimal effect on the estimates.

common inputs. Class sizes have fallen, qualifications of teachers have risen, and expenditures have increased. Unfortunately, little evidence exists to suggest that any significant changes in student outcomes have accompanied this growth in resources devoted to schools. Because many find the limited relationship between school inputs and student outcomes surprising and hard to believe, this section delves into the evidence available on this score in some detail.

These data on aggregate cost and performance provide strong *prima facie* evidence that simple resource policies are not generally effective. Much of the current policy discussion argues that with additional resources it would be possible to implement programmes or approaches that lift student achievement. Of course, these are precisely the same arguments made over the past decades. The validity of current proposals rests on these current proposals being notably superior to the policies of the past (which were hypothesised at the time also to be superior policies).

1.1. *Aggregate US Data*

The simplest and perhaps clearest demonstration of the resource story is found in the aggregate US data over the past few decades. The US, operating under a system that is largely decentralised to the 50 separate states, has pursued the conventionally advocated resource policies vigorously. Table 1 tracks the patterns of pupil-teacher ratios, teacher education, and teacher experience. Between 1960 and 2000, pupil-teacher ratios fell by almost 40%. The proportion of teachers with a master's degree or more over doubled so that a majority of all US teachers today have at least a master's degree. Finally, median teacher experience – which is more driven by demographic cycles than active policy – increased significantly, almost doubling since its trough in 1970.

American teachers are heavily unionised, and the most common structure of teacher contracts identifies teacher education levels and teacher experience as the driving force behind salaries. Thus, as teacher inputs rise and as the numbers of students per teachers decline, expenditure per pupil rises. As seen in the bottom row of Table 1, real expenditures per pupil more than tripled over this period.[4] In fact, this period is not special in US schools. Over the entire 100 years of 1890-1990, real spending per pupil rose by at a remarkably steady pace of 3½% per year (Hanushek and Rivkin, 1997). Over this longer period, real per student expenditure in 1990 dollars goes from $164 in 1890 to $772 in 1940 to $4,622 in 1990 – roughly quintupling in each 50 year period.[5]

[4] The calculation of real expenditures deflates by the Consumer Price Index. If the alternative of a wage deflator were employed, the calculated rate of real increase over this period would not change much. Baumol's disease (Baumol, 1967) is frequently cited at this point to explain increases in input costs without increasing real inputs. Specifically, if service sectors are ones where productivity growth is necessarily low – say, for technological reasons – they will face cost pressures in the hiring of inputs, putting the service sector (technologically backward) at a disadvantage. Over this period, however, such pressures cannot explain the patterns of inputs and outputs to schooling (Hanushek, 1997*b*).

[5] These calculations differ from those in Table 1 both in using a different deflator (GDP deflator in 1990 dollars) and in calculating spending per pupil on a membership rather than an attendance basis.

Table 1

Public School Resources in the US, 1960–2000

	1960	1970	1980	1990	2000
Pupil-teacher ratio	25.8	22.3	18.7	17.2	16.0
% teachers with master's degree or more	23.5	27.5	49.6	53.1	56.2*
median years teacher experience	11	8	12	15	15*
current expenditure/ADA (2000/2001 $s)	$2,235	$3,782	$5,124	$6,867	$7,591

Note: *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 question remains, what was obtained for these spending increases? Since the early 1970s, a random sample of students in the US has been given tests at differing ages in various subjects under the auspices of the National Assessment of Educational Progress, or NAEP. These tests have been designed to provide a consistent measure of performance over time. Figure 1 gives performance data for the same period as the previously described input data. In this Figure the pattern of average performance by 17-year-olds is traced for reading, mathematics, and science. The performance of students in mathematics and reading is ever so

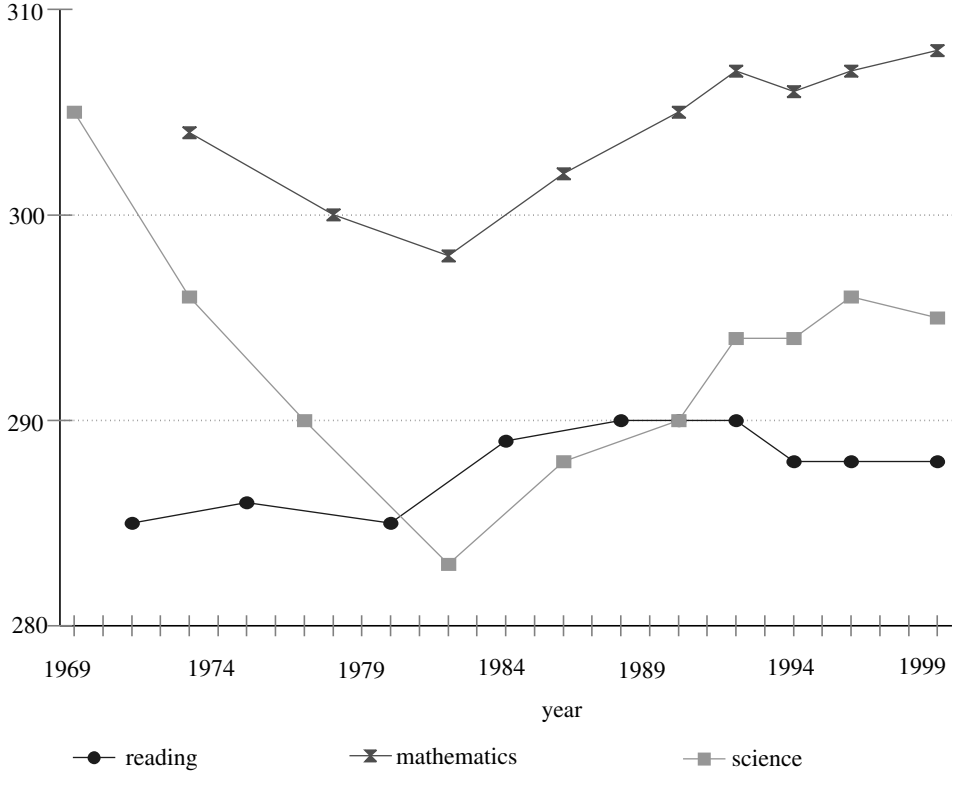

Fig. 1. *Scores by 17-year-olds on National Assessment of Educational Progress, 1969–99*

slightly higher in 1999 than 30 years before when spending was dramatically lower.[6] The performance of students in science is significantly lower in 1999 than it was in 1970. Writing performance (not shown) was first tested in 1984 and declined steadily until 1996 when testing was discontinued.

The only other test that provides a national picture of performance over a long period of time is the Scholastic Aptitude Test, or SAT. This college admissions test has the advantage of providing data going back to the 1960s but the disadvantage of being a voluntary test taken by a selective subset of the population.[7] Scores on this test actually plunged from the mid-1960s until the end of the 1970s, suggesting that the NAEP scores that begin in the 1970s may understate the magnitude of the performance problem.[8]

In simplest terms, input policies have been vigorously pursued over a long period of time, but there is no evidence that the added resources have improved student performance, at least for the most recent three decades when it has been possible to compare quantitative outcomes directly. This evidence suggests that the efficacy of further input-based policies depends crucially on improved use of resources compared to past history.

Two arguments are made, however, for why the simple comparison of expenditures and student performance might be misleading:

1. The characteristics of students may have changed such that they are more difficult (and expensive) to educate now than in the past:
2. Other expansions of the requirements on schools have driven up costs but would not be expected to influence observed student performance.

1.1.1. *Changes in students*

One simple explanation for why added resources yield no apparent performance improvement is that students are more poorly prepared or motivated for school over time, requiring added resources just to stay even. For example, there have been clear increases in the proportion of children living in single-parent families and, relatedly, in child poverty rates – both of which are hypothesised to lead to lower student achievement. Between 1970 and 1990, children living in poverty families rose from 15 to 20%, while children living with both parents declined from 85 to 73%. The percentage of children not speaking English at home also rose from 9% in 1980 to 17% in 2000. But, there have also been other trends that appear to be positive forces on student achievement. Family sizes have fallen and

[6] The cumulative nature of the educational process implies that scores will reflect both current and past spending. A 17-year-old in 1970, for example, would have entered school in the late 1950s, implying that the resource growth in Table 1 that goes back to 1960 is relevant for comparison with the NAEP performance data.

[7] NAEP samples are not tainted by selection. The school completion rate and the rate of attendance of private schools have been essentially constant over the period of the NAEP tests and testing involves a random sample of public school children.

[8] Analyses of the changes in SAT scores suggest that a portion of the decline in scores comes from increases in the rate of test taking but that the decline also has a real component of lesser average performance over time (Wirtz, 1977; Congressional Budget Office, 1986).

parental education levels have improved. Among all families with children, the percentage with three or more children fell from 36 to 20%. Moreover, over the same period, adults aged 25–29 with a high school or greater level of schooling went from 74 to 86% (up from 61% in 1960). Finally, enrollment in kindergarten and pre-school increased dramatically over the period.

It is difficult to know how to net out these opposing trends with any accuracy. Extensive research, beginning with the Coleman Report (Coleman *et al.*, 1966) and continuing through today (Hanushek, 1997*a*), has demonstrated that differences in families are very important for student achievement. Most of these studies have not focused their primary attention on families, however, and thus have not delved very far into the measurement and structure of any family influences. Grissmer *et al.* (1994) attempt to sort out the various factors in a crude way. That analysis uses econometric techniques to estimate how various family factors influence children's achievement at a point in time. It then applies these cross-sectionally estimated regression coefficients as weights to the trended family background factors identified above. Their overall findings are that black students performed better over time than would be expected from the trends in black family factors. They attribute this better performance to improvements in schools. On the other hand, white students, who make up the vast majority, performed worse over time than would be expected, leading presumably to the opposite conclusion that schools for the majority of students actually got worse over time.

While there are reasons to be sceptical about these precise results, they do suggest that the spending-performance relationship is not driven in any simple way by changes in student preparation.[9] Changes in family inputs have occurred over time, making it possible that a portion of the increased school resources has gone to offset adverse factors. The evidence is nonetheless quite inconclusive about even the direction of any trend effects, let alone the magnitude. The only available quantitative estimates indicate that changing family effects are unable to offset the large observed changes in pupil-teacher ratios and school resources and may have even worked in the opposite direction, making the performance of schools appear better than it was.

1.1.2. *Exogenous cost increases*

The most discussed cost concern involves 'special education', programmes to deal with students who have various disabilities. The issue is that these programmes are

[9] Scepticism about the results from Grissmer *et al.* (1994) comes from methodological problems. First, they do not observe or measure differences in schools but instead simply attribute unexplained residual differences in the predicted and observed trends to school factors. In reality any factor that affects achievement, that is unmeasured, and that has changed over their analysis period would be mixed with any school effects. Second, in estimating the cross-sectional models that provide the weights for the trending family factors, no direct measures of school inputs are included. In the standard analysis of misspecified econometric models, this omission will lead to biased estimates of the influence of family factors if school factors are correlated with the included family factors in the cross-sectional data that underlie their estimation. For example, better educated parents might systematically tend to place their children in better schools. In this simple example, a portion of the effects of schools will be incorrectly attributed to the education of parents, and this will lead to inappropriate weights for the trended family inputs. Third, one must believe either that the factors identified are the true causal influences or that they are stable proxies of the true factors, but there is doubt about this (Mayer, 1997).

expensive but the recipients tend not to take standardised tests. Thus, even if special education programmes are effective (Hanushek *et al.*, 2002), the increased expenditures on special education will not show up in measured student performance.

Concerns about the education of children with both physical and mental disabilities were translated into federal law with the enactment of the Education for All Handicapped Children Act in 1975. This Act prescribed a series of diagnostics, counselling activities, and educational services to be provided for handicapped students. To implement this and subsequent laws and regulations school systems expanded staff and programmes, developing entirely new administrative structures in many cases to handle 'special education'. The general thrust of the educational services has been to provide regular classroom instruction where possible ('mainstreaming') along with specialised instruction to deal with specific needs. The result has been growth of students classified as the special education population even as the total student population fell. Between 1977 and 1999, the percentage of students classified as disabled increases from 9.3 to 13.0%. Moreover, the number of special education teachers increases much more rapidly than the number of children classified as disabled.

The magnitude of special education spending and its growth, however, are insufficient to reconcile the cost and performance dilemma. Using the best available estimate of the cost differential for special education – 2.3 times the cost of regular education (Chaikind *et al.*, 1993), the growth in special education students between 1980 and 1990 can explain less than 20% of the expenditure growth (Hanushek and Rivkin, 1997). In other words, while special education programmes have undoubtedly influenced overall expenditures, they remain a relatively small portion of the total spending on schools.

Direct estimates of other exogenous programmes and changes resulting from other academic aspects of schools such as language instruction for immigrants or nonacademic programmes such as sports, art, or music are not readily available. Nonetheless, no evidence suggests that these can explain the magnitude of spending growth.

1.2. *Aggregate International Data*

Most other countries of the world have not tracked student performance over any length of time, making analyses comparable to the US discussion impossible. Nonetheless, international testing over the past four decades permits an overview of spending across countries. Seven different mathematics and science tests have been given between the early 1960s and 1995 to students at different grade levels in a varying set of voluntarily participating nations. (Only the US and the UK participated in all testing.) The test performance across time, updated from Hanushek and Kimko (2000), is summarised in Figure 2. In this Figure the scores for each test have been aggregated across grade levels and subtests and the world average in each year is set to 50.[10] While the tests were not designed to track

[10] A description of the individual tests and the aggregation of scores is given in Hanushek and Kim (1995). The figure drops off the first year of testing (1965) when there are questions about representativeness of the sampling. It also does not include the most recent testing (TIMSS-R in 1999).

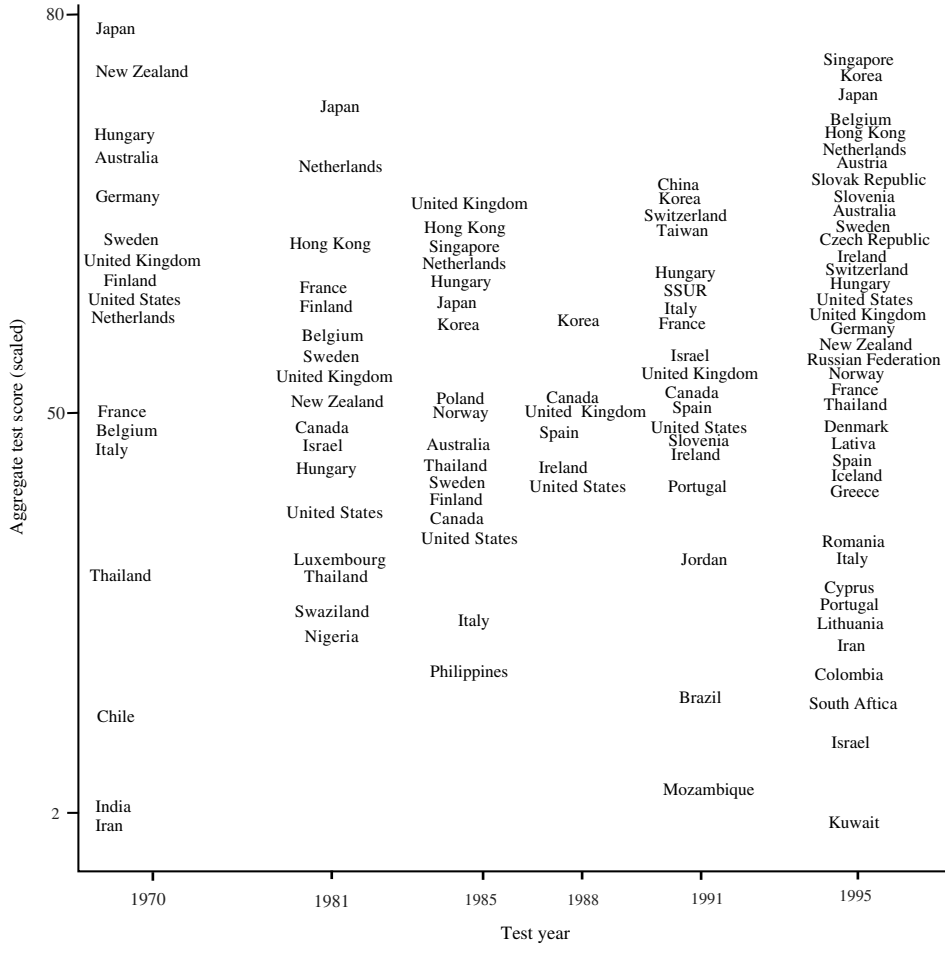

Fig. 2. *Performance on International Mathematics and Science Examinations*

performance over time and while they have been taken by varying countries, they can be equated using the patterns of US test performance reported in Figure 1. This alternative normalisation does not affect the pattern because the pattern of performance of US students is essentially the same on both national and international exams.

Performance bears little relationship to the patterns of expenditure across the countries. Table 2 provides the distribution of 1998 primary and secondary school spending per pupil across a set of countries participating in the recent Third International Mathematics and Science Study (TIMSS). These countries are sorted by order of aggregate performance on TIMSS, and a quick glance at the Table highlights the incongruity of spending and performance.[11] The simple

[11] Data from OECD (2001) provide a consistent set of spending figures converted to US dollars on a purchasing power parity basis. A total of 23 TIMSS countries have reported spending figures.

Table 2

Primary and Secondary School Spending per Pupil in 1998, Sorted by Overall TIMSS Performance (United States dollars in purchasing power parity)

Country	Primary School Spending	Secondary School Spending
Korea	2,838	3,544
Japan	5,075	5,890
Belgium*	3,771	6,104
Netherlands	3,795	5,304
Austria	6,065	8,163
Australia	3,981	5,830
Sweden	5,579	5,648
Czech Republic	1,645	3,182
Ireland	2,745	3,934
Switzerland	6,470	9,348
Hungary	2,028	2,140
United States	6,043	7,764
United Kingdom	3,329	5,230
Germany	3,531	6,209
Norway	5,761	7,343
France	3,752	6,605
Thailand	1,048	1,177
Denmark	6,713	7,200
Spain	3,267	4,274
Greece	2,368	3,287
Italy	5,653	6,458
Portugal	3,121	4,636
Israel	4,135	5,115

* Flemish and French speaking Belgium combined into single average expenditure.
Source: OECD (2001).

correlation between secondary school spending and TIMSS score is an insignificant 0.06.

International comparisons, of course, amplify the problems of possible contamination of the influence of factors other than schools that was considered previously in the case of the US. As a preliminary attempt to deal with some of these issues, Hanushek and Kimko (2000) estimate models that relate spending, family backgrounds, and other characteristics of countries to student performance for the tests prior to 1995.[12] This estimation consistently indicates a statistically significant negative effect of added resources on performance after controlling for other influences.

Gundlach *et al.* (2001) consider changes in scores of a set of developed nations between 1970 and 1995 and their relationship to spending changes. They conclude that productivity of schools has fallen dramatically across these countries. Woessman (2000, 2001) also performs a related analysis that relies on just the 1995 performance information from TIMSS. His analysis suggests that traditional

[12] The estimation includes average schooling of parents, population growth rates, school participation rates, and separate intercepts for each of the different tests. Several measures of school resources including spending as a proportion of GNP, current expenditures per student, and class size in elementary and secondary schools were also included.

resource measures bear little consistent relationship to differences in scores among the 39 nations participating in TIMSS for 13-year-olds.

1.3. *Conclusions from Aggregate Data*

Analysis of aggregate performance data is subject to a variety of problems. Any relationship between resources and student achievement – whether within a single country or across different countries – might be distorted by other influences on performance. Nonetheless, the variations in resources are huge, suggesting that any effect should be apparent in even crude comparisons. No significant effect of spending comes through in the aggregate, even when consideration of family background differences is introduced.

Any claim that a given set of estimated resource effects provides support for broad increases in specific inputs – such as argued by Krueger (2002, 2003) – must be reconciled with the aggregate data that show no past effects of extensive pursuit of such policies.

2. Econometric Evidence

The aggregate story is supported by an extensive body of direct evidence coming from detailed econometric analyses of student achievement. This evidence has been motivated by a monumental governmental study of US achievement that was conducted in the mid-1960s. The 'Coleman Report' (Coleman *et al.*, 1966) presented evidence that was widely interpreted as saying that schools did not matter. The most important factor in achievement was the family, followed by peers in school. This study led to a great amount of research – research that has supported part of the Coleman study but, more importantly, has clarified the interpretation.

2.1. *US Estimates*

The statistical analyses relevant to this work have a common framework that has been well-understood for some time (Hanushek, 1979). Student achievement at a point in time is related to the primary inputs: family influences, peers, and schools. The educational process is also cumulative, so that both historical and contemporaneous inputs influence current performance.

With the exception of the Coleman Report, the subsequent analysis seldom has relied on data collected specifically for the study of the educational process. Instead, it has tended to be opportunistic, employing available data to gain insights into school operations. The focus of much of this work has been the effect of varying resources on student achievement. This focus flows from the underlying perspective of production functions; from its obvious relevance for policy; and from the prevalence of relevant resource data in the administrative records that are frequently used.

The summary of production in US schools begins with all of the separate estimates of the effects of resources on student performance, and then

concentrates on a more refined set of estimates.[13] The underlying work includes all published analyses prior to 1995 that include one of the resource measures described below, that have some measure of family inputs in addition to schools, and that provides the sign and statistical significance of the resource relationship with a measurable student outcome. The 89 individual publications that appeared before 1995 and that form the basis for this analysis contain 376 separate production function estimates. While a large number of analyses were produced as a more or less immediate reaction to the Coleman Report, half of the available estimates have been published since 1985. Of course, a number of subsequent analyses have also appeared since 1995. While not formally assessed, it is clear that including them would not significantly change any of the results reported here, given their mixed results and the large number of prior estimates.

Understanding the character of the underlying analyses is important for the subsequent interpretation. Three-quarters of the estimates rely on student performance measured by standardised tests, while the remainder uses a variety of different measures including such things as continuation in school, dropout behaviour, and subsequent labour market earnings. Not surprisingly, test score performance measures are more frequently employed for studying education in primary schools, while a vast majority of the analyses of other outcomes relate to secondary schools. The level of aggregation of the school input measures is also an issue considered in detail below. One-quarter of the estimates consider performance in individual classrooms, while 10% focus on school inputs only at the level of the state. Moreover, fully one-quarter of the estimates employing nontest measures rely solely on interstate variations in school inputs.

Table 3 presents the overall summary of basic results about the key resources that form the basis for most overall policy discussions.[14] The standard hypothesis driving policy initiatives is that each of these resources should have a positive

[13] Individual publications typically contain more than one set of estimates, distinguished by different measures of student performance, by different grade levels, and frequently by entirely different sampling designs. If, however, a publication includes estimates of alternative specifications employing the same sample and performance measures, only one of the alternative estimates is included. As a general rule, the tabulated results reflect the estimates that are emphasised by the authors of the underlying papers. In some cases, this rule did not lead to a clear choice, at which time the tabulation emphasised statistically significant results among the alternatives preferred by the original author. An alternative approach, followed by Betts (1996), aggregates all of the separate estimates of a common parameter that are presented in each individual paper. Still another approach, followed by Krueger (2002, 2003), aggregates all estimates in a given publication into a single estimate, regardless of the underlying parameter that is being estimated (see discussion below).

[14] A more complete description of the studies can be found in Hanushek (1997*a*), which updates the analysis in Hanushek (1986). The tabulations here correct some of the original miscoding of effects in these publications. They also omit the estimates from Card and Krueger (1992*b*). In reviewing all of the studies and estimates, it was discovered that the results of that paper were based on models that did not include any measures of family background differences and thus could not be interpreted as identifying any resource parameter. As a minimal quality criterion, tabulated estimates must come from statistical models that include some measure of family background, since omission will almost certainly lead to biased resource estimates. Family backgrounds have been shown to be quite generally correlated with school resources and have been shown to have strong effects on student outcomes.

Table 3

Percentage Distribution of Estimated Effect of Key Resources on Student Performance, Based on 376 Production Function Estimates

Resources	Number of estimates	Statistically significant (%)		Statistically insignificant (%)
		Positive	Negative	
Real classroom resources				
Teacher-pupil ratio	276	14	14	72
Teacher education	170	9	5	86
Teacher experience	206	29	5	66
Financial aggregates				
Teacher salary	118	20	7	73
Expenditure per pupil	163	27	7	66
Other				
Facilities	91	9	5	86
Administration	75	12	5	83
Teacher test scores	41	37	10	53

Source: Hanushek (1997*a*) (revised, see text and footnote 14).

effect on student performance.[15] In terms of real classroom resources, only 9% of the estimates considering the level of teachers' education and 14% of the estimates investigating teacher-pupil ratios find positive and statistically significant effects on student performance.[16] These relatively small numbers of statistically significant positive results are balanced by another set finding statistically significant negative results – reaching 14% in the case of teacher-pupil ratios.[17] A higher proportion of estimated effects of teacher experience are positive and statistically significant: 29%. Importantly, however, 71% still indicate either worsening performance with experience or less confidence in any positive effect. Because more experienced teachers can frequently choose their school and/or students, a portion of the positive effects could actually reflect reverse causation (Greenberg and McCall, 1974; Murnane, 1981; Hanushek *et al.*, 2001*b*). In sum, the vast number of estimated real resource effects gives little confidence that just adding more of any of the specific resources to schools will lead to a boost in student achievement. Moreover, this statement does not even get into whether or not any effects are 'large'. Given the small confidence in just getting

[15] It is possible that the level and shape of the salary schedule with respect to experience are set to attract and retain an optimal supply of teachers and that the year-to-year changes in salaries do not reflect short run productivity differences. This possibility would introduce some ambiguity about expectations of estimates of experience and salary effects.

[16] The individual studies tend to measure each of these inputs in different ways. With teacher-pupil ratio, for example, some measure actual class size, while the majority measure teacher-pupil ratio. In all cases, estimated signs are reversed if the measure involves pupil-teacher ratios or class size instead of teacher-pupil ratio.

[17] While a large portion of the studies merely note that the estimated coefficient is statistically insignificant without giving the direction of the estimated effect, those statistically insignificant studies reporting the sign of estimated coefficients are split fairly evenly between positive and negative.

noticeable improvements, it seems somewhat unimportant to investigate the size of any estimated effects.

The financial aggregates and other inputs provide a similar picture. There is very weak support for the notion that simply providing higher teacher salaries or greater overall spending will lead to improved student performance. Per pupil expenditure has received the most attention, but only 27% of the estimated coefficients are positive and statistically significant. In fact, 7% even suggest some confidence in the fact that spending more would harm student achievement. In reality, as discussed below, analyses involving per pupil expenditure tend to be the lowest quality, and there is substantial reason to believe that even these results overstate the true effect of added expenditure.

2.1.1. *Study quality*

The tabulated analyses of educational performance clearly differ in quality and their potential for yielding biased results. Two elements of quality, both related to model specification and estimation, are particularly important. First, education policy in the US is made chiefly by the separate 50 states, and the resulting variations in spending, regulations, graduation requirements, testing, labour laws, and teacher certification and hiring policies are large. These important differences – which are also the locus of most current policy debates – imply that any estimates of student performance across states must include descriptions of the policy environment of schools or else they will be subject to standard omitted variables bias. The misspecification bias of models that ignore variations in state education policy (and other potential state differences) will be exacerbated by aggregation of the estimation sample. Second, as noted, education is a cumulative process, but a majority of analyses are purely cross-sectional with only contemporaneous measures of inputs. In other words, when looking at performance at the end of secondary schooling, many analyses include just measures of the current teachers and school resources and ignore the dozen or more prior years of inputs. Obviously, current school inputs will tend to be a very imperfect measure of the resources that went into producing ending achievement. This mismeasurement is strongest for any children who changed schools over their career (a sizable majority in the US) but also holds for students who do not move because of the heterogeneity of teachers within individual schools; see Hanushek *et al.* (2001*a*); Rivkin *et al.* (2001). Even if contemporaneous measures were reasonable proxies for the stream of cumulative inputs, uncertainty about the interpretation and policy implications would remain. But there is little reason to believe that they are good proxies.

While judgments about study quality often have a subjective element, it is possible to make straightforward distinctions based on violations of these two problems. We begin with the issue of measuring the policy environment. States differ dramatically in their policies, and ignoring any policies that have a direct impact will bias the statistical results if important policies tend to be correlated with the resource usage across states. While the direction of any bias depends on the magnitude and sign of correlation, under quite general circumstances, the severity will increase with the level of aggregation of the school inputs.

That is, any bias will tend to be more severe if estimation is conducted at the state level than if conducted at the classroom level (Hanushek *et al.*, 1996).[18]

Table 4 provides insight into the pattern and importance of the specific omitted variables bias resulting from lack of information about key educational policy differences. This Table considers two input measures: teacher-pupil ratio and expenditure per pupil. These inputs, on top of being important for policy, are included in a sufficient number of analyses at various levels of aggregation that they can point to the potential misspecification biases. As discussed previously, the overall percentage of all estimates of teacher-pupil ratios that are statistically significant and positive is evenly balanced by those that are statistically significant and negative. But this is not true for estimates relying upon samples drawn entirely within a single state, where the overall policy environment is constant and thus where any bias from omitting overall state policies is minimised or eliminated. For single state estimates, the statistically significant effects are disproportionately negative. Yet, as the samples are drawn across states, the relative proportion positive and statistically significant rises. For those aggregated to the state level where the expected bias is largest, almost two-thirds of the estimates are positive and

Table 4

Percentage Distribution of Estimated Effect of Teacher-Pupil Ratio and Expenditure per Pupil by State Sampling Scheme and Aggregation

Level of aggregation of resources	Number of estimates	Statistically significant (%)		Statistically insignificant (%)
		Positive	Negative	
a. Teacher-Pupil Ratio				
Total	276	14	14	72
Single state samples*	157	11	18	71
Multiple state samples†	119	18	8	74
Disaggregated within states‡	109	14	8	78
State level aggregation§	10	60	0	40
b. Expenditure per pupil				
Total	163	27	7	66
Single state samples*	89	20	11	69
Multiple state samples†	74	35	1	64
Disaggregated within states‡	46	17	0	83
State level aggregation§	28	64	4	32

*Estimates from samples drawn within single states.
†Estimates from samples drawn across multiple states.
‡Resource measures at level of classroom, school, district, or country, allowing for variation within each state.
§Resource measures aggregated to state level with no variation within each state.

[18] The discussion of aggregation is part of a broader debate trying to reconcile the findings of Card and Krueger (1992*a*) with those presented here. For a fuller discussion, see Burtless (1996). Of particular relevance is Heckman *et al.* (1996*a*, *b*), which raises other issues with the Card and Krueger estimation. Specifically, their key identifying assumption of no selective migration is violated. Similarly, assumptions about homogeneity of effects across schooling categories are found not to hold.

statistically significant. The pattern of results also holds for estimates of the effects of expenditure differences (which are more likely to come from highly aggregate analyses involving multiple states).[19]

This pattern of results is consistent with expectations from considering specification biases when favourable state policies tend to be positively correlated with resource usage. The initial assessment of effects indicated little reason to be confident about overall resource policies. This refinement on quality indicates that a number of the significant effects may further be artifacts of the sampling and methodology.

The second problem, improper consideration of the cumulative nature of the educational process, is a different variant of model specification. Relating the level of performance at any point in time just to the current resources is likely to be very misleading. The standard approach for dealing with this is the estimation of value-added models where attention is restricted to the growth of achievement over a limited period of time (where the flow of resources is also observed). By concentrating on achievement gains over, say, a single grade, it is possible to control for initial achievement differences, which will be determined by earlier resources and other educational inputs. In other words, fixed but unmeasured factors are eliminated.

Table 5 displays the results of estimates that consider value-added models for individual students. The top panel shows all such results, while the bottom panel follows the earlier approach of concentrating just on estimates within an individual state. With the most refined investigation of quality, the number of analyses gets quite small and selective. In these, however, there is no support for systematic improvements through increasing teacher-pupil ratios and hiring teachers with more graduate education. The effects of teacher experience are largely unaffected from those for the universe of estimates.

The highest quality estimates indicate that the prior overall results about the effects of school inputs were not simply an artifact of study quality. If anything, the total set of estimates understates the ineffectiveness of pure resources differences in affecting student outcomes.

The methodology of Krueger (2002, 2003) takes a different approach is tabulating the results – recording a single composite estimate for each publication.[20] He implies that he is making overall quality judgments in his tabulations when he selectively contrasts a few publications with both a large number of estimates and potentially damaging statistical problems with an analysis that has both a small

[19] Expenditure studies virtually never direct analysis at performance across different classrooms or schools, since expenditure data are typically available only at the district level. Thus, they begin at a more aggregated level than many studies of real resources. An alternative explanation of the stronger estimates with aggregation is that the disaggregated studies are subject to considerable errors-in-measurement of the resource variables. The analysis in Hanushek *et al.* (1996), however, suggests that measurement error is not the driving force behind the pattern of results.

[20] A separate approach to aggregating the econometric results, referred to as 'meta-analysis', has been proposed by Greenwald *et al.* (1996). Instead of just tabulating results, they propose formal statistical analysis. This approach, however, typically considers the wrong hypothesis for policy discussions (i.e. that all estimated coefficients for a given parameter are simultaneously zero). Further, these approaches invariably lack the necessary statistical information when they rely on just published results: see Hanushek (1996).

Table 5

Percentage Distribution of Other Estimated Influences on Student Performance, Based on Value-added Models of Individual Student Performance

Resources	Number of estimates	Statistically significant (%)		Statistically insignificant (%)
		Positive	Negative	
a. All estimates				
Teacher-pupil ratio	79	11	9	80
Teacher education	41	0	10	90
Teacher experience	62	37	2	61
b. Estimates within a single state				
Teacher-pupil ratio	24	4	17	79
Teacher education	34	0	9	91
Teacher experience	37	41	3	56

number of estimates and better statistical modelling (Summers and Wolfe, 1977). This impression is, however, very deceptive, and the mechanical tabulation approaches simply do not provide any effective overall quality assessment.[21]

A review of the available estimates clarifies how Krueger's tabulation of teacher-pupil ratio results differs: 17 of the 59 publications (29%) contained a single estimate of the effect of the teacher-pupil ratio — but these estimates are only 6% of the 277 total available estimates.[22] Krueger wants to increase the weight on these 17 estimates (publications) and commensurately decrease the weight on the remaining 260 estimates. Note, however, that over 40% of the single-estimate publications use state aggregate data, compared to only 4% of all estimates. Relatedly, the single-estimate publications are more likely to employ multistate estimates (which consistently ignore any systematic differences in state policies) than the publications with two or more estimates. Weighting by publications rather than separate estimates, as Krueger promotes, heavily weights low-quality estimates that suffer from the two major quality problems discussed above.

The implications of the different weighting schemes are perhaps easiest seen by noting the effect on the weights attached to his own estimates (Card and Krueger, 1992*a*, *b*). Each of these state-level analyses contributes one positive and statistically significant estimate of teacher-pupil ratios (although, as noted, Card and Krueger (1992*b*) should not be included because it lacks any measure of family

[21] A central motivation for Krueger (2002) is the assertion that publications containing more estimates will have smaller sample sizes and thus will typically have larger standard errors. Sample sizes do *not*, however, fall on average with the number of estimates. The median sample size for estimates in publications with just one is indistinguishable from that for publications with 8 or more estimates (Hanushek, 2002). The single estimate in Card and Krueger (1992*a*), for example, is based on *just 147 state aggregate data points*, placing its sample size at less than half used by the median of all of the estimates of the effects of teacher-pupil ratios. Moreover, if one took seriously that results should be weighted by sample size, such calculations can be easily done instead of relying on the imprecise weighting of the number of estimates in each publication.

[22] Note, to facilitate comparisons with Krueger (2003), this discussions includes the estimate of the effects of pupil-teacher ratios in Card and Krueger (1992*b*) that was excluded from the tabulations previously displayed; see footnote 14. This one estimate in fact has a huge impact on the Krueger calculations because it comes from a frequently cited publication with a single estimate.

background). On the basis of available estimates, these would represent 0.7% of the findings. This rises to 3.4% based on weighting by publications. But, on the basis of citations, these estimates go to a remarkable 17% of the weighted findings; see Hanushek (2002).

The explicit quality considerations made in the bottom panel of Table 5 in fact eliminate all of the publications and estimates Krueger identifies as being problematic (i.e. the nine publications with eight or more estimates) – although they are eliminated on grounds of statistical quality and not because they simply provided too many separate estimates of class size effects. That panel does include the Summers and Wolfe (1977) estimate, along with a number of other equally high quality analyses of student achievement. But, most importantly, it also eliminates the 11 highly problematic estimates that come from estimates of the effect of teacher-pupil ratios using state level analyses that ignore differences in the state policy environment. These latter estimates have a disproportionate impact on each of his tabulations even though they are arguably the poorest estimates of the effect of class size on student performance.

In sum, Krueger's re-analysis of the econometric evidence achieves different results by emphasising low-quality estimates. The low-quality estimates are demonstrably biased toward finding significant positive effects of class size reduction and of added spending. His discussion tries to suggest that one is caught on the horns of a dilemma: either weight heavily the estimates from the nine publications with the most estimates (as in the overall estimates of Table 3) or weight heavily the low-quality state aggregate estimates (as he favours). In reality, another option is available: weight *neither* heavily because both suffer from serious statistical problems, as shown in the bottom of Table 5. Instead, concentrate on the highest quality studies.

Remarkably, just re-weighting by the Krueger technique still provides weak support for the overall class size reduction policies that Krueger advocates. Most of the estimates, no matter how tabulated, are not statistically different from zero at conventional levels. Even weighting by publications instead of estimates, *three*-quarters of the estimates are insignificant or have the wrong sign, and barely more than half the results indicate a positive effect of smaller classes. Thus, even when heavily weighting low-quality estimates, he can only achieve his rhetorical purpose of emphasising that 'class size is systematically related to student performance' by giving equal weight to statistically insignificant and statistically significant results and discarding estimates where the sign of insignificant estimates is unavailable.

2.1.2. *Overall econometric specification*

A key issue in considering the results of the educational production function analyses is whether they provide the necessary guidance for policy purposes. Specifically, while they show a pattern of association, is it reasonable to infer that they identify causal relationships?

The issue is particularly important when put into the context of educational policy. Resource allocations are determined by a complicated series of political and behavioural choices by schools and parents. The character of these choices

could influence the estimates of the effectiveness of resources. Consider, for example, the result of systematically assigning school resources in a compensatory manner. If low achieving kids are given extra resources – say smaller classes, special remedial instruction, improved technology, and the like – there is an obvious identification problem. Issues of this kind suggest both care in interpretation of results and the possible necessity of alternative approaches.

Before continuing, however, it is important to be more precise about the nature and potential importance of these considerations. Funding responsibility for schools in the US tends on average to be roughly equally divided between states and localities with the federal government contributing only 7% of overall spending. Huge variation in funding levels and formulae nonetheless exists across states. In most state funding of schools in the US, the distribution of expenditure does not depend on the actual performance of individual students, but instead (inversely) on the wealth and income of the community. In models of achievement that include the relevant family background terms (such as education, income, or wealth), this distribution of state resources would simply increase the correlations among the exogenous variables but would not suggest any obvious simultaneity problems for the achievement models. In fact, while the compensatory nature of funding often motivates some concerns, even this correlation of background and resources is not clear. Much of the funding debate in the US has revolved around a concern that wealthier communities and parents can afford to spend more for schools, and in fact almost all state financing formula are designed to offset this tendency at least partially. Thus, the actual correlations of resources and family backgrounds often are not very high.[23]

At the individual student level, correlations with aggregate district resources through either formula allocations or community decisions are not a major cause of concern. The individual classroom allocations may, however, be a concern. For example, within a school, low achievers may be placed in smaller classes, suggesting the possibility of simultaneity bias. Any such problems should be largely ameliorated by value-added models, which consider the student's prior achievement directly. The only concern then becomes allocations made on the basis of unmeasured achievement influences that are unrelated to prior achievement.

Particularly in the area of class size analysis, a variety of approaches do go further in attempting to identify causal effects, and the results are quite varied. Hoxby (2000) used de-trended variations in the size of birth cohorts to identify exogenous changes in class size in small Connecticut towns. Changes in cohort sizes, coupled

[23] The distribution of state funds varies across the states, but one fairly common pattern is that major portions of state funds are distributed inversely to the property wealth of the community. Because community wealth includes the value of commercial and industrial property within a community, the correlation of community wealth with the incomes of local residents tends to be low and sometimes even negative.

with the lumpiness of classes in small school districts, can provide variations in class size that are unrelated to other factors.[24] Other estimates have also explicitly considered exogenous factors affecting class size within the context of instrumental variables estimators for the effects of class size (Akerhielm, 1995; Boozer and Rouse, 1995). Unfortunately, identification of truly exogenous determinants of class size, or resource allocations more generally, is sufficiently rare that other compromises in the data and modelling are frequently required. These coincidental compromises jeopardise the ability to obtain clean estimates of resource effects and may limit the generalisability of any findings. Rivkin *et al.* (2001), employing an approach similar in spirit to that used by Hoxby, make use of exogenous variations in class sizes within Texas schools across multiple cohorts of varying sizes.[25] They find some small class size effects, but the effects vary significantly across grades and specifications.

These alternative approaches yield inconsistent results both in terms of class size effects and in terms of the effects of alternative methodologies. The results in each of these analyses tend to be quite sensitive to estimation procedures and to model specification. Further, they are inconsistent in terms of statistical significance, grade pattern, and magnitude of any effects. As a group, the results are more likely to be statistically significant with the expected sign than those presented previously for all estimates, but the typical estimate (for statistically significant estimates) tends to be very small in magnitude (see below).

2.2. *International Econometric Evidence*

The evidence for countries other than the US is potentially important for a variety of reasons. Other countries have varying institutional structures, so different findings could help to identify the importance of organisation and overall incentives. Moreover, other countries frequently have much different levels of resources and exhibit larger variance in resource usage, offering the prospect of understanding better the importance of pure resource differences. For example, one explanation of the lack of relationship between resources and performance in the US is its schools there are generally operating in an area of severe diminishing marginal productivity, placing most on the 'flat of the curve'. Thus, by observing schools at very different levels of resources, it would be possible to distinguish between technological aspects of the production relationship and other possible interpretations of the evidence such as imprecise incentives for students and teachers.

[24] While pertaining directly to the international evidence below, in a related approach Angrist and Lavy (1999) note that Maimonides' Rule requires that Israeli classes cannot exceed forty students, so that, again, the lumpiness of classrooms may lead to large changes in class size when the numbers of students in a school approaches multiples of forty (and the preferred class size is greater than forty). They formulate a regression discontinuity approach to identify the effects of class size, but many of their estimates also use class size variation other than that generated by the discontinuities. Similarly, Case and Deaton (1999) concentrate on the impact of white decision making on black schools in South Africa (where endogeneity from compensatory policies is arguably less important). They conclude that smaller classes have an impact on student outcomes in that setting.

[25] The nature of this analysis is discussed further below in the section on teacher quality.

While the international evidence has been more limited, this situation is likely to be reversed profitably in the future. A key problem has been less available performance data for different countries, but this lack of information is being corrected. As student outcome data become more plentiful – allowing investigation of value added by teachers in schools in different environments, international evidence can be expected to grow in importance.

2.2.1. *Developing countries*

Existing analyses in less developed countries have shown a similar inconsistency of estimated resource effects as that found in the US. While these estimates typically come from special purpose analyses and are frequently not published in refereed journals, they do provide insights into resource use at very different levels of support. Table 6 provides evidence on resource effects from estimates completed by 1990.[26] Two facets of these data compared to the previous US data stand out: (i) in general, a minority of the available estimates suggests much confidence that the identified resources positively influence student performance; (ii) there is generally somewhat stronger support for these resource policies than that existing in US analyses. Thus, the data hint that the importance of resources may vary with the level of resources, a natural presumption. Nonetheless, the evidence is not conclusive that pure resource policies can be expected to have a significant effect on student outcomes.

2.2.2. *Developed countries*

The evidence on developed countries outside of the US is more difficult to compile. The review by Vignoles *et al.* (2000) points to a small number of analyses outside of the US and shows some variation them similar to that already reported among estimates elsewhere. Dustman *et al.* (2003) provide an additional set of estimates.

Table 6

Percentage Distribution of Estimated Expenditure Parameter Coefficients from 96 Educational Production Function Estimates: Developing Countries

Input	Number of estimates	Statistically Significant (%)		Statistically Insignificant (%)
		Positive	Negative	
Teacher/Pupil Ratio	30	27	27	46
Teacher Education	63	56	3	41
Teacher Experience	46	35	4	61
Teacher Salary	13	31	15	54
Expenditure/Pupil	12	50	0	50
Facilities	34	65	9	26

Source: Hanushek (1995).

[26] This compilation of results from Hanushek (1995) incorporates information from Fuller (1985), Harbison and Hanushek (1992), and a variety of studies during the 1980s.

One set of consistent estimates for the TIMSS data is presented in Hanushek and Luque (2003). They employ the data on variations in scores across schools within individual countries. The 17 countries with complete data for 9-year-olds and the 33 countries with complete data for 13-year-olds are weighted toward more developed countries but do include poor countries. As shown in Table 7, they find little evidence that any of the standard resource measures for schools are related to differences in mathematics scores within countries, although a majority of the class size results for the youngest age group do have the expected negative sign. An extension of the estimation considers the possibility of compensatory allocation of students to varying class sizes. Specifically, estimation for rural schools with a single classroom – where compensatory placement is not feasible – yields little change in the overall results.[27] The lack of significant resource effects when corrected for selection does differ from the findings of Angrist and Lavy (1999) and of Case and Deaton (1999), which find more significant resource effects in Israel and South Africa (see footnote 24 for details).

Moreover, there is no evidence in this consistent work that there are different effects of resources by income level of the country or by level of the resources. Thus, contrary to the conclusions of Heyneman and Loxley (1983), schools do not appear relatively more important for poorer countries.

Woessman (2000, 2001) looks at cross national differences in TIMSS mathematics and science scores and concludes that the institutional structure matters importantly for achievement. By pooling the individual student test scores across countries and estimating models that include both school and national characteristics, he finds suggestive evidence that the amount of competition from private schools and the amount of decentralisation of decision making to individuals schools have significant beneficial impacts, while union strength is detrimental and standard differences in resources across countries are not clearly related to student performance. The limited number of national observations for institutions nevertheless leaves some uncertainty about the estimates and calls for replication in other samples that permit, say, variations within individual countries in the key institutional features.

3. Project STAR and Experimental Data[28]

A different form of evidence – that from random assignment experiment – has recently been widely circulated in the debates about class size reduction. In assessing resource effects, concern about selection frequently remains, even in the instrumental approaches. Following the example of medicine, one large scale experimental investigation in the State of Tennessee in the mid-1980s (Project STAR) pursued the effectiveness of class size reductions. Random-assignment experiments in principle have considerable appeal. The underlying idea is that we can obtain valid evidence about the impact of a given well-defined treatment by

[27] An additional check analyses whether smaller classes in a given grade seem to be allocated on compensatory or elitist grounds and finds countries split on this. The impact of such considerations on the estimated effects is nonetheless minimal.

[28] For a more extensive discussion of Project STAR, see Hanushek (1999*a*, *b*).

Table 7

Distribution of Estimated Production Function Parameters across Countries and Age Groups, by Sign and Statistical Significance (10% level) Dependent variable: classroom average TIMSS mathematics score

	Age 9 population					Age 13 population				
	Negative		Positive			Negative		Positive		
	Significant	Not significant	Not significant	Significant	Number of countries	Significant	Not significant	Not significant	Significant	Number of countries
Class Size	*3*	11	2	0	17	*2*	8	6	17	33
Teacher with at least a bachelor's degree	0	3	12	*0*	15	2	11	12	*2*	32
Teacher with special training	0	7	4	*1*	12	0	12	11	*2*	25
Teacher experience	0	7	6	*4*	17	3	9	17	*4*	33

Note: Number of statistically significant results with the expected sign of the effect shown in italics. Because these estimates rely on actual class size, the expected sign is negative (and not reversed as for teacher-pupil ratios in the prior tables).
Source: Hanushek and Luque (2003).

randomly assigning subjects to treatment and control groups, eliminating the possible contaminating effects of other factors and permitting conceptually cleaner analysis of the outcomes of interest across these groups. With observations derived from natural variations in individual selection, one must be able to distinguish between the treatment and other differences that might directly affect the observed outcomes and that might be related to whether or not they receive the treatment. Randomisation seeks to eliminate any relationship between selection into a treatment programme and other factors that might affect outcomes. (See, however, the caution provided in Todd and Wolpin (2003)).

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. Schools were solicited for participation, with the stipulation that any school participating must be large enough to have at least one class in each treatment group. The initial sample included 6,324 kindergarten pupils, split between 1,900 in small classes and 4,424 in regular classes. (After the first year, the two separate regular class treatments were effectively combined, because there were no perceived differences in student performance. The result about the ineffectiveness of classroom aides has received virtually no attention.) The initial sample included 79 schools, although this subsequently fell to 75. The initial 326 teachers grew slightly to reflect the increased sample size in subsequent grades, although of course most teachers are new to the experiment at each new grade.

The results of the Project STAR experiment have been widely publicised. The simplest summary is that:

1. Pupils in small classes perform better than those in regular classes or regular classes with aides starting in kindergarten;
2. The kindergarten performance advantage of small classes widens a small amount in first grade but then either remains quantitatively the same (reading) or narrows (mathematics) by third grade; and,
3. Taking each grade separately, the difference in performance between small and regular classes is statistically significant.

This summary reflects the typical reporting, focusing on the differences in performance at each grade and concluding that small classes are better than large, e.g., Finn and Achilles (1990); Mosteller (1995). But, it ignores the fact that under the common conceptual discussions one would expect the differences in performance to become wider through the grades because they continue to get more resources (smaller classes) and that should keep adding an advantage. This issue was first raised by Prais (1996), who framed the discussion in terms of the value-added. As Krueger (1999) demonstrates, the small class advantage is almost exclusively obtained in the first year of being in a small class – suggesting that the advantages of small classes are not generalisable to any other grades.

Importantly, this pattern of effects is at odds with the normal rhetoric about smaller classes permitting more individualised instruction, allowing improved class room interactions, cutting down on disruptions, and the like. If these were the important changes, small classes should confer continuing benefits in any grades

where they are employed. Instead, the results appear more consistent with socialisation or introduction into the behaviour of the classroom – one time effects that imply more general class size reduction policies across different grades will not be effective – or with simple problems in the randomisation and implementation of the experiment.

The actual gains in performance from the experimental reduction in class size were relatively small (less than 0.2 standard deviations of test performance), especially when the gains are compared to the magnitude of the class size reduction (around 8 students per class). Thus, even if Project STAR is taken at face value, it has relatively limited policy implications.

While the experimental approach has great appeal, the actual implementation in the case of Project STAR introduces considerable uncertainty into these estimates (Hanushek, 1999*b*). The uncertainty arises most importantly from questions about the quality of the randomisation over time. In each year of the experiment, there was sizable attrition from the prior year's treatment groups, and these students were replaced with new students. Of the initial experimental group starting in kindergarten, 48% remained in the experiment for the entire four years. No information, such as pretest scores before entry to the experiment, is available to assess the quality of student randomisation for the initial experimental sample or for the subsequent additions to it. Throughout the four years of the experiment there was substantial and nonrandom treatment group crossover (about 10% of the small class treatment group in grades 1–3). There is also substantial, nonrandom test taking over the years of the experiment, exceeding 10% on some tests. Most important, the results depend fundamentally on the choice of teachers. While the teachers were to be randomly assigned to treatment groups, there is little description of how this was done. Nor is it easy to provide any reliable analysis of the teacher assignment, because only a few descriptors of teachers are found in the data and because there is little reason to believe that they adequately measure differences in teacher quality.[29] The schools themselves were self-selected and are clearly not random. Small schools were precluded from the study, as were those schools that were unwilling to provide their own partial funding to cover the full costs. (This issue is also important, because the STAR experiment heavily oversampled urban and minority schools where the response to the programme is thought to be largest.)[30] The net result of each of these effects is difficult to ascertain, but there is prima facie evidence that the total impact is to overstate the impact of reduced class size (Hanushek, 1999*b*). Hoxby (2000) further points out that because teachers and administrators knew they were participating in an experiment that could have significant implications for future resources, their behaviour in the experiment could be affected.

[29] The teacher data include race, gender, teaching experience, highest degree, and position on the Tennessee career ladder. While there is no information about the effect of career ladder position on student performance, as summarised above, none of the other measures have been found to be reliable indicators of quality. For estimates of the magnitude of variation in teacher quality, see below.

[30] Krueger (1999) identifies significantly stronger effects for disadvantaged students, which will then be overweighted in calculating programme average treatment effects.

The importance of the methodology does deserve emphasis. Because of questions about effectiveness and causality in the analysis of schools, further use of random assignment experimentation would have high value. As Todd and Wolpin (2003) point out, random assignment experiments do not answer all of the policy questions. Nonetheless, it would seem natural to develop a range of experiments that could begin to provide information about what kinds of generalisations can be made.

The one limited and flawed experiment in Tennessee cannot be taken as providing the definitive evidence needed for policy changes that cost billions of dollars annually. At best it provides evidence about the potential impact of very large changes in class size applied to kindergarten students, and there is direct evidence that these findings do not generalise to other grades and other situations.

4. Interpreting the Resource Evidence

A wide range of analyses indicate that overall resource policies have not led to discernible improvements in student performance. It is important to understand what is and is not implied by this conclusion. First, it does not mean that money and resources *never* matter. There clearly are situations where small classes or added resources have an impact. It is just that no good description of when and where these situations occur is available, so that broad resource policies such as those legislated from central governments may hit some good uses but also hit bad uses that generally lead to offsetting outcomes. Second, this statement does not mean that money and resources *cannot* matter. Instead, as described below, altered sets of incentives could dramatically improve the use of resources.

The evidence on resources is remarkably consistent across countries, both developed and developing. Had there been distinctly different results for some subsets of countries, issues of what kinds of generalisations were possible would naturally arise. Such conflicts do not appear particularly important.

There is a tendency by researchers and policy makers to take a single study and to generalise broadly from it. By finding an analysis that suggests a significant relationship between a specific resource and student performance, they conclude that, while other resource usage might not be productive, the usage that is identified would be, e.g, Grissmer *et al.* (2000). If this is so, it leads to a number of important questions. Why is that schools have failed to employ such a policy? Is it just that they do not have the information that the researcher has? That of course seems unlikely since schools in fact constantly experiment with a variety of approaches and resource patterns. Alternatively, consistent with the discussion below, it seems more likely that schools have limited incentives to seek out and to employ programmes that consistently relate to student achievement.

It is just this tendency to overgeneralise from limited evidence that lies behind the search for multiple sources of evidence on the effectiveness of different resource usage. That broader body of evidence provides little support for the input policies that continue to be the most common approach to decision making.

5. Teacher Quality

Starting with the Coleman Report on *Equality of Educational Opportunity* (Coleman *et al.*, 1966), many have argued that schools do not matter and that only families and peers affect performance. Unfortunately, that report and subsequent interpretations of it have generally confused 'measurability' with true effects. Specifically, as described above for more recent work, characteristics of schools and classrooms were not closely related to student performance – leading to the conclusion that schools do not matter. This conclusion not only led to the extensive subsequent research but also probably more than anything else to a prevailing view that differences among schools are not very important.

The extensive research over the past 35 years has made it clear that there are very important differences among teachers and, by implication, schools. This finding, of course, does not surprise many parents who are well aware of quality differences of teachers but it eluded many researchers.

The simple definition of teacher quality used here is an output based measure that focuses on student performance, instead of the more typical input measures based on characteristics of the teacher and school. High quality teachers are ones who consistently obtain higher than expected gains in student performance, while low quality teachers are ones who consistently obtain lower than expected gains. Using that definition, variations in teacher quality can be obtained by estimating fixed effects models of student performance after conditioning on entering student performance and other factors that affect achievement gains. When this approach has been used in studying US schools, large variations in performance have been uncovered, e.g., Hanushek (1971, 1992); Murnane (1975); Murnane and Phillips (1981); Armor *et al.* (1976); Rivkin *et al.* (2001).[31] The only related study internationally pertains to rural Brazil, where similarly large differences among teachers are found (Harbison and Hanushek, 1992).

The magnitude of differences in teacher quality is impressive. Looking at the range of quality for teachers within a single large urban district, 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 (Hanushek, 1992).[32] That is, a good teacher will get a gain of 1½ grade level equivalents while a bad teacher will get ½ year for a single academic year.

A second set of estimates comes from recent work on students in Texas (Rivkin *et al.*, 2001). The analysis follows several entire cohorts of students and permits

[31] In the general fixed effect formulation, identification and interpretation of teacher and school effects is nonetheless complicated. For example, teacher effects, school effects and classroom peer effects are 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 *et al.* (2001) remove separate school and grade fixed effects and observe the consistency of teacher effects across different cohorts – thus isolating the impact of teachers.

[32] 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.

multiple observations of different classes with a given teacher. We look at just the variations in performance from differences in teacher quality within a typical school and do not consider any variations across schools, making them very much a lower bound on teacher effects. 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. (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 which all tend to lead to understatement of the systematic teacher differences.)

A third indication of magnitude is found in the Project STAR results. The average difference in performance of students in small kindergartens has been the focus of all attention, 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.

The teacher differences estimated in Texas are huge compared to any of the estimates for measured teacher and school attributes. For example, a one standard deviation reduction in class size implies a 0.01–0.03 standard deviation improvement in student achievement (Rivkin *et al.*, 2001). The lower bound estimate on teacher quality summarised implies a one standard deviation change in quality leads to a 0.11 standard deviation increase in achievement. The fact that even the lower bound estimate of teacher quality effects dwarfs either class size or experience effects should give policy makers pause.

These estimates of teacher quality can also be related to the popular argument that family background is overwhelmingly important and that schools cannot be expected to make up for bad preparation from home. The latter estimates of teacher performance suggest that having five years of good teachers in a row (one standard deviation above average, or at the 85th quality percentile) would overcome the average achievement deficit between low income kids (those on free or reduced price lunch) and others from higher income families. In other words, high quality teachers can make up for the typical deficits that we see in the preparation of kids from disadvantaged backgrounds.

We do not tend to observe these deficits disappearing, however, because the current school system does not ensure any streaks of such high quality teachers – particularly for disadvantaged students. In fact, it is currently as likely that the typical student gets a run of bad teachers – with the symmetric achievement losses – as a run of good teachers.

6. Policy Alternatives

Much of economic analysis is built on a presumption that higher expenditure yields better outcomes. Thus, many people are surprised to find evidence that school resources are not closely related to student performance. Indeed, a variety of

mechanisms might conceptually push schools toward better resource use. Parents undoubtedly care about the performance of their children. Democratic political pressures might force responsive government actions. Household locational decisions allow families some latitude to select schools that are performing well.

But this is not a frictionless market with knowledgeable consumers making decisions with perfect information. We are considering government provision of a service whose quality is difficult to judge. Moving one's residence or forcing better governmental performance in a specific service area is difficult and expensive. Moreover, it is frequently difficult to separate the quality of the school from the quality of the students. We know that parents as well as students exert a powerful influence on student achievement. In order to make quality judgments, it is necessary to separate the school from the nonschool influences. Parents can generally tell the differences among the different teachers within a given school, but comparing the average quality of teachers in one school to those in another is a more difficult task. It is especially difficult because residential and school choice decisions frequently involve an element of sorting along socio-economic lines and, on average, along lines of student preparation. Conventionally defined 'good schools' are often schools with the best-prepared students going into them but not necessarily ones where the value-added of the school is particularly high and *vice versa* for 'bad schools'.

The clearest contrast in policy perspectives is between input policy and output, or incentive, policies. In the US and elsewhere, for example, a very popular recent policy is funding or mandating smaller class sizes. But, as the evidence indicates, this is an expensive and generally unproductive policy.

In recognition of the importance of quality teachers, a variety of recommendations and policy initiatives have been introduced. Unfortunately, the currently most popular proposals in the US are likely to lower teacher quality rather than improve it. The idea that has been picked up by US policy makers at all levels is to increase the requirements to become a teacher. The notion is simple: if we can insist on better prepared and more able teachers, teacher quality will necessarily rise, and student performance will respond. This argument – at least as implemented – proves as incorrect as it is simple. The range of options being pushed forward include raising the course work requirement for teacher certification, testing teachers on either general or specific knowledge, requiring specific kinds of undergraduate degrees, and requiring master's degrees. Each of these has surface plausibility, but little evidence exists to suggest that these are strongly related to teacher quality and to student achievement.

More pernicious, these input requirements almost certainly act to reduce the supply of teachers. In other words, while the proposed requirements do little or nothing to ensure high quality teachers, they do cut down on the group of people who might enter teaching. Teacher certification requirements are generally advertised as making sure that there is some minimum floor on quality, but, if the requirements end up keeping out high quality teachers who do not want to take the specific courses required, they instead act more like a ceiling on quality.

A related policy proposal is to raise teacher salaries in order to compensate for any costs of added preparation or simply to attract a larger group of higher quality

people into teaching. By itself, however, such an undifferentiated pay policy would do little to ensure that the quality of teaching would improve, at least for a long time. Current teachers, both good and bad, would be encouraged not to leave teaching, and any specific shortages such as of high quality teachers or of teachers in more technical areas would not be relieved unless these salaries could be targeted.

The alternative set of potential policies emphasises performance incentives. Few employees of US public schools find that their jobs are at all dependent on the performance of students. Pay, promotion, retention in a job, and the like appear to be little different for high quality teachers and low quality teachers. Similarly jobs for school principals or other administrative and support personnel do not seem closely related to any student outcomes. A simple idea that pervades economics is that incentives have powerful effects. In the case of schools few incentives relate to the object of interest – student performance. Thus, it should not be particularly surprising if added resources do not translate into better performance, because there is little feedback from performance.

Much of the larger debate about school policy actually revolves around proposed changes in the structure of school incentives. The range of incentive policies currently under debate fall into three generic types.[33] First, merit pay for teachers – such as that recently introduced into British schools – or rewards to entire schools imply moving to a direct pay-for-performance relationship. Second, privatisation or contracting arrangements involve hiring private firms to provide given academic or nonacademic functions with their rewards based upon outcomes. Finally, expanded choice of schools by students relies on the underlying idea that schools that do well will attract more students and those that do poorly will lose students and that this mechanism will provide incentives to improve student performance (Friedman, 1962). (Choice actually comes in different forms identified chiefly by whether or not private schools can compete with public schools.)

Each of these generic approaches has considerable appeal compared to the current system. Each focuses attention on what is desired, instead of trying to guess at a set of inputs that will lead to the desired result. Contrary to the current structure, the general outline of each of the incentive structures makes economic sense.

Designing good incentives, however, is not easy. For example, voucher opponents point to a variety of issues including the prospect of further racial and economic segregation in schools; the chance that schools offer undesirable courses of study; and the possibility that the competition does not have much impact on public schools. Similarly, merit pay opponents argue that there is little research supporting positive outcomes (Cohen and Murnane, 1986); that its award is likely to be too subjective and political; and that individual rewards lead to undesirable competition among teachers.

In most cases, it would be possible to design incentive schemes that circumvent the largest problems, at least if the problems are anticipated. Unfortunately, incentive contracts can be very complicated, and some of the reactions to the specified incentives might be surprising. For example, an early experiment with

[33] An expanded discussion of such incentives can be found in Hanushek *et al.* (1994).

performance contracting (that involved hiring private firms to teach basic subjects to disadvantaged students and paying the firms based on results) failed to yield much information because of fundamental flaws in the incentive contract (Gramlich and Koshel, 1975).[34] In other cases, such as the limited exploration of the use of vouchers in Milwaukee, Wisconsin, a single highly constrained scheme was employed and even then considerable controversy over the outcomes continues (Peterson *et al.*, 1996; Greene *et al.*, 1998; Rouse, 1998; Witte, 1999).[35]

In sum, there is ample evidence that the currently employed input policies are failures. No one, of course, argues that they wish to pursue the failed policies of the past. Typically, proponents of input policies either argue they have new ideas that are not repeating the mistakes of the past policies. However, these new policies, like those in the past, are seldom based on evidence of superiority. An alternative form used is to say that 'money spent wisely will yield favourable outcomes,' but this is tautological.

At the same time the generic incentive approaches, with the exception of certain specific kinds of merit pay schemes, have not been tried very often, so there is little experience with developing good contracts. Moreover, no systematic approach to developing information about incentives has been employed. Therefore, the superiority of using performance incentives instead of relying on just input policies remains largely untapped.

A significant issue in the discussion of incentives is the slowness with which evidence accumulates. There is not a strong scientific evaluation tradition within education. Further, because education is such a potent political issue, there are constant pressures to go immediately to new universal policies without worrying too much about the evidence supporting them. This has two features. First, many mistakes are made (leading to the results described previously). Second, no new evidence accumulates to aid in making future decisions.

The State of California provides an informative if discouraging case study. In 1997, the State provided financial incentives to school districts to reduce class size. This politically popular programme, offered to all districts simultaneously, defies any evaluation because no baseline performance data are available and because all districts received the same treatment. It continues with appropriations of $1.5 billion annually with no information about its effectiveness.

If educational policies are to be improved, much more serious attention must be given to developing solid evidence about what things work and what things do not. Developing such evidence means that regular high quality information about student outcomes must be generated. In particular, it must be possible to infer the value-added of schools. Improvement also would be advanced significantly by the introduction and general use of random assignment experiments and other

[34] The experimental contract did not offer firms a fair chance to make a profit, provided no payment to the firms if achievement gains were below the national average, and capped the maximum reward. These provisions provided poor incentives and led firms to do a variety of educationally inappropriate things.

[35] Assessment controversies have arisen over the length of time before achievement gains should be expected, over the appropriate comparison groups, and over the costs of private schooling.

well-defined evaluation methods. Without incentives and without adequate evaluation, there should be no expectation that schools improve, regardless of the resources added to the current structure.

Stanford University and NBER

References

Akerhielm, K. (1995). 'Does class size matter?', *Economics of Education Review*, vol. 14 (3) (September), pp. 229–41.

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

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.

Baumol, W.J. (1967). 'Macroeconomics of unbalanced growth: the anatomy of urban crisis', *American Economic Review*, vol. 57 (3) (June), pp. 415–26.

Behrman, J.R., Kletzer, L.G., McPherson, M.S. and Schapiro, M.O. (1998). 'The microeconomics of college choice, careers, and wages: measuring the impact of higher education', *Annals of the American Academy of Political and Social Science*, vol. 559 (September), pp. 12–23.

Betts, J.R. (1996). 'Is there a link between school inputs and earnings? Fresh scrutiny of an old literature', in (G. Burtless, ed.) *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success*, pp. 141–191. Washington, DC: Brookings.

Bishop, J. (1989). 'Is the test score decline responsible for the productivity growth decline?', *American Economic Review*, vol. 79 (1) pp. 178–97.

Bishop, J. (1991). 'Achievement, test scores, and relative wages', in (M.H. Kosters, ed.) *Workers and Their wages*, pp. 146–86. Washington, DC: The AEI Press.

Bishop, J. (1992). 'The impact of academic competencies of wages, unemployment, and job performance', *Carnegie-Rochester Conference Series on Public Policy*, vol. 37 (December), pp. 127–94.

Blackburn, M.L. and Neumark, D. (1993). 'Omitted-ability bias and the increase in the return to schooling', *Journal of Labor Economics*, vol. 11 (3) (July), pp. 521–44.

Blackburn, M.L., and Neumark, D. (1995). 'Are OLS estimates of the return to schooling biased downward? Another look', *Review of Economics and Statistics*, vol. 77 (2) (May), pp. 217–30.

Boissiere, M.X., Knight, J.B., and Sabot, R.H. (1985). 'Earnings, schooling, ability, and cognitive skills', *American Economic Review*, vol. 75 (5), pp. 1016–30.

Boozer, M.A., and Rouse, C. (1995). 'Intraschool variation in class size: patterns and implications', NBER Working Paper 5144, June.

Burtless, G. (1996). *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success*. Washington, DC: Brookings.

Card, D. and Krueger, A.B. (1992*a*). Does school quality matter? Returns to education and the characteristics of public schools in the United States', *Journal of Political Economy*, vol. 100 (1) (February), pp. 1–40.

Card, D. and Krueger, A.B. (1992*b*). 'School quality and black-white relative earnings: a direct assessment', *Quarterly Journal of Economics*, vol. 107 (1) (February), pp. 151–200.

Case, A. and Deaton, A. (1999). 'School inputs and educational outcomes in South Africa', *Quarterly Journal of Economics*, vol. 114 (3) (August), pp. 1047–84.

Cawley, J., Heckman, J.J., Lochner, L. and Vytlacil, E. (2000). 'Understanding the role of cognitive ability in accounting for the recent rise in the economic return to education', in (K. Arrow, S. Bowles and S. Durlauf, eds.) *Meritocracy and Economic Inequality*, pp. 230–65. Princeton, NJ: Princeton University Press.

Chaikind, S., Danielson, L.C. and Brauen, M.L. (1993). 'What do we know about the costs of special education? A selected review', *Journal of Special Education*, vol. 26 (4), pp. 344–70.

Cohen, D.K. and Murnane, R.J. (1986). 'Merit pay and the evaluation problem: understanding why most merit pay plans fail and a few survive', *Harvard Educational Review*, vol. 56 (1) (February), pp. 1–17.

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, D.C.: U.S. Government Printing Office.

Congressional Budget Office. (1986). *Trends in Educational Achievement*. Washington, D.C.: Congressional Budget Office.

Currie, J. and Thomas, D. (2000). 'Early test scores, socioeconomic status, school quality, and future outcomes', Department of Economics, UCLA (mimeo) (September).

Dugan, D.J. (1976). 'Scholastic achievement: its determinants and effects in the education industry', in (J.T. Froomkin, D.T. Jamison and R. Radner, eds.) *Education as an Industry*, pp. 53–83. Cambridge, MA: Ballinger.

Dustmann, C., Rajah, N. and van Soest, A. (2003). 'Class size, education and wages', Economic Journal, vol. 113, pp. F99–120.

Finn, J.D. and Achilles, C.M. (1990). 'Answers and Questions about class size: a statewide experiment', *American Educational Research Journal*, vol. 27 (3) (Fall), pp. 557–77.

Friedman, M. (1962). *Capitalism and Freedom*, Chicago: University of Chicago Press.

Fuller, B. (1985). *Raising School Quality in Developing Countries: What Investments Boost Learning?* Washington, DC: The World Bank.

Gramlich, E.M. and Koshel, P.P. (1975). *Educational Performance Contracting*. Washington, D.C.: The Brookings Institution.

Greenberg, D. and McCall, J. (1974). 'Teacher mobility and allocation', *Journal of Human Resources*, vol. 9(4) (Fall), pp. 480–502.

Greene, J.P., Peterson, P.E. and Du, J. (1998). 'School choice in Milwaukee: a randomized experiment', in (P.E. Peterson and B.C. Hassel, eds.) *Learning from School Choice*, pp. 335–56, Washington, DC: Brookings Institution.

Greenwald, R., Hedges, L.V. and Laine, R.D. (1996). 'The effect of school resources on student achievement', *Review of Educational Research*, vol. 66 (3) (Fall), pp. 361–96.

Griliches, Z. (1974). 'Errors in variables and other unobservables', *Econometrica*, vol. 42 (6) (November), pp. 971–98.

Grissmer, D.W., Flanagan, A., Kawata, J. and Williamson, S. (2000). *Improving Student Achievement: What NAEP State Test Scores Tell Us*, Santa Monica, CA: Rand Corporation.

Grissmer, D.W., Kirby, S.N., Berends, M. and Williamson, S. (1994). *Student Achievement and the Changing American Family*, Santa Monica, CA: Rand Corporation.

Grogger, J.T. and Eide, E. (1993). 'Changes in college skills and the rise in the college wage premium', *Journal of Human Resources*, vol. 30 (2) (Spring), pp. 280–310.

Gundlach, E., Woessmann, L. and Gmelin, J. (2001). 'The decline of schooling productivity in OECD countries', Economic Journal, vol. 111 (May), pp. C135–47.

Hanushek, E.A. (1971). 'Teacher characteristics and gains in student achievement: Estimation using micro data', *American Economic Review*, vol. 60 (2) (May), pp. 280–8.

Hanushek, E.A. (1979). 'Conceptual and empirical issues in the estimation of educational production functions', *Journal of Human Resources*, vol. 14 (3) (Summer), pp. 351–88.

Hanushek, E.A. (1986). 'The economics of schooling: production and efficiency in public schools', *Journal of Economic Literature*, vol. 24 (3) (September), pp. 1141–77.

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

Hanushek, E.A. (1995). 'Interpreting recent research on schooling in developing countries', *World Bank Research Observer*, vol. 10 (2) (August), pp. 227–46.

Hanushek, E.A. (1996). 'A more complete picture of school resource policies', *Review of Educational Research*, vol. 66 (3) (Fall), pp. 397–409.

Hanushek, E.A. (1997*a*). 'Assessing the effects of school resources on student performance: an update', *Educational Evaluation and Policy Analysis*, vol. 19 (2) (Summer), pp. 141–64.

Hanushek, E.A. (1997*b*). 'The productivity collapse in schools', in (W.J. Fowler, Jr., ed.) *Developments in School Finance, 1996*, pp. 185–95, Washington, DC: National Center for Education Statistics.

Hanushek, E.A. (1999*a*). 'The evidence on class size', in (S.E. Mayer and P.E. Peterson, eds.) *Earning and Learning: How Schools Matter*, pp. 131–68. Washington, DC: Brookings Institution.

Hanushek, E.A. (1999*b*). 'Some findings from an independent investigation of the Tennessee STAR experiment and from other investigations of class size effects', *Educational Evaluation and Policy Analysis*, vol. 21 (2) (Summer), pp. 143–63.

Hanushek, E.A. (2002). 'Evidence, politics, and the class size debate', in (L. Mishel and R. Rothstein, eds.) *The Class Size Debate*, pp. 37–65. Washington, DC: Economic Policy Institute.

Hanushek, E.A., Kain, J.F. and Rivkin, S.G. (2001*a*). 'Disruption versus Tiebout improvement: the costs and benefits of switching schools', WP 8479, National Bureau of Economic Research (September).

Hanushek, E.A., Kain, J.F. and Rivkin, S.G. (2001*b*). 'Why public schools lose teachers', WP 8599, National Bureau of Economic Research (November).

Hanushek, E.A., Kain, J.F. and Rivkin, S.G. (2002). 'Inferring program effects for specialized populations: does special education raise achievement for students with disabilities?' *Review of Economics and Statistics*, vol. 84 (4) (November), pp. 584–99.

Hanushek, E.A. and Kim, D. (1995). 'Schooling, labor force quality, and economic growth', Working Paper 5399, National Bureau of Economic Research (December).

Hanushek, E.A. and Kimko, D.D. (2000). 'Schooling, labor force quality, and the growth of nations', *American Economic Review*, vol. 90 (5) (December), pp. 1184–208.

Hanushek, E.A. and Luque, J.A. (2003). 'Efficiency and equity in schools around the world', *Economics of Education Review*, vol. 22 (4) (August), forthcoming.
Hanushek, E.A. and others, (1994). *Making Schools Work: Improving Performance and Controlling Costs*, Washington, DC: Brookings Institution.
Hanushek, E.A. and Pace, R.R. (1995). 'Who chooses to teach (and why)?', *Economics of Education Review*, vol. 14 (2) (June), pp. 101–17.
Hanushek, E.A. and Rivkin, S.G. (1997). 'Understanding the twentieth-century growth in U.S. school spending', *Journal of Human Resources*, vol. 32 (1) (Winter), pp. 35–68.
Hanushek, E.A., Rivkin, S.G. and Taylor, L.L. (1996). 'Aggregation and the estimated effects of school resources', *Review of Economics and Statistics*, vol. 78 (4) (November), pp. 611–27.
Harbison, R.W. and Hanushek, E.A. (1992). *Educational Performance of the Poor: Lessons from Rural Northeast Brazil*, New York: Oxford University Press.
Heckman, J.J., Layne-Farrar, A. and Todd, P. (1996*a*). 'Does measured school quality really matter? An examination of the earnings-quality relationship', in (G. Burtless, ed.) *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success*, pp. 192–289. Washington, DC: Brookings.
Heckman, J.J., Layne-Farrar, A. and Todd, P. (1996*b*). 'Human capital pricing equations with an application to estimating the effect of schooling quality on earnings', *Review of Economics and Statistics*, vol. 78(4) (November), pp. 562–610.
Heckman, J.J. and Vytlacil, E. (2001). 'Identifying the role of cognitive ability in explaining the level of and change in the return to schooling', *Review of Economics and Statistics*, vol. 83 (1) (February), pp. 1–12.
Heyneman, S.P. and Loxley, W. (1983). 'The effect of primary school quality on academic achievement across twenty-nine high and low income countries', *American Journal of Sociology*, vol. 88 (May), pp. 1162–94.
Hoxby, C.M. (2000). 'The effects of class size on student achievement: new evidence from population variation', *Quarterly Journal of Economics*, vol. 115 (3) (November), pp. 1239–85.
Krueger, A.B. (1999). 'Experimental estimates of education production functions', *Quarterly Journal of Economics*, vol. 114 (2) (May), pp. 497–532.
Krueger, A.B. (2002). 'Understanding the magnitude and effect of class size on student achievement', in (L. Mishel and R. Rothstein, eds.) *The Class Size Debate*, pp. 7–35. Washington, DC: Economic Policy Institute.
Krueger, A.B. (2003). 'Economic considerations and class size', Economic Journal, vol. 113, pp. F34–63.
Manski, C.F. and Wise, D.A. (1983). *College Choice in America*, Cambridge: Harvard University Press.
Mayer S.E. (1997). What Money Can't Buy: Family Income and Children's Life Chances, Cambridge, MA: Harvard University Press.
Mosteller, F. (1995). 'The Tennessee study of class size in the early school grades', *The Future of Children*, vol. 5 (2) (Summer/Fall), pp. 113–27.
Murnane, R.J. (1975). *Impact of School Resources on the Learning of Inner City Children*, Cambridge, MA: Ballinger.
Murnane, R.J. (1981). 'Teacher mobility revisited', *Journal of Human Resources*, vol. 16 (1) (Winter), pp. 3–19.
Murnane, R.J. and Phillips, B. (1981). 'What do effective teachers of inner-city children have in common?', *Social Science Research*, vol. 10 (1) (March), pp. 83–100.
Murnane, R.J., Willett, J.B., Braatz, M.J. and Duhaldeborde, Y. (2001). 'Do different dimensions of male high school students' skills predict labor market success a decade later? Evidence from the NLSY', *Economics of Education Review*, vol. 20 (4) (August), pp. 311–20.
Murnane, R.J., Willett, J.B., Duhaldeborde, Y. and Tyler, J.H. (2000). 'How important are the cognitive skills of teenagers in predicting subsequent earnings?', *Journal of Policy Analysis and Management*, vol. 19 (4) (Fall), pp. 547–68.
Murnane, R.J., Willett, J.B. and Levy, F. (1995). 'The growing importance of cognitive skills in wage determination', *Review of Economics and Statistics*, vol. 77 (2) (May), pp. 251–66.
Neal, D.A. and Johnson, W.R. (1996). 'The role of pre-market factors in black-white differences', *Journal of Political Economy*, vol. 104 (5) (October), pp. 869–95.
O'Neill, J. (1990). 'The role of human capital in earnings differences between black and white men', *Journal of Economic Perspectives*, vol. 4 (4) (Fall), pp. 25–46.
OECD. (2001). *Education at a Glance*, Paris: OECD.
Peterson, P.E., Greene, J.P. and Noyes, C. (1996). 'School choice in Milwaukee', *Public Interest*, vol. 125 (Fall), pp. 38–56.
Pierce, B. and Welch, F. (1996). 'Changes in the structure of wages', in (E.A. Hanushek and D.W. Jorgenson, eds.) *Improving America's Schools: The Role of Incentives*, pp. 53–73. Washington, DC: National Academy Press.

Prais, S.J. (1996). 'Class-size and learning: the Tennessee experiment – what follows?', *Oxford Review of Education*, vol. 22 (4) pp. 399–414.
Psacharopoulos, G. (1989). 'Time trends of the returns to education: cross-national evidence', *Economics of Education Review*, vol. 8 (3), pp. 225–31.
Psacharopoulos, G. (1994). 'Returns to investment in education: a global update', *World Development*, vol. 22, pp. 1325–44.
Rivkin, S.G. (1995). 'Black/white differences in schooling and employment', *Journal of Human Resources*, vol. 30 (4) (Fall), pp. 826–52.
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).
Rouse, C.E. (1998). 'Private school vouchers and student achievement: an evaluation of the Milwaukee Parental Choice Program', *Quarterly Journal of Economics*, vol. 113 (2), (May), pp. 553–602.
Summers, A.A. and Wolfe, B.L. (1977). 'Do schools make a difference?', *American Economic Review*, vol. 67 (4) (September), pp. 639–52.
Todd, P.E. and Wolpin, K.I. (2003). 'On the specification and estimation of the production function for cognitive achievement', Economic Journal, vol. 113, pp. F3–33.
U.S. Department of Education. (2002). *Digest of Education Statistics, 2001*, Washington, DC: National Center for Education Statistics.
Vignoles A., Levacic, R., Walker, J., Machin, S. and Reynolds, D. (2000). 'The relationship between resource allocation and pupil attainment: a review', DPUQ, Centre for the Economics of Education, LSE (November).
Wirtz, W. (1977). *On Further Examination: Report of the Advisory Panel and the Scholastic Aptitude Test Score Decline*, NY: College Entrance Examination Board.
Witte, J.F., Jr. (1999). *The Market Approach to Education*, Princeton, NJ: Princeton University Press.
Woessmann, L. (2000). 'Schooling resources, educational institutions, and student performance: the international evidence', Kiel Working Paper No. 983, Kiel Institute of World Economics, Kiel, (December).
Woessmann, L. (2001). 'Why students in some countries do better', *Education Matters*, vol. 1 (2) (Summer), pp. 67–74.