Chapter 14

SCHOOL RESOURCES

ERIC A. HANUSHEK

Stanford University,

National Bureau of Economic Research

Contents

Handbook of the Economics of Education, Volume 2
Edited by Eric A. Hanushek and Finis Welch

DOI: 10.1016/S1574-0692(06)02014-9

Abstract

Although there is intense policy interest in improving educational outcomes around the world, there is much greater uncertainty about how to accomplish this. The primary governmental decisions often relate to the resources that are devoted to schooling, but the research indicates little consistent relationship between resources to schools and student achievement. Much of the research considers how resources affect student achievement as measured by standardized test scores. These scores are strongly related to individual incomes and to national economic performance, making them a good proxy for longer run economic impacts. But, the evidence – whether from aggregate school outcomes, econometric investigations, or a variety of experimental or quasiexperimental approaches – suggests that pure resource policies that do not change incentives are unlikely to be effective. Importantly, the results appear similar across both developed and developing countries.

Keywords

school resources, class size, achievement, experimental evidence, economic growth, incomes

Introduction

Perhaps no issue in the economics of education is as contentious as debates about the role and impact of school resources. Governments, legislatures, and at times courts routinely decide on the amount of money and resources to go to schools. Parents in part assess the quality of schools by the resources available. Taxpayers consider how resources relate to school performance and to other uses of those funds. As such, research into the effectiveness of various school resources and the efficiency of educational production are relevant to the policy debates, and many research results enter swiftly into the discussions. Yet, available research raises questions about the effectiveness of current spending.

This chapter considers the underlying approach to understanding the impact of resources and provides an overview of the current state of knowledge. Because of the interest in the topic and the recent dramatic expansions in the data available about schools, the field is rapidly expanding. Thus, the empirical results are soon dated, even if the general approach and conclusions remain viable.

The discussion is broad in its coverage – identifying research around the globe. But it is also narrow in its focus – emphasizing resource effects while leave details such as the impacts of teacher quality to other chapters. The chapter also attempts to link research and policy implications, since the attention given to the area is largely driven by its closeness to actual policy making.

1. Overview and motivation

A wide variety of policy discussions about schools revolve around the quality of schools. Most schooling around the world is publicly provided, and governments routinely and regularly make decisions about the support of schools. In that, they are frequently motivated by goals of student performance, but they cannot directly affect the level of outcomes. This leads to a concentration on items that can be directly controlled or affected by policy. Perhaps the most common focus is the level of resources provided to schools, although this clearly is just one part of the overall policy picture. Governmental policy can and does have considerable impact through the regulations on schools and the incentives that are set up by the funding, monitoring, and regulation of schools [see, for example, the international evidence in Wößmann (2001, 2003a, 2003b)].

Indeed, one major thrust of policy over recent years has been the concentration more directly on outcomes through the specification of objectives (or standards) for schools and through the assessment of student accomplishment of these objectives through a range of accountability systems.[1] The policies of individual US states, for example, were reinforced by federal policy to hold schools responsible for achievement results.

[1] See, for example, Hanushek and Raymond (2005) for the US or Burgess et al. (2005) for England.

Nonetheless, even when the focus is on student outcomes, the debates invariably return to questions about whether the resources provided to the schools are sufficient.

The discussion of resource usage in schools is closely related to questions about the efficiency of schools. It is quite natural to think of the level of resource usage as an index of school quality. By analogy to a profit maximizing firm, if schools are effectively maximizing outcomes and are producing efficiently, then the resources into schools will be an index of the outcomes. Moreover, more inputs to the firm can be expected to raise the level of outputs. The key element in this analogy is that the market competition forces the for-profit firm to efficient production.

The situation with government provision of schools and without direct competition changes the perspective dramatically. The presumption of efficient provision is suspect when government produces the services. And, if the resource use is inefficient, the relationship between added resources and outcomes is unclear. This simple observation motivates a direct investigation of the relationship between outcomes and inputs to schools.

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 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 chapter delves into the evidence available on this score in some detail.

2. Measurement of outcomes

Economists have devoted considerable attention to understanding how human capital affects a variety of economic outcomes. The underlying notion is that individuals make investment decisions in themselves through schooling and other routes. The accumulated skills that are relevant for the labor market from these investments over time represent an important component of the human capital of an individual. The investments made to improve skills then return future economic benefits in much the same way that a firm's investment in a set of machines (physical capital) returns future production and income. In the case of public education, parents and public officials act as trustees for their children in setting many aspects of the investment paths.

In looking at human capital and its implications for future outcomes, economists are frequently agnostic about where these skills come from or how they are produced. Although we return to that below, it is commonly presumed that formal schooling is one of several important contributors to the skills of an individual and to human capital. It is not the only factor. Parents, individual abilities, and friends undoubtedly contribute.

Schools nonetheless have a special place because they are most directly affected by public policies. For this reason, we frequently emphasize the role of schools.

The human capital perspective immediately makes it evident that the real issues are ones of long-run outcomes. Future incomes of individuals are related to their past investments. It is not their income while in school or their income in their first job. Instead, it is their income over the course of their working life.

The distribution of income in the economy similarly involves both the mixture of people in the economy and the pattern of their incomes over their lifetime. Specifically, most measures of how income and well-being vary in the population do not take into account the fact that some of the low-income people have low incomes only because they are just beginning a career. Their lifetime income is likely to be much larger as they age, gain experience, and move up in their firms and career. What is important is that any noticeable effects of the current quality of schooling on the distribution of skills and income will only be realized years in the future, when those currently in school become a significant part of the labor force. In other words, most workers in the economy were educated years and even decades in the past – and they are the ones that have the most impact on current levels of productivity and growth, if for no reason other than that they represent the larger share of active workers.

Much of the early and continuing development of empirical work on human capital concentrates on the role of school attainment, that is, the quantity of schooling. The revolution in the United States during the twentieth century was universal schooling. This has spread around the world, encompassing both developed and developing countries. Quantity of schooling is easily measured, and data on years attained, both over time and across individuals, are readily available.

Today, however, policy concerns in most corners of the world revolve much more around issues of quality than issues of quantity.

Most economists tend to emphasize labor market outcomes when thinking about differences in individual skills (as with the basic human capital models). For most schooling discussions, however, direct analysis of labor market outcomes is not practical, because these outcomes are only observed years after the schooling takes place. This fact makes it difficult to relate schooling experiences or other background factors to outcomes.

The widely adopted alternative considers, at least conceptually, a two step analytical procedure. The first step involves considering how schools and other influences on students relate to proxy measures of individual skills, of which measures of cognitive skills are the most readily available and most common object of analysis. The second step, rarely done within the same analysis, considers how the proxy relates to labor market outcomes. In fact, the second step has been infrequently considered at all and has largely been just assumed.

This initial discussion fills in the relationship between standardized tests of cognitive skills and later outcomes. While this step is also difficult, because it generally requires fairly long panel observations, there is now a reasonable amount of evidence that has accumulated.

Two other observations help in the interpretation of these results. First, many general discussions of schools consider a range of outcomes that include cognitive skills but also a variety of other factors such as creativity, teamwork, political knowledge, and the like.[2] Concentration on cognitive skills is not meant to deny other potential outcomes, although the limited analysis of other things makes it difficult to say much about them. Second, while part of the subsequent discussion refers to differences in test scores as reflecting school quality, it is also clear that test score variations can come from a variety of nonschool factors including families, peers and neighborhoods. The reasoning in using the narrower language about school quality is simply that we generally believe (and have evidence) that school quality differences are directly related to test score differences.

2.1. Impacts of quality on individual incomes – developed countries

One of the challenges in understanding the impact of quality differences in human capital has been simply knowing how to measure quality. Much of the discussion of quality – in part related to new efforts to provide better accountability – has identified cognitive skills as the important dimension. And, while there is ongoing debate about the testing and measurement of these skills, most parents and policy makers alike accept the notion that cognitive skills are a key dimension of schooling outcomes. The question is whether this proxy for school quality – students' performance on standardized tests – is correlated with individuals' performance in the labor market and the economy's ability to grow. Until recently, little comprehensive data have been available to show any relationship between differences in cognitive skills and any related economic outcomes. Such data are now becoming available.

Beginning with Mincer (1970, 1974), economists have employed readily available census data to estimate what is now simply referred to as a "Mincer equation":

$$\ln(y_i) = a_0 + \rho S_i + a_1 Exp + a_2 Exp^2 + X_i \beta + \varepsilon_i, \tag{1}$$

where y_i is earnings, S_i is years of schooling, Exp_i is labor market (or potential) experience, X_i is a vector of other individual attributes and ε_i is an error term. The object of attention, ρ, is interpreted as the rate of return to a year of schooling, and this has been estimated for a very large number of countries around the world [see Psacharopoulos (1994)].[3]

[2] Indeed, much of the discussion of human capital and schooling makes a distinction between private and social returns to schooling. The social returns often include a broad set of factors that might have externalities such as the impact on crime, the functioning of democracy, and so forth. See Hanushek (2002).

[3] There has been some controversy over exactly how to estimate the rate of return to school attainment. The main issue has revolved around whether or not a causal interpretation can be given to ρ. The argument has been that higher-ability students are more likely to continue in schooling. Therefore, part of the higher earnings observed for those with additional schooling really reflects pay for added ability and not for the additional schooling. Early discussion of ability bias can be found in Griliches (1974). Economists have pursued

A variety of researchers, however, have investigated how quality enters, and they document that the earnings advantages to higher achievement on standardized tests are quite substantial. These results are derived from different specific approaches, but the basic underlying analysis involves estimating a standard "Mincer" earnings function and adding a measure of individual cognitive skills,

$$\ln(y_i) = a_0 + \rho S_i + \gamma T_i + a_1 Exp + a_2 Exp^2 + X_i \beta + \varepsilon_i, \tag{2}$$

where T_i is the individual's measured cognitive skill and γ is the return to quality.

There is mounting evidence that quality measured by test scores is directly related to individual earnings, productivity, and economic growth. A variety of researchers documents that the earnings advantages to higher achievement on standardized tests are quite substantial.[4] While these analyses emphasize different aspects of individual earnings, they typically find that measured achievement has a clear impact on earnings after allowing for differences in the quantity of schooling, the experiences of workers, and other factors that might also influence earnings. In other words, higher quality as measured by tests similar to those currently being used in accountability systems around the country is closely related to individual productivity and earnings.

Three recent US studies provide direct and quite consistent estimates of the impact of test performance on earnings [Mulligan (1999), Murnane et al. (2000), Lazear (2003)]. These studies employ different nationally representative data sets that follow students after they leave schooling and enter the labor force. When scores are standardized, they suggest that one standard deviation increase in mathematics performance at the end of high schools translates into 12 percent higher annual earnings.

Murnane et al. (2000) provide evidence from the High School and Beyond and the National Longitudinal Survey of the High School Class of 1972. Their estimates suggest some variation with males obtaining a 15 percent increase and females a 10 percent increase per standard deviation of test performance. Lazear (2003), relying on a somewhat younger sample from NELS88, provides a single estimate of 12 percent. These estimates are also very close to those in Mulligan (1999), who finds 11 percent for the normalized AFQT score in the NLSY data. By way of comparison, estimates of the value of an additional year of school attainment are typically 7–10 percent.

a variety of analytical approaches for dealing with this. The approaches have included looking for circumstances where the amount of schooling is affected by things other than the student's valuation of continuing and considering the income differences among twins [see Card (1999)]. The various adjustments for ability differences typically make small differences on the estimates of the value of schooling, and Heckman and Vytlacil (2001) argue that it is not possible to separate the effects of ability and schooling.

4 These results are derived from different specific approaches, but the basic underlying analysis involves estimating a standard "Mincer" earnings function and adding a measure of individual cognitive skills. This approach relates the logarithm of earnings to years of schooling, experience, and other factors that might yield individual earnings differences. The clearest analyses are found in the following references [which are analyzed in Hanushek (2002)]. See Bishop (1989, 1991), O'Neill (1990), Grogger and Eide (1993), Blackburn and Neumark (1993, 1995), Murnane, Willett and Levy (1995), Neal and Johnson (1996), Mulligan (1999), Murnane et al. (2000, 2001), Altonji and Pierret (2001) and Lazear (2003).

There are reasons to believe that these estimates provide a lower bound on the impact of higher achievement. First, these estimates are obtained fairly early in the work career (age from mid-twenties to early-thirties), and other analysis suggests that the impact of test performance becomes larger with experience.[5] Second, the labor market experiences that are observed begin the mid-1980s and extend into the mid-1990s, but other evidence suggests that the value of skills and of schooling has grown throughout and past that period. Third, future general improvements in productivity are likely to lead to larger returns to skill.[6]

A limited number of additional studies are available for developed countries outside of the United States. McIntosh and Vignoles (2001) study wages in the United Kingdom and find strong returns to both numeracy and literacy.[7] Finnie and Meng (2002) and Green and Riddell (2003) investigate returns to cognitive skills in Canada. Both suggest that literacy has a significant return, but Finnie and Meng (2002) find an insignificant return to numeracy. This latter finding stands at odds with most other analyses that have emphasized numeracy or math skills.

Another part of the return to school quality comes through continuation in school.[8] There is substantial US evidence that students who do better in school, either through grades or scores on standardized achievement tests, tend to go farther in school.[9]

[5] Altonji and Pierret (2001) find that the impact of achievement grows with experience, because the employer has a chance to observe the performance of workers.

[6] The earnings analyses typically compare workers of different ages at one point in time to obtain an estimate of how earnings will change for any individual. If, however, productivity improvements occur in the economy, these will tend to raise the earnings of individuals over time. Thus, the impact of improvements in student skills are likely to rise over the work life instead of being constant as portrayed here, at least if the technologies expand similar to the past with biases toward skilled labor.

[7] Because they look at discrete levels of skills, it is difficult to compare the quantitative magnitudes directly to the US work.

[8] Much of the work by economists on differences in worker skills has actually been directed at the issue of determining the average labor market returns to additional schooling. The argument has been that higher-ability students are more likely to continue in schooling. Therefore, part of the higher earnings observed for those with additional schooling really reflects pay for added ability and not for the additional schooling. Economists have pursued a variety of analytical approaches for dealing with this, including adjusting for measured cognitive test scores, but this work generally ignores issues of variation in school quality. The approaches have included looking for circumstances where the amount of schooling is affected by things other than the student's valuation of continuing and considering the income differences among twins [see Card (1999)]. The various adjustments for ability differences typically make small differences on the estimates of the value of schooling, and Heckman and Vytlacil (2001) argue that it is not possible to separate the effects of ability and schooling. The only explicit consideration of school quality typically investigates expenditure and resource differences across schools, but these are known to be poor measures of school quality differences [Hanushek (2002)].

[9] 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, Rivkin and Taylor (1996), in considering the factors that influence school attainment, find that individual achievement scores are highly correlated with continued school attendance. Neal and

Murnane et al. (2000) separate the direct returns to measured skill from the indirect returns of more schooling and suggest that perhaps one-third to one-half of the full return to higher achievement comes from further schooling. Note also that the effect of quality improvements on school attainment incorporates concerns about drop out rates. Specifically, higher student achievement keeps students in school longer, which will lead among other things to higher graduation rates at all levels of schooling.[10]

The impact of test performance on individual earnings provides a simple summary of the primary economic rewards to an individual. This estimate combines the impacts on hourly wages and on employment/hours worked. It does not include any differences in fringe benefits or nonmonetary aspects of jobs. Nor does it make any allowance for aggregate changes in the labor market that might occur over time.

2.2. Impacts of quality on individual incomes – developing countries

Questions remain about whether the clear impacts of quality in the US and other developed countries generalize further, particularly developing countries. The literature on returns to cognitive skills in developing countries is restricted to a relatively limited number of countries: Ghana, Kenya, Morocco, Pakistan, South Africa and Tanzania. Moreover, a number of studies actually employ the same basic data, albeit with different analytical approaches, but come up with somewhat different results.

Table 1 provides a simple summary to the quantitative estimates available for developing countries. The summary of the evidence permits a tentative conclusion that the returns to quality may be even larger in developing countries than in developed countries. This of course would be consistent with the range of estimates for returns to quantity of schooling [e.g., Psacharopoulos (1994)], which are frequently interpreted as indicating diminishing marginal returns to schooling.

There are some reasons for caution in interpreting the precise magnitude of estimates. First, the estimates appear to be quite sensitive to the estimation methodology itself. Both within individual studies and across studies using the same basic data, the results are quite sensitive to the techniques employed in uncovering the fundamental parameter

Johnson (1996) in part use the impact of achievement differences of blacks and whites on school attainment to explain racial differences in incomes. Their point estimates of the impact of cognitive skills (AFQT) on earnings and school attendance appear to be roughly comparable to that found in Murnane et al. (2000). 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 various determinants of achievement. Hanushek and Pace (1995) find that college completion is significantly related to higher test scores at the end of high school.

[10] This work has not, however, investigated completely how achievement affects the ultimate outcomes of additional schooling. For example, if over time lower-achieving students tend increasingly to attend further schooling, these schools may be forced to offer more remedial courses, and the variation of what students know and can do at the end of school may expand commensurately.

Table 1
Summary of estimated returns to a standard deviation increase in cognitive skills

Country	Study	Estimated effect[a]	Notes
Ghana	Glewwe (1996)	0.21**–0.3** (government) 0.14–0.17 (private)	Alternative estimation approaches yield some differences; math effects shown generally more important than reading effects, and all hold even with Raven's test for ability
Ghana	Jolliffe (1998)	0.05–0.07*	Household income related to average math score with relatively small variation by estimation approach; effect from off-farm income with on-farm income unrelated to skills
Ghana	Vijverberg (1999)	?	Income estimates for math and reading with nonfarm self-employment; highly variable estimates (including both positive and negative effects) but effects not generally statistically significant
Kenya	Boissiere, Knight and Sabot (1985), Knight and Sabot (1998)	0.19**–0.22**	Total sample estimates: small variation by primary and secondary school leavers
Morocco	Angrist and Lavy (1997)	?	Cannot convert to standardized scores because use indexes of performance; French writing skills appear most important for earnings, but results depend on estimation approach
Pakistan	Alderman et al. (1996)	0.12–0.28*	Variation by alternative approaches and by controls for ability and health; larger and more significant without ability and health controls
Pakistan	Behrman, Ross and Sabot (2006)	0.25	Estimates of structural model with combined scores for cognitive skill; index significant at 0.01 level
South Africa	Moll (1998)	0.34**–0.48**	Depending on estimation method, varying impact of computation; comprehension (not shown) generally insignificant
Tanzania	Boissiere, Knight and Sabot (1985), Knight and Sabot (1998)	0.07–0.13*	Total sample estimates: smaller for primary than secondary school leavers

*Significant at 0.05 level.
**Significant at 0.01 level.
[a]Estimates indicate proportional increase in wages from a one standard deviation increase in measured test scores.

for cognitive skills.[11] Second, the evidence on variations within developing countries is not entirely clear. For example, Jolliffe (1998) finds little impact of skills on farm income, while Behrman, Ross and Sabot (2006) suggest an equivalence across sectors at least on theoretical grounds.

Nonetheless, the overall summary is that the available estimates of the impact of cognitive skills on outcomes suggest strong economic returns within developing countries. The substantial magnitude of the typical estimates indicates that quality concerns are very real for developing countries and that this aspect of schools simply cannot be ignored – a topic that comes up below.

2.3. Impacts of quality on economic growth

The relationship between measured labor force quality and economic growth is perhaps even more important than the impact of human capital and school quality on individual productivity and incomes. Economic growth determines how much improvement will occur in the overall standard of living of society. Moreover, the education of each individual has the possibility of making others better off (in addition to the individual benefits just discussed). Specifically, a more educated society may lead to higher rates of invention; may make everybody more productive through the ability of firms to introduce new and better production methods; and may lead to more rapid introduction of new technologies. These externalities provide extra reason for being concerned about the quality of schooling.

The current economic position of the United States, for example, is largely the result of its strong and steady growth over the twentieth century. Economists have developed a variety of models and ideas to explain differences in growth rates across countries – invariably featuring the importance of human capital.[12]

The empirical work supporting growth analyses has emphasized school attainment differences across countries. Again, this is natural because, while compiling comparable data on many things for different countries is difficult, assessing quantity of schooling is more straightforward. The typical study finds that quantity of schooling is highly related to economic growth rates. But, quantity of schooling is a very crude measure of the knowledge and cognitive skills of people – particularly in an international context.

Hanushek and Kimko (2000) go beyond simple quantity of schooling and delve into quality of schooling.[13] Kimko and I incorporate the information about international differences in mathematics and science knowledge that has been developed through testing over the past four decades. And we find a remarkable impact of differences in school quality on economic growth.

[11] The sensitivity to estimation approach is not always the case; see, for example, Jolliffe (1998). A critique and interpretation of the alternative approaches within a number of these studies can be found in Glewwe (2002).

[12] Barro and Sala-i-Martin (2003) review recent analyses and the range of factors that are included.

[13] Barro and Lee (2001) provide an analysis of qualitative differences that also includes literacy.

The international comparisons of quality come from piecing together results of a series of tests administered over the past four decades. In 1963 and 1964, the International Association for the Evaluation of Education al Achievement (IEA) administered the first of a series of mathematics tests to a voluntary group of countries. These initial tests suffered from a number of problems, but they did prove the feasibility of such testing and set in motion a process to expand and improve on the undertaking.[14]

Subsequent testing, sponsored by the IEA and others, has included both math and science and has expanded on the group of countries that have been tested. In each, the general model has been to develop a common assessment instrument for different age groups of students and to work at obtaining a representative group of students taking the tests. An easy summary of the participating countries and their test performance is found in Figure 1. This figure tracks performance aggregated across the age groups and subject area of the various tests and scaled to a common test mean of 50.[15] (The United States and the United Kingdom are the only countries to participate in all of the testing.)

There is some movement across time of country performance on the tests, but for the one country that can be checked – the United States – the pattern is consistent with other data. The National Assessment of Educational Progress (NAEP) in the United States is designed to follow performance of US students for different subjects and ages. NAEP performance over this period, shown in Figure 2, also exhibits a sizable dip in the seventies, a period of growth in the eighties, and a leveling off in the nineties.

Kimko's and my analysis of economic growth is very straightforward. We develop a consistent measure of labor force quality based on information about international differences in mathematics and science knowledge. We combine all of the available earlier test scores into a single composite measure of quality and consider statistical models that explain differences in growth rates across nations during the period 1960 to 1990.[16] The basic statistical models relate annual growth rates of GDP per capita (g_c) to our measure of labor force quality (T_c), the initial level of income (Y^0), the quantity of schooling (S_c), and a vector of other control variables (Z_c) which includes

[14] The problems included issues of developing an equivalent test across countries with different school structure, curricula and language; issues of selectivity of the tested populations; and issues of selectivity of the nations that participated. The first tests did not document or even address these issues in any depth.

[15] The details of the tests and aggregation can be found in Hanushek and Kimko (2000) and Hanushek and Kim (1995). Figure 1 excludes the earliest administration and runs through the Third International Mathematics and Science Study (TIMSS) administered in 1995. Other international tests have been given and are not included in the figure. First, reading and literacy tests have been given in 1991 and very recently. The difficulty of unbiased testing of reading across languages plus the much greater attention attached to math and science both in the literature on individual earnings and in the theoretical growth literature led to the decision not to include these test results in the empirical analysis. Second, the most recent follow-up to the 1995 TIMSS in math and science (given in 1999) plus the 2003 TIMSS and the PISA tests for 2000 and 2003 are excluded from the figure simply for presentational reasons.

[16] We exclude the recent TIMSS tests from 1995 through 2003 and the OECD's PISA tests because they were taken outside of the analytical period on economic growth. We combine the test measures over the 1965–1991 period into a single measure for each country. The underlying objective is to obtain a measure of quality for the labor force in the period during which growth is measured.

Figure 1. Performance on international mathematics and science examinations.

in different specifications the population growth rates, political measures, openness of the economies, and the like:

$$g_c = \alpha_0 + \eta T_c + \alpha_1 Y_c^0 + \alpha_2 S_c + Z_c \phi + \nu_c. \tag{3}$$

Most important, the impact of the quality of the labor force as measured by math and science scores (η) is extremely important. One standard deviation difference on test performance is related to 1 percent difference in annual growth rates of gross domestic product (GDP) per capita.[17] This quality effect, while possibly sounding small, is actually very large and significant. Because the added growth compounds, it leads to powerful effects on national income and on societal well-being.

One common concern in analysis such as this is that schooling might not be the actual cause of growth but, in fact, may just reflect other attributes of the economy that are beneficial to growth. For example, as seen in Figure 1, the East Asian countries consistently score very highly on the international tests, and they also had extraordinarily high growth over the 1960–1990 period. It may be that other aspects of these East Asian economies have driven their growth and that the statistical analysis of labor force quality simply is picking out these countries. But in fact, even if the East Asian countries are excluded from the analysis, a strong – albeit slightly smaller – relationship is still observed with test performance. This test of sensitivity of the results seems to reflect a basic importance of school quality, a factor that contributes also to the observed growth of East Asian countries.

Another concern might be that other factors that affect growth, such as efficient market organizations, are also associated with efficient and productive schools – so that, again, the test measures are really a proxy for other attributes of the country. In order to investigate this, we concentrate on immigrants to the United States who received their education in their home countries. We find that immigrants who were schooled in countries that have higher scores on the international math and science examinations earn more in the United States. This analysis makes allowance for any differences in school attainment, labor market experience, or being native English-language speakers. In other words, skill differences as measured by the international tests are clearly rewarded in the United States labor market, reinforcing the validity of the tests as a measure of individual skills and productivity.[18]

In sum, although cognitive test scores may not measure all of the various outcomes expected from schools, they do provide important information on quality as related to the labor market returns. Further, no other measure provides such a consistent and validated assessment of the quality of educational outcomes.

[17] The details of this work can be found in Hanushek and Kimko (2000) and Hanushek (2003b). Importantly, adding other factors potentially related to growth, including aspects of international trade, private and public investment, and political instability, leaves the effects of labor force quality unchanged.

[18] Finally, the observed relationships could simply reflect reverse causality, that is, that countries that are growing rapidly have the resources necessary to improve their schools and that better student performance is the result of growth, not the cause of growth. This in fact is closely related to the analysis below about the impact of resources on achievement, and thus the discussion is left until later.

3. Aggregate United States performance

Given that student assessments provide a measure of school outcomes, it is possible to begin the investigation of how school resources (and other factors) relate to student performance. It is instructive to begin with the simplest overall evidence that comes from aggregate scores and then to move to more detailed analytical studies.

The simplest and perhaps clearest demonstration of the resource story is found in the aggregate United States data over the past few decades. The United States, operating under a system that is largely decentralized to the fifty separate states, has pursued the conventionally advocated resource policies vigorously. Table 2 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 unionized, 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 2, real expenditures per pupil more than tripled over this period.[19] 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 3.5% 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 fifty-year period.[20]

Table 2
Public school resources in the United States, 1960–2000

	1960	1970	1980	1990	2000
Pupil–teacher ratio	25.8	22.3	18.7	17.2	16.0
Percent of teachers with master's degree or more	23.5	27.5	49.6	53.1	56.2[a]
Median years teacher experience	11	8	12	15	15[a]
Current expenditure/ADA (2000/2001)	$2,235	$3,782	$5,124	$6,867	$7,591

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

[19] The calculation of real expenditures deflates by the Consumer Price Index. Use of a wage deflator (see the discussion of prices below) does not significantly change this picture.
[20] 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.

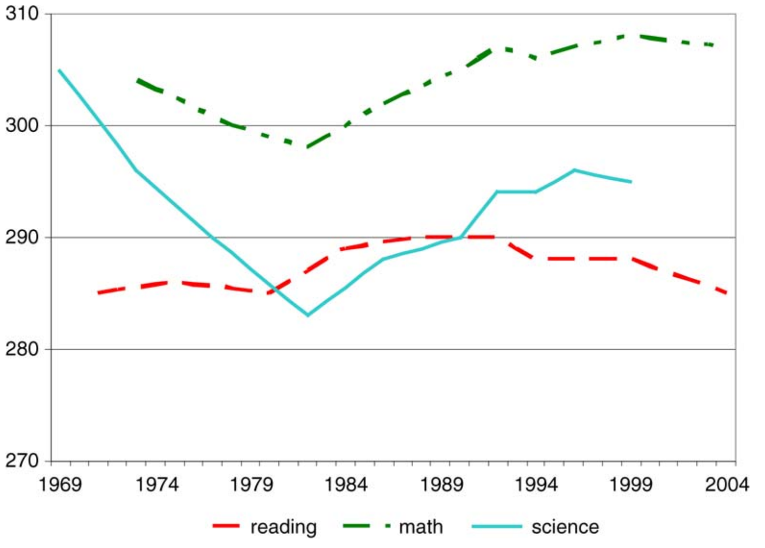

Figure 2. Trends in NAEP performance by subject, 17-year-olds.

The question remains, what was obtained for these spending increases? As mentioned above, 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 2 gives performance data for the relevant period as the previously described input data.[21] In this figure the pattern of average performance by 17-year-olds is traced for reading, mathematics and science. The performance of students in math is slightly higher (less than 0.1 standard deviations) in 2004 than thirty years earlier when spending was dramatically lower. The performance in reading in 2004 is exactly where it was in 1971. The performance of students in science is significantly lower in 1999 (the latest observation) than it was in 1970. Writing performance (not shown) was first tested in 1984 and declined steadily until 1996 when testing was discontinued.[22]

The only other test that provides a national picture of US 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

[21] 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 2 that goes back to 1960 is relevant for comparison with the NAEP performance data.

[22] With writing, test reliability is an issue, and this led to suspension of the writing testing in NAEP.

voluntary test taken by a selective subset of the population.[23] 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.[24]

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.

Three 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; and (3) changing relative prices have increased schooling costs.

3.1. 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 hypothesized 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 percent 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 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 (1997a)], has demonstrated that differences in families are very important for student achievement. Most of these studies have not focused

[23] 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.

[24] 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)].

their primary attention on families, however, and thus have not delved very far into the measurement and structure of any family influences.[25]

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 [Grissmer et al. (1994a)] 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, but there are reasons to be skeptical about these results.[26]

3.2. Exogenous cost increases

The most discussed cost concern involves "special education", programs to deal with students who have various disabilities. The issue is that these programs are expensive but the recipients tend not to take standardized tests. Thus, even if special education programs are effective [Hanushek, Kain and Rivkin (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, counseling activities, and educational services to be provided for handicapped students. To implement this and subsequent laws and regulations, school systems expanded staff and programs, developing entirely new administrative structures in many cases to handle

[25] Grissmer et al. (1994b) attempts 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 expected, leading presumably to the opposite conclusion that schools for the majority of students actually got worse over time.

[26] Skepticism 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 [cf. Mayer (1997)].

"special education". The general thrust of the educational services has been to provide regular classroom instruction where possible ("mainstreaming") along with specialized 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 2004, the percentage of students classified as disabled increases from 8.3% to 13.8%. 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, Danielson and Brauen (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 programs have undoubtedly influenced overall expenditures, they remain a relatively small portion of the total spending on schools.

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

3.3. Changing prices

A series of well-known arguments emphasize the cost implications of differential technological change and productivity growth [Scitovsky and Scitovsky (1959), Baumol and Bowen (1965), Baumol (1967)]. The focus of this work is the cost disadvantage of a sector that experiences little apparent technological change while other sectors undergo regular productivity improvements. Because the rise in real wages – increases above general inflation – are roughly proportional to the average growth rate of labor productivity in all sectors, the technologically stagnant sector faces increased real labor costs. In other words, industries with rapid improvements in their ability to produce outputs can afford to pay more for workers and will bid up the wages of workers. It is often assumed that the nature of production prevents the stagnant sector from hiring fewer of the increasingly costly labor inputs, thus leading to increases in the price of output. The lack of substitutability of machines for workers can arise either because of some necessity (e.g., the need for four musicians in a horn quartet) or because the quantity of labor input is directly related to perceived quality (e.g., class sizes and the demand for teachers in schools).[27] These simple predictions of increasing costs in low

[27] Measurement issues abound. For example, while musical groups may be constrained to a relatively fixed mix of musicians, some believe the advent of recordings, radio, television, and now the Internet have led to a very large expansion of output for the same number of musicians. If defined solely in terms of concert performances, there may be little substitutability, but this does not hold if defined in terms of total music output.

productivity growth sectors, often termed simply "Baumol's disease", dominate explanations for cost growth in government services, the arts, many nonprofit activities, and other industries in which labor services are the most significant input factor. Part of this price increase in schools might simply reflect Baumol's disease. Schools rely heavily on college-trained workers, and the relative pay of college workers has risen dramatically since the mid-1970s [Murphy and Welch (1992), Hanushek et al. (1994)].

This argument cannot, however, change the conclusions on inputs and school outcomes. First, it is important to note that these arguments do not apply to the changes in real resources in Table 2 but only to the cost aggregation in the final line. Second, in terms of real spending, the arguments would imply that the cost deflator understates the change in input costs. But, if costs are deflated on a wage basis for recent periods (1967–1991), the results are quite ambiguous because schools were actually drawing from a lower point in the wage distribution over time [see Hanushek and Rivkin (1997) and Hanushek (1997b)]. Finally, as shown in Table 2, schools have actually substituted *more* of the expensive input (teachers) rather than less over time.

4. Aggregate international data

Most other countries of the world have not tracked student performance over any length of time, making analyses comparable to the United States discussion impossible. Nonetheless, international testing over the past four decades permits an overview of spending across countries. As discussed above, a series of international tests – given from 1963 through 2003 – provide some indication of national performance. (Only the US and UK participated in all tests.) The test performance across time, updated from Hanushek and Kimko (2000), were summarized in Figure 1.

The important feature of these is that cross-country performance bears little relationship to the patterns of expenditure across the countries. Figure 3 shows the comparison of scores on the PISA tests in 2003 and the spending of nations in purchasing power parity terms.[28] Countries are ranked in terms of the average score on the PISA tests, and the height of the bars indicates spending. Except for the developing countries, which both spend noticeably less than the others and perform noticeably poorer, there is little association between spending and performance. For all countries spending at least $1,500 per student, the correlation of spending and performance is 0.18. (At the low end of performance, there are also questions about the validity of the assessments. Scores in the developing countries are very far from the mean of the distribution for developed countries, the set for which the tests were developed.)

[28] The Programme for International Student Assessment (PISA) is conducted by the OECD and tests 15-year-olds in mathematics, science and reading. It has been conducted in 2000 and 2003. The scores in the figure are the average across the three subjects. The spending calculations come from Organisation for Economic Co-operation and Development (2003).

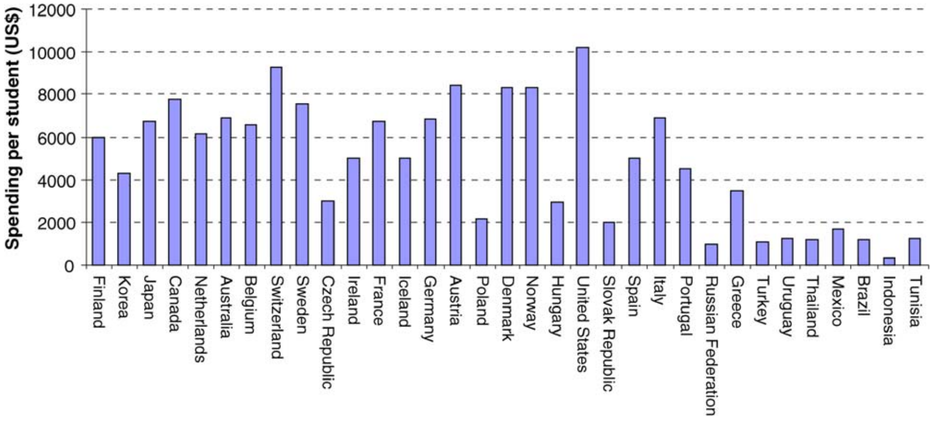

Figure 3. Expenditure per student at all levels (countries ranked by combined PISA 2003 scores).

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 United States. 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.[29] This estimation consistently indicates a statistically significant negative effect of added resources on performance after controlling for other influences.[30]

Gundlach, Wossmann and Gmelin (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. Wößmann (2001, 2003b) also performs a related analysis that relies on just the 1995 performance information from the Third International Mathematics and Science Study (TIMSS). His analysis suggests that traditional resource measures bear little consistent relationship to differences in scores among the 39 nations participating in TIMSS for 13-year-olds.

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.

[29] 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.

[30] This can also be thought of as an additional causality test in the growth models, because growing countries that spend more on schools to not see higher achievement – the concern with reverse causality.

5. Econometric approach

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 conclusions of the Coleman study but, more importantly, has clarified the interpretation.

This initial study led to much follow on work of both an empirical and a conceptual variety. This genre is often labeled "education production function" analysis and has pursued a wide variety of specialized analyses. Various conceptual discussions and reviews currently exist, and the focus here is understanding the results and their interpretation.[31]

The framework of analysis of educational performance considers a general production function such as:

$$O_{it} = f\left(F_i^{(t)}, P_i^{(t)}, S_i^{(t)}, A_i\right) + \upsilon_{it}, \tag{4}$$

where O_{it} – performance of student i at time t, $F_i^{(t)}$ – family inputs cumulative to time t, $P_i^{(t)}$ – cumulative peer inputs, $S_i^{(t)}$ – cumulative school inputs, A_i – innate ability, and a stochastic term υ_{it}.

This general structure has motivated an extensive series of empirical studies. The typical empirical study collects information about student performance and about the various educational inputs and then attempts to estimate the characteristics of the production function using econometric techniques.

Two aspects of this formulation are important to point out. First, a variety of influences outside of schools enter into the production of achievement. Second, the production process for achievement is cumulative, building on a series of inputs over time. Both of these are important in the specification and interpretation of educational production functions.

The relevance of many factors outside schools highlights the necessity of going beyond simple comparisons of student performance across schools. Most of the attention in analytical studies has focused on the measurement of school attributes. This focus seems natural from a policy point of view. It also reflects the common use of administrative data in estimating production functions, because administrative data are frequently short of many measures of family background. Nonetheless, this lack of attention to other inputs is unfortunate. First, increasing attention has been given to

[31] Conceptual discussions can be found in Hanushek (1979) and Todd and Wolpin (2003). Prior reviews of different strands of work along with more detailed considerations of the range of studies can be found in Hanushek (1986, 2003a) and Betts (1996).

potential policies related to families – such as preschool and daycare programs, after school programs, parent education and the like. Second, because families frequently exert preferences for the schools that their children will attend, incomplete measurement of external influences on performance raise intense issues of selection bias and preclude simple statements about causal influences of schools. Such an observation of course does not seem very profound, but, as discussed below, many empirical studies give little attention to nonschool influences in addressing the impact of school factors. Moreover, public policy debates surprisingly frequently rely on simple accounting of performance across schools. For example, much of the current movement toward increased school accountability often relies on just aggregate student scores for a school.[32] Just the level of student performance is not a reliable indicator of the quality of schools students are attending.

The cumulative nature of achievement, where the learning in any time period builds on prior learning, implies that any analysis must take into account the time path of inputs. This places heavy demands on measurement and data collection, because historical information is frequently difficult to obtain.

The cumulative nature of the production process has been a prime motivation for considering a value-added formulation. At least in a linear version of Equation (4), it is possible to look at the growth in contemporaneous performance over some period of time, instead of the level of performance, and relate that to the flow of specific inputs. The general value-added formulation can be written as

$$O_{it} - O_{it^*} = f^*\big(F_i^{(t-t^*)}, P_i^{(t-t^*)}, S_i^{(t-t^*)}\big) + \upsilon_{it} - \upsilon_{it^*}, \tag{5}$$

where outcome changes over the period $(t - t^*)$ are related to inputs over the same period. Note that this formulation dramatically lessens the data requirements and eliminates anything that appears as a fixed effect in the level of achievement [Equation (5)].

This formulation presumes that innate abilities are constant and thus fall out of achievement growth. With more information on variations over time, it is also possible to allow for ability differences in growth [Rivkin, Hanushek and Kain (2005)]. Alternative formulations have prior achievement, O_{it^*}, on the right-hand side, allowing for coefficient different than one [Hanushek (1979)]. This latter approach has the advantages of allowing for different scales of measurement in achievement during different years and introducing the possibility that growth in performance differs by starting point. It has the disadvantages of introducing measurement error on the right-hand side and of complicating the error structure, particularly in models relying on more than a single year of an individual's achievement growth.

[32] With the increasing popularity of publishing average performance of students in different schools, the interpretation of scores becomes more important. In fact, without consideration of the various inputs that go beyond just schools, alternative accountability systems can have perverse effects [cf. Hanushek and Raymond (2001)]. The integration of the underlying theoretical and empirical analysis of the determination of achievement with accountability and incentive systems is an important but underdeveloped area of investigation.

When this formulation is generalized by moving lagged achievement to the right-hand side, the coefficient on lagged achievement indicates the persistence of any prior input effects. The estimation becomes problematic if one is including fixed effects for individual students, since this implies biases from the endogeneity of prior achievement. In this case, it is necessary to use an instrumental variables approach. The statistical problems in estimation using the difference form of Equation (5) depend directly on how far the coefficient on lagged achievement is from one.

In any event, a key element in the consideration of estimation results is the likelihood that they are biased because of the estimation approach, model specification, or available samples of observations. These issues, along with a variety of approaches for dealing with them, are discussed below.

6. United States econometric evidence

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 United States schools here 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.[33] 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

[33] Individual publications included in the following summary 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 emphasized by the authors of the underlying papers. In some cases, this rule did not lead to a clear choice, at which time the tabulation emphasized 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.

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. No attempt has been made to catalog the newer general analyses, because they have not yielded markedly different results. Individual studies of note will, however, be separately discussed.

Understanding the character of the underlying analyses is important for the subsequent interpretation. Three-quarters of the estimates rely on student performance (O_{it}) measured by standardized tests, while the remainder uses a variety of different measures including such things as continuation in school, dropout behavior, and subsequent labor 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.[34]

Table 3 presents the overall summary of basic results about the key resources that form the basis for most overall policy discussions.[35] The standard hypothesis driving

Table 3
Percentage distribution of estimated effect of key resources on student performance, based on 376 studies

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

Source: Hanushek (1997a), revised.

[34] A more complete description of the universe of studies that are reviewed can be found in Hanushek (1997a).

[35] A more complete description of the studies can be found in Hanushek (1997a), which updates the analysis in Hanushek (1986). The tabulations here correct some miscoding of effects in these original publications. They also omit the estimates from Card and Krueger (1992b). 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.

policy initiatives is that each of these resources should have a positive effect on student performance.[36] 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.[37] 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.[38] 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. And, 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, Kain and Rivkin (2004b) and Hanushek et al. (2005)]. 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 noticeable improvements, it seems somewhat unimportant to investigate the size of any estimated effects.

The financial aggregates 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.

The studies that include measures of facilities or administration fail to show much of significance. These factors are generally, however, quite poorly measured – thus complicating the interpretation.

Finally, this review does not concentrate on specific measures of teachers, since that is covered in the chapter on teacher quality in this volume. It is true that measures of teacher test scores and their impact on student performance are more consistently related

[36] 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.

[37] 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.

[38] 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.

to student outcomes than the factors identified previously.[39] But, they are generally not paid for directly or indirectly and thus are not directly related to resources. Thus, the general character of the evidence and implications of these analyses are left for the chapter on teacher quality.

6.1. 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 United States is made chiefly by the separate 50 states, and the resulting variations in spending, regulations, graduation requirements, testing, labor 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 United States) but also holds for students who do not move because of the heterogeneity of teachers within individual schools [see Hanushek, Kain and Rivkin (2004a), Rivkin, Hanushek and Kain (2005)].[40] 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

[39] Of the 41 studies with measures of teacher test scores, 41 percent are positive and statistically significant, while 10 percent are negative and statistically significant.

[40] A third argument on quality is made in Krueger (2002). He argues that summaries of econometric results should be weighted by publication counts, i.e., a published study that provides separate estimates on resource effects in grades 3, 6 and 9 should received the same weight as a study that provides a single estimate for grade 3. He provides no theoretical or empirical justification for this weighting, and it is not employed here.

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 rather than if conducted at the classroom level [Hanushek, Rivkin and Taylor (1996)].[41]

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

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[a]	157	11	18	71
Multiple state samples[b]	119	18	8	74
Disaggregated within states[c]	109	14	8	78
State level aggregation[d]	10	60	0	40
B. *Expenditure per pupil*				
Total	163	27%	7%	66%
Single state samples[a]	89	20	11	69
Multiple state samples[b]	74	35	1	64
Disaggregated within states[c]	46	17	0	83
State level aggregation[d]	28	64	4	32

Source: Hanushek (1997a), revised.
[a]Estimates from samples drawn within single states.
[b]Estimates from samples drawn across multiple states.
[c]Resource measures at level of classroom, school, district, or county, allowing for variation within each state.
[d]Resource measures aggregated to state level with no variation within each state.

[41] The discussion of aggregation is part of a broader debate trying to reconcile the findings of Card and Krueger (1992a) with those presented here. For a fuller discussion, see Burtless (1996). Of particular relevance is Heckman, Layne-Farrar and Todd (1996a, 1996b), 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.

minimized 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 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).[42]

This pattern of results is consistent with expectations from considering specification biases when favorable 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 discussion by 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 high quality estimates paints a stronger picture. Therefore, a more careful set of econometric analyses confirms the basic picture presented in the aggregate data.

[42] Expenditure studies virtually never include include 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, Rivkin and Taylor (1996), however, suggests that measurement error is not the driving force behind the pattern of results.

Table 5
Percentage distribution of 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 studies*				
Teacher–pupil ratio	78	12%	8%	80%
Teacher education	40	0	10	90
Teacher experience	61	36	2	62
B. *Studies within a single state*				
Teacher–pupil ratio	23	4%	13%	83%
Teacher education	33	0	9	91
Teacher experience	36	39	3	58

Source: Hanushek (1997a), revised.

6.2. *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 behavioral 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 United States tends on average to be roughly equally divided between states and localities with the federal government contributing only about 7% of overall spending. Huge variation in funding levels and formulae nonetheless exists across states. In most state funding of schools in the United States, 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 United States 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.[43]

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 detrended variations in the size of birth cohorts to identify exogenous changes in class size in small Connecticut towns. Changes in cohort sizes, coupled with the lumpiness of classes in small school districts, can provide variations in class size that are unrelated to other factors.[44] 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 (2001)]. Unfortunately, identification of truly exogenous determinants of class size, or resource allocations more generally, is sufficiently rare that other compromises in the data and modeling are frequently required. These coincidental compromises jeopardize the ability to obtain clean estimates of resource effects and may limit the generalizability of any findings. Rivkin, Hanushek and Kain (2005), 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.[45] They find some small class size effects, but the effects vary significantly across grades and specifications.

43 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.

44 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.

45 The nature of this analysis is discussed further in the chapter on teacher quality.

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.

The inconsistency of the limited existing set of analyses should not, however, detract from the importance of the issue. Continued close attention to the nature of the statistical models and the estimation approach is warranted, and the area of attention is a fruitful one for future research efforts.

7. International econometric evidence

The evidence for countries other than the United States is potentially important for a variety of reasons. Other countries have varying institutional structures, so different findings could help to identify the importance of organization 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 United States is its schools there are generally operating in an area of severe diminishing marginal productivity, placing most observations 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.

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 many other countries introduce regular student assessment. 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.

7.1. Developing countries

Existing analyses in less developed countries have shown a similar inconsistency of estimated resource effects as that found in the United States. 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.[46] Two

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

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).

facets of these data compared to the previous United States data stand out: (1) in general, a minority of the available estimates suggests much confidence that the identified resources positively influence student performance; and (2) there is generally somewhat stronger support for these resource policies than that existing in United States 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.

The major concern with the work on developing countries is study quality. Virtually no research in developing countries has had longitudinal data on individuals.[47] Further, the data collected are often limited considerably by the survey and sampling procedures. These issues imply real concerns about any causal interpretation of the estimated relationships.

7.2. Developed countries

The evidence on developed countries outside of the United States 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. And, new articles continue to appear [e.g., Dustmann, Rajah and van Soest (2003)].

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

47 An exception is the work on Brazil in Harbison and Hanushek (1992), where a subsample of students was followed over time.

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		Number of countries	Negative		Positive		Number of countries
	Significant	Not significant	Not significant	Significant		Significant	Not significant	Not significant	Significant	
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: Bold indicates the number of statistically significant results with the expected sign of the effect. 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).

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. Note, however, that these estimates are all cross-sectional in nature and subject to a variety of concerns about specification.

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.[48] 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. It is nevertheless inconsistent with the overall findings of Wößmann and West (2006), who use within school variations to identify the impacts of class size.

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.

Wößmann (2001, 2003b) looks at cross national differences in TIMSS math 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 decentralization 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.

8. Project STAR and experimental data[49]

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

[48] An additional check analyzes 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.

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

given well-defined treatment by 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. Randomization seeks to eliminate any relationship between selection into a treatment program 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 students, 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 publicized. The simplest summary is that: (1) students 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 (math) 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 generalizable to any other grades.

Importantly, this pattern of effects is at odds with the normal rhetoric about smaller classes permitting more individualized 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 socialization or introduction into the behavior 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 randomization 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.

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 generalizations can be made.

Project STAR also teaches the difficulty in conducting true random assignment experiments. First, many people resist experiments, largely on ethical grounds. While seemingly less important than similar issues in medical research where experimentation is well established, this issue remains important. Second, it is difficult to ensure that the implementation of experiments matches the conceptual ideal.[50] Interestingly, experimentation appears to be more common in developing countries than in developed countries [e.g., see Kremer (2003)].

[50] While the experimental approach has great appeal, the actual implementation in the case of Project STAR introduces considerable uncertainty into these estimates [Hanushek (1999b)]. The uncertainty arises most importantly from questions about the quality of the randomization 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 randomization 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 assignment 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. 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 (1999b)]. 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 behavior in the experiment could be affected.

9. 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 generalizations were possible would naturally arise. Such conflicts do not appear particularly important, although there are some obvious qualifications. When considering countries that do not have a fully functioning education system even at the primary level, there clearly is some minimal level of resources for the definition of a school.[51] Nonetheless, even in the poorest areas of the world, it is difficult to identify a minimum threshold of resources where there are clear impacts on student outcomes. More refined policies that go beyond simply adding resources with no concomitant sets of policies and incentives still have high payoffs in areas with undeveloped school systems.

There is a tendency by researchers and policy makers to take a single study and to generalize 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. 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 don't 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, it seems more likely that schools have limited incentives to seek out and to employ programs that consistently relate to student achievement. It also appears, as discussed below, that much of the research employed in active policy debates has not adequately identified the causal structure and thus cannot be generalized in useful ways.

It is just this tendency to overgeneralize 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.

51 As an example, in their study of the rural Northeast of Brazil, Harbison and Hanushek (1992) find that some of the "schools" lack an permanent physical structure, lack textbooks, and have teachers with a fourth grade education. In this situation, modest resources can in fact have readily discernible effects.

10. Implications for research

The recent rapid expansion of research into the economics of education in general and the determination of school performance in particular is clearly warranted by the value of the undertaking. Education remains one of the largest expenditures on government budgets around the world. Thus, information that could improve the efficiency of resource usage could have enormous impacts.

The review of existing data and studies demonstrates quite conclusively that the conclusions about the general inefficiency of resource usage are unlikely to be overturned by new data, by new methodologies, or the like. On the other hand, much remains to be learned about when and where resources are most productively used.

There appears to be enormous value in analyzing the implications and outcomes of alternative incentive structures. The observed inefficiency represented by the inconsistent relationship between resources and outcomes almost certainly reflects the nature of current incentives within education. But recognizing the importance of incentives, particularly ones related to student outcomes, is quite different from knowing exactly what to do to improve the situation.

A wide variety of altered incentives currently are being instituted in schools, ranging from varying use of choice to bonus pay schemes for teachers to overall school accountability plans with different rewards and sanctions. Consideration of the consequences of these plans offers an opportunity to investigate who schools, teachers, and students respond to varying incentives.

Focusing on such alternative analyses also highlights one of the key elements for future work. It is common within educational policy discussions to hear that "plan X worked where it was first tried but it could not be brought to scale". This shorthand discussion refers to the frequent observation that promising looking ideas cannot be transported to other places with the same success. The best way to think of this is that much of the research that moves into policy has not adequately identified the underlying causal structure. Considerable progress has recently been made in understanding alternative approaches to investigating causal relations, and pushing this work forward in terms of schooling outcomes is an obvious high-priority item.

References

Akerhielm, K. (1995). "Does class size matter?". Economics of Education Review 14 (3, September), 229–241.

Alderman, H., Behrman, J.R., Ross, D.R., Sabot, R. (1996). "The returns to endogenous human capital in Pakistan's rural wage labor market". Oxford Bulletin of Economics and Statistics 58, 29–55.

Altonji, J.G., Pierret, C.R. (2001). "Employer learning and statistical discrimination". Quarterly Journal of Economics 116 (1, February), 313–350.

Angrist, J.D., Lavy, V. (1997). "The effect of a change in language of instruction on the returns to schooling in Morocco". Journal of Labor Economics 15, S48–S76.

Angrist, J.D., Lavy, V. (1999). "Using Maimondides' rule to estimate the effect of class size on scholastic achievement". Quarterly Journal of Economics 114 (2, May), 533–575.

Barro, R.J., Lee, J.-W. (2001). "International data on educational attainment: Updates and implications". Oxford Economic Papers 53 (3, July), 541–563.

Barro, R.J., Sala-i-Martin, X. (2003). Economic Growth, second ed. MIT Press, Cambridge, MA.

Baumol, W.J. (1967). "Macroeconomics of unbalanced growth: The anatomy of urban crisis". American Economic Review 57 (3, June), 415–426.

Baumol, W.J., Bowen, W.G. (1965). "On the performing arts: The anatomy of their economic problems". American Economic Review 55 (May), 495–502.

Behrman, J.R., Kletzer, L.G., McPherson, M.S., 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 559 (September), 12–23.

Behrman, J.R., Ross, D., Sabot, R. (2006). "Improving the quality versus increasing the quantity of schooling: Estimates of rates of return from rural Pakistan", in preparation.

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

Bishop, J. (1989). "Is the test score decline responsible for the productivity growth decline?". American Economic Review 79 (1), 178–197.

Bishop, J. (1991). "Achievement, test scores, and relative wages". In: Kosters, M.H. (Ed.), Workers and Their Wages. AEI Press, Washington, DC, pp. 146–186.

Blackburn, M.L., Neumark, D. (1993). "Omitted-ability bias and the increase in the return to schooling". Journal of Labor Economics 11 (3, July), 521–544.

Blackburn, M.L., Neumark, D. (1995). "Are OLS estimates of the return to schooling biased downward? Another look". Review of Economics and Statistics 77 (2, May), 217–230.

Boissiere, M.X., Knight, J.B., Sabot, R.H. (1985). "Earnings, schooling, ability, and cognitive skills". American Economic Review 75 (5), 1016–1030.

Boozer, M.A., Rouse, C. (2001). "Intraschool variation in class size: Patterns and implications". Journal of Urban Economics 50 (1, July), 163–189.

Burgess, S., Propper, C., Slater, H., Wilson, D. (2005). "Who wins and who loses from school accountability? The distribution of educational gain in English secondary schools". Working Paper No. 05/128. Centre for Market and Public Organisation, University of Bristol (July).

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

Card, D. (1999). "Causal effect of education on earnings". In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics. North-Holland, Amsterdam, pp. 1801–1863.

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

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

Case, A., Deaton, A. (1999). "School inputs and educational outcomes in South Africa". Quarterly Journal of Economics 114 (3, August), 1047–1084.

Chaikind, S., Danielson, L.C., Brauen, M.L. (1993). "What do we know about the costs of special education? A selected review". Journal of Special Education 26 (4), 344–370.

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

Congressional Budget Office (1986). Trends in Educational Achievement. Congressional Budget Office, Washington, DC.

Dugan, D.J. (1976). "Scholastic achievement: Its determinants and effects in the education industry". In: Froomkin, J.T., Jamison, D.T., Radner, R. (Eds.), Education as an Industry. Ballinger, Cambridge, MA, pp. 53–83.

Dustmann, C., Rajah, N., van Soest, A. (2003). "Class size, education, and wages". Economic Journal 113 (485, February), F99–F120.

Finn, J.D., Achilles, C.M. (1990). "Answers and questions about class size: A statewide experiment". American Educational Research Journal 27 (3, Fall), 557–577.
Finnie, R., Meng, R. (2002). "Minorities, cognitive skills, and incomes of Canadians". Canadian Public Policy 28, 257–273.
Fuller, B. (1985). Raising School Quality in Developing Countries: What Investments Boost Learning. The World Bank, Washington, DC.
Glewwe, P. (1996). "The relevance of standard estimates of rates of return to schooling for educational policy: A critical assessment". Journal of Development Economics 51, 267–290.
Glewwe, P. (2002). "Schools and skills in developing countries: Education policies and socioeconomic outcomes". Journal of Economic Literature 40 (2, June), 436–482.
Green, D.A., Riddell, W.C. (2003). "Literacy and earnings: An investigation of the interaction of cognitive and unobserved skills in earnings generation". Labour Economics 10, 165–184.
Greenberg, D., McCall, J. (1974). "Teacher mobility and allocation". Journal of Human Resources 9 (4, Fall), 480–502.
Griliches, Z. (1974). "Errors in variables and other unobservables". Econometrica 42 (6, November), 971–998.
Grissmer, D.W., Kirby, S.N., Berends, M., Williamson, S. (1994a). Student Achievement and the Changing American Family. RAND Corporation, Santa Monica, CA.
Grissmer, D.W., Kirby, S.N., Berends, M., Williamson, S. (1994b). Student Achievement and the Changing Family. RAND Corporation, Santa Monica, CA.
Grogger, J.T., Eide, E. (1993). "Changes in college skills and the rise in the college wage premium". Journal of Human Resources 30 (2, Spring), 280–310.
Gundlach, E., Wossmann, L., Gmelin, J. (2001). "The decline of schooling productivity in OECD countries". Economic Journal 111 (May), C135–C147.
Hanushek, E.A. (1979). "Conceptual and empirical issues in the estimation of educational production functions". Journal of Human Resources 14 (3, Summer), 351–388.
Hanushek, E.A. (1986). "The economics of schooling: Production and efficiency in public schools". Journal of Economic Literature 24 (3, September), 1141–1177.
Hanushek, E.A. (1995). "Interpreting recent research on schooling in developing countries". World Bank Research Observer 10 (2, August), 227–246.
Hanushek, E.A. (1997a). "Assessing the effects of school resources on student performance: An update". Educational Evaluation and Policy Analysis 19 (2, Summer), 141–164.
Hanushek, E.A. (1997b). "The productivity collapse in schools". In: Fowler Jr., W.J. (Ed.), Developments in School Finance 1996. National Center for Education Statistics, Washington, DC, pp. 185–195.
Hanushek, E.A. (1999a). "The evidence on class size". In: Mayer, S.E., Peterson, P.E. (Eds.), Earning and Learning: How Schools Matter. Brookings Institution Press, Washington, DC, pp. 131–168.
Hanushek, E.A. (1999b). "Some findings from an independent investigation of the Tennessee STAR experiment and from other investigations of class size effects". Educational Evaluation and Policy Analysis 21 (2, Summer), 143–163.
Hanushek, E.A. (2002). "Publicly provided education". In: Auerbach, A.J., Feldstein, M. (Eds.), Handbook of Public Economics. Elsevier, Amsterdam, pp. 2045–2141.
Hanushek, E.A. (2003a). "The failure of input-based schooling policies". Economic Journal 113 (485, February), F64–F98.
Hanushek, E.A. (2003b). "The importance of school quality". In: Peterson, P.E. (Ed.), Our Schools and Our Future: Are We Still at Risk?. Hoover Institution Press, Stanford, CA, pp. 141–173.
Hanushek, E.A., Kain, J.F., O'Brien, D.M., Rivkin, S.G. (2005). "The market for teacher quality". Working Paper 11154. National Bureau of Economic Research (February).
Hanushek, E.A., Kain, J.F., Rivkin, S.G. (2002). "Inferring program effects for specialized populations: Does special education raise achievement for students with disabilities?". Review of Economics and Statistics 84 (4, November), 584–599.
Hanushek, E.A., Kain, J.F., Rivkin, S.G. (2004a). "Disruption versus Tiebout improvement: The costs and benefits of switching schools". Journal of Public Economics 88 (9–10), 1721–1746.

Hanushek, E.A., Kain, J.F., Rivkin, S.G. (2004b). “Why public schools lose teachers”. Journal of Human Resources 39 (2), 326–354.
Hanushek, E.A., Kim, D. (1995). “Schooling, labor force quality, and economic growth”. Working Paper 5399. National Bureau of Economic Research (December).
Hanushek, E.A., Kimko, D.D. (2000). “Schooling, labor force quality, and the growth of nations”. American Economic Review 90 (5, December), 1184–1208.
Hanushek, E.A., Luque, J.A. (2003). “Efficiency and equity in schools around the world”. Economics of Education Review 22 (5, August), 481–502.
Hanushek, E.A., Pace, R.R. (1995). “Who chooses to teach (and why)?”. Economics of Education Review 14 (2, June), 101–117.
Hanushek, E.A., Raymond, M.E. (2001). “The confusing world of educational accountability”. National Tax Journal 54 (2, June), 365–384.
Hanushek, E.A., Raymond, M.E. (2005). “Does school accountability lead to improved student performance?”. Journal of Policy Analysis and Management 24 (2, Spring), 297–327.
Hanushek, E.A., Rivkin, S.G. (1997). “Understanding the twentieth-century growth in U.S. school spending”. Journal of Human Resources 32 (1, Winter), 35–68.
Hanushek, E.A., Rivkin, S.G., Taylor, L.L. (1996). “Aggregation and the estimated effects of school resources”. Review of Economics and Statistics 78 (4, November), 611–627.
Hanushek, E.A. with others (1994). Making Schools Work: Improving Performance and Controlling Costs. Brookings Institution Press, Washington, DC.
Harbison, R.W., Hanushek, E.A. (1992). Educational Performance of the Poor: Lessons From Rural Northeast Brazil. Oxford University Press, New York.
Heckman, J.J., Layne-Farrar, A., Todd, P. (1996a). “Does measured school quality really matter? An examination of the earnings–quality relationship”. In: Burtless, G. (Ed.), Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success. Brookings Institution Press, Washington, DC, pp. 192–289.
Heckman, J.J., Layne-Farrar, A., Todd, P. (1996b). “Human capital pricing equations with an application to estimating the effect of schooling quality on earnings”. Review of Economics and Statistics 78 (4, November), 562–610.
Heckman, J.J., 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 83 (1, February), 1–12.
Heyneman, S.P., Loxley, W. (1983). “The effect of primary school quality on academic achievement across twenty-nine high and low income countries”. American Journal of Sociology 88 (May), 1162–1194.
Hoxby, C.M. (2000). “The effects of class size on student achievement: New evidence from population variation”. Quarterly Journal of Economics 115 (3, November), 1239–1285.
Jolliffe, D. (1998). “Skills, schooling, and household income in Ghana”. World Bank Economic Review 12, 81–104.
Knight, J.B., Sabot, R.H. (1998). Education, Productivity, and Inequality. Oxford University Press, New York.
Kremer, M. (2003). “Randomized evaluations of educational programs in developing countries: Some lessons”. American Economic Review 93 (2, May), 102–104.
Krueger, A.B. (1999). “Experimental estimates of education production functions”. Quarterly Journal of Economics 114 (2, May), 497–532.
Krueger, A.B. (2002). “Understanding the magnitude and effect of class size on student achievement”. In: Mishel, L., Rothstein, R. (Eds.), The Class Size Debate. Economic Policy Institute, Washington, DC, pp. 7–35.
Krueger, A.B. (2003). “Economic considerations and class size”. Economic Journal 113 (485, February), F34–F63.
Lazear, E.P. (2003). “Teacher incentives”. Swedish Economic Policy Review 10 (3), 179–214.
Manski, C.F., Wise, D.A. (1983). College Choice in America. Harvard University Press, Cambridge.
Mayer, S.E. (1997). What Money Can’t Buy: Family Income and Children’s Life Chances. Harvard University Press, Cambridge, MA.

McIntosh, S., Vignoles, A. (2001). "Measuring and assessing the impact of basic skills on labor market outcomes". Oxford Economic Papers 53, 453–481.

Mincer, J. (1970). "The distribution of labor incomes: A survey with special reference to the human capital approach". Journal of Economic Literature 8 (1, March), 1–26.

Mincer, J. (1974). Schooling Experience and Earnings. NBER, New York.

Moll, P.G. (1998). "Primary schooling, cognitive skills, and wage in South Africa". Economica 65, 263–284.

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

Mulligan, C.B. (1999). "Galton versus the human capital approach to inheritance". Journal of Political Economy 107 (6, December), S184–S224. Part 2.

Murnane, R.J. (1981). "Teacher mobility revisited". Journal of Human Resources 16 (1, Winter), 3–19.

Murnane, R.J., Willett, J.B., Braatz, M.J., 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 20 (4, August), 311–320.

Murnane, R.J., Willett, J.B., Duhaldeborde, Y., Tyler, J.H. (2000). "How important are the cognitive skills of teenagers in predicting subsequent earnings?". Journal of Policy Analysis and Management 19 (4, Fall), 547–568.

Murnane, R.J., Willett, J.B., Levy, F. (1995). "The growing importance of cognitive skills in wage determination". Review of Economics and Statistics 77 (2, May), 251–266.

Murphy, K.M., Welch, F. (1992). "The structure of wages". Quarterly Journal of Economics 107 (1, February), 285–326.

Neal, D.A., Johnson, W.R. (1996). "The role of pre-market factors in black–white differences". Journal of Political Economy 104 (5, October), 869–895.

O'Neill, J. (1990). "The role of human capital in earnings differences between black and white men". Journal of Economic Perspectives 4 (4, Fall), 25–46.

Organisation for Economic Co-operation and Development (2003). Education at a Glance: OECD Indicators 2003. Organisation for Economic Co-operation and Development, Paris, France.

Prais, S.J. (1996). "Class-size and learning: the Tennessee experiment – what follows?". Oxford Review of Education 22 (4), 399–414.

Psacharopoulos, G. (1994). "Returns to investment in education: A global update". World Development 22, 1325–1344.

Rivkin, S.G. (1995). "Black/white differences in schooling and employment". Journal of Human Resources 30 (4, Fall), 826–852.

Rivkin, S.G., Hanushek, E.A., Kain, J.F. (2005). "Teachers, schools, and academic achievement". Econometrica 73 (2, March), 417–458.

Scitovsky, T., Scitovsky, A. (1959). "What price economic progress?". Yale Law Review 49 (Autumn), 95–110.

Todd, P.E., Wolpin, K.I. (2003). "On the specification and estimation of the production function for cognitive achievement". Economic Journal 113 (485, February), F3–F33.

U.S. Department of Education (2002). Digest of Education Statistics, 2001. National Center for Education Statistics, Washington, DC.

Vignoles, A., Levacic, R., Walker, J., Machin, S., Reynolds, D. (2000). "The relationship between resource allocation and pupil attainment: A review". DP 02. Centre for the Economics of Education, London School of Economics and Political Science (November).

Vijverberg, W.P.M. (1999). "The impact of schooling and cognitive skills on income from non-farm self-employment". In: Glewwe, P. (Ed.), The Economics of School Quality Investments in Developing Countries: An Empirical Study of Ghana. St. Martin's Press (with University of Oxford), New York.

Wirtz, W. (1977). On Further Examination: Report of the Advisory Panel and the Scholastic Aptitude Test Score Decline. College Entrance Examination Board, New York.

Wößmann, L. (2001). "Why students in some countries do better". Education Matters 1 (2, Summer), 67–74.

Wößmann, L. (2003a). “Central exit exams and student achievement: International evidence”. In: Peterson, P.E., West, M.R. (Eds.), No Child Left Behind? The Politics and Practice of School Accountability. Brookings Institution Press, Washington, DC, pp. 292–323.

Wößmann, L. (2003b). “Schooling resources, educational institutions, and student performance: The international evidence”. Oxford Bulletin of Economics and Statistics 65 (2), 117–170.

Wößmann, L., West, M.R. (2006). “Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS”. European Economic Review 50 (3), 695–736.