Forthcoming: ANNALS of the American Academy of Political and Social Science

Testing and Accountability

Eric A. Hanushek*

November 2018

Abstract

Recent research highlights the importance of academic achievement as a determinant of economic well-being. Individual earnings, income growth in states, and national growth rates for GDP are each significantly determined by the population's cognitive skills, which in turn are proxied by scores on standardized achievement tests. This well-documented relationship between education and economic outcomes underscores the importance of using test information to guide both school policy and school operations. While test-based accountability has been controversial, scientific evidence about the economic value of school improvement and about the efficacy of various accountability approaches point to holding schools and teachers accountable for their contributions to the academic performance of students.

Keywords: school accountability, teacher value-added, standardized test scores, economic growth, individual incomes, cognitive skills

* Hoover Institution, Stanford University. Many helpful suggestions and comments came from Michael Feuer, Richard Murnane, and participants of the Workshop on Educational Assessment as Useful and Useable Evidence.

Introduction

Testing has long been a polarizing concept. Some educators and policy makers view testing as a necessary component of any effort to improve the quality of schools and to lessen inequality of opportunities. Others see the reliance on testing, and particularly test-based accountability, as narrowing the curriculum, leading teachers to substitute test preparation for deep instruction and more generally making teaching an undesirable occupation. These opposing viewpoints, which often take on an ideological flavor, guide much discussion of testing and accountability. This chapter introduces evidence supporting the argument that, especially from an economic standpoint, a focus on achievement as measured by standardized tests is appropriate. The evidence strongly supports policies directly related to these measures of performance – both of schools and teachers.

I begin with a consideration of the relationship between the skills measured by standardized tests and economic outcomes. My argument is that there is great economic value to skills as measured by tests. Although not the only thing to be concerned about, tests can provide a useful measuring rod for skills valued in the economy. I turn then to the case for using standardized tests for *accountability* in the educational system, including both school and teacher accountability.

Economic Value of Achievement

Existing research shows a very strong and consistent relationship between scores on common standardized tests and economic outcomes. This linkage with future economic well-being motivates the attention to alternative approaches to improving student performance. Research pinpoints key issues that are important when looking at the economic impacts of improved schools for individual states

Economic Growth of Nations

Economic growth determines the future economic wellbeing of nations. Economists have considered the process of economic growth for much of the last 100 years, but until recently little attention was given to large differences of growth rates across nations. Over the past quarter century, however, economists have linked the analysis of economic growth more closely to empirical observations of country differences, which has yielded insights relevant for government policy.

Early efforts at extracting fundamental factors underlying growth differences proved difficult, leading some researchers to abandon the attempt (see, for example, Levine and Renelt (1992), Levine and Zervos (1993)). Recent analyses suggest that a significant portion of the difficulty came from issues surrounding the measurement of skills of a nation's population.

Virtually all empirical studies of the long-run growth of countries have highlighted a role for human capital. The early literature focusing on cross-country differences in economic growth overwhelmingly employed measures related to school attainment, or years of schooling, to test the human capital aspects of growth models. This work tended to find a significant positive association between quantitative measures of schooling and economic growth (see, for example, Sala-i-Martin, Doppelhofer, and Miller (2004)). However, the overall validity and reliability of the empirical analysis remained open to question (Pritchett (2006).

Conceptually, average years of schooling is an incomplete and potentially misleading measure of education when comparing different countries. It implicitly assumes that a year of schooling delivers the

same increase in knowledge and skills regardless of the education system. For example, a year of schooling in Peru is assumed to create the same increase in productive human capital as a year of schooling in Japan. Additionally, growth formulations relying exclusively on measures of school attainment assume that formal schooling is the only source of education and that variations in non-school factors have negligible effects on education outcomes and skills. This neglect of cross-country differences in the quality of schools and in the strength of family, health, and other influences is a major drawback in such research.

An attractive alternative is measuring skills of the population in different countries by the cognitive skills found in international achievement tests. As laid out in prior work (Hanushek and Kimko (2000), Hanushek and Woessmann (2015a)), the prior validity concerns with growth analyses can be substantially alleviated once skills are correctly measured. International achievement test scores can be thought of as measures of human capital differences, regardless of the source of such differences. Indeed, once long run growth rates across countries are related to international test scores, three-quarters of the cross-country variation in growth rates can be explained by differences in scores on international math and science tests.[1] Moreover, there is reason to believe that this relationship is causal – i.e., if cognitive skills can be raised, growth rates will increase (e.g., Hanushek and Woessmann (2012). These estimates indicate that just increasing school attainment without also increasing the amount of learning has no impact. In other words, just getting students through more schooling without ensuring high levels of learning is not an effective policy.

The historical impact on economic growth of differences in test scores is large. One easy way to see the importance of cognitive skills is to project the economic value of school improvement on the U.S. economy (Hanushek, Peterson, and Woessmann (2013), Hanushek and Woessmann (2015b)). Consider, for example, the estimated impact of bringing just the bottom of the achievement distribution up to a basic skill level – i.e., a policy similar to the ideas behind NCLB except stretched out over a fifteen year period in the future. Hanushek and Woessmann (2015b) estimate that, according to historical growth patterns, this would lead to average GDP levels that were 3.3 percent higher across the remainder of the century when compared to expected GDP levels with current skill levels. Such increases would be sufficient to deal with, for example, the financial deficits of the Social Security program and the Medicare program if the added resources were so invested.

As an alternative (which is relevant for the discussion below), consider the economic impact of bringing the achievement of U.S. students up to the level of Canadian students, whose performance on PISA is almost one-half of one standard deviation higher than U.S. students. Other things equal, reaching the Canadian level of achievement would, by historical growth relationships, yield an average 20 percent boost in every worker's paycheck for the rest of the century (Hanushek, Peterson, and Woessmann (2013).

The challenge to the United States is clear from these growth estimates. Currently, U.S. students rank slightly below the average developed country in the OECD. While the U.S. economy in the past has done better than would be expected by student performance, this outcome cannot be counted

[1] International tests of math were first conducted in 1964. These early tests evolved into regular assessments by the Programme for International Student Assessment (PISA) and the Trends in Mathematics and Science Study (TIMSS); see Hanushek and Woessmann (2011), and Singer, Brown, and Chudowsky (2018). The estimation of growth models using these tests is described in Hanushek and Woessmann (2015a).

on in the future (Hanushek, Peterson, and Woessmann (2013)). In simplest terms, the evidence suggests that future wellbeing of U.S. society and the future position of the U.S. in the world are highly dependent on improving student achievement and implicitly the quality of U.S. schools.

Economic Growth of States

Education policy has long been the provenance of the U.S. states rather than the federal government. Although historians argue about the origins and virtues of this situation (see, e.g., Vinovskis in this volume), the federal system sets up weakened incentives for state investment. Given the high levels of mobility in the U.S., where the work location of somebody might be very different from where the person grew up and went to school, states do not directly experience all of the results of their school systems. Therefore, while improving schools might be in the national interest, individual states might benefit less and thus might not have strong incentives to invest in better schools. The tension in America between centralized and decentralized education policy has been a pivotal policy issue for decades.

How schools affect state-level measures of economic output is a high priority concern for policy makers (and researchers), especially in the light of the most recent reauthorization of federal education law: the Every Student Succeeds Act (ESSA), which has shifted policy control again back toward the state level. The evidence is compelling. In a series of studies, Hanushek, Ruhose, and Woessmann (2016, 2017a, 2017b) show that *economic growth of individual states, just like nations, is dependent on the quality of the labor force as measured by standardized tests.* Moreover, the relationship between worker skills and growth at the state level is virtually identical to that found internationally.

Because a majority of students educated in a given state remain in the state when entering the labor force, even with migration, it pays for each state to invest in improved school quality. But since the labor force in each state is comprised of both locally educated workers and workers educated in other states, the largest gains come when all states improve their school quality, as opposed to a single state.

Again, measuring quality differences with standardized test scores, rather than relying just on attainment statistics, is important. For example, the data from the National Assessment of Educational Progress (NAEP), a trusted source of information on the condition of student learning in America (see chapter by Fahle et al, in this volume), suggest that an average 8th grader in the lowest performing state on mathematics in 2017 (Alabama) was achieving at the level of a 5th grader in Massachusetts. Increasing schooling levels (enrollment and attainment) without addressing *quality* is unlikely to yield desired economic results.

Individual Incomes

The previous sections focused on the effects of improved school quality on aggregate economic gains at the state and national level. More research has focused on the relationship between education and individual earnings. Innumerable economic studies show that school attainment affects earnings and income. These studies, pioneered by Jacob Mincer (1970, 1974), showed that economic success depends heavily on schooling. Nonetheless, they suffer from many of the same problems described in the previous aggregate studies. In particular, they ignore quality differences in schools, and they ignore sources of skills outside of schools. As demonstrated by the landmark "Equality of Educational Opportunity" report, commonly known as "the Coleman Report," families are very important, as are

peers in schools, neighborhood influences, and more (Coleman et al. (1966). An extensive body of research documents the multiplicity of inputs in educational production (e.g., Hanushek (2002).

The alternative, as with the aggregate studies, is to use measured skill from standardized tests to capture the totality of individual skills from families, schools, and other influences. This approach also relates the research more directly to educational policy. It has not been pursued extensively in the past, largely because few data sources combine information on both skills and individual earnings. (For the effects of so-called "noncognitive" skills, see, e.g., Cunha and Heckman (2008) and Deming (2017)).

Recent international data provide the ability to estimate the economic value to individuals of higher educational achievement. The OECD surveyed random samples of adults age 15-65 across 32 countries in the Program for International Assessment of Adult Competencies (PIAAC). This survey contained information on backgrounds of individuals and their labor market experiences along with giving them a series of standardized tests (see Hanushek et al. (2015, 2017)).

Hanushek et al. (2015, 2017) estimate the economic returns to greater individual skills. The U.S. has high returns, exceeding those found in almost all of the developed countries that are observed. These returns imply that an individual in the U.S. who has skills as defined and measured on international comparative assessments that are one standard deviation above the mean will on average see 28 percent higher earnings across the lifetime compared to the median person. But these high returns also imply that somebody one standard deviation below the mean can expect 28 percent lower earnings across a lifetime. In other words, the U.S. provides high rewards to acquired skills as measured by standardized tests, but it also severely punishes those with low skills. These estimates are consistent with research about the growing importance of basic cognitive skills from a quarter of a century ago (Murnane, Willett, and Levy (1995)).

In sum, a wide range of evidence shows the substantial economic value of improved cognitive skills. This in turn suggests that student test scores merit policy attention.

Test-based School Accountability

While student test scores may be good indicators of skills that are ultimately valued in the economy, this does not by itself indicate how test scores might enter into policy. The most contentious aspect of testing revolves around the extent to which test performance enters into policy actions and into school system operations.

Throughout the 1990s an increasing number of states introduced formal annual testing across multiple grades and specified how these would enter into the evaluation of schools and potential policy decisions. Figure 1 shows the expansion of state accountability systems and indicates that most states had already put in place their own systems by the time the federal government became involved with the No Child Left Behind Act of 2001 (NCLB).

**********Figure 1 about here **********

State accountability systems

Understanding the impact of test-based accountability systems is challenging. First, the use of tests may not be independent of different policy options. For example, a state may reduce the extent of regulations and may grant more decision making autonomy to local districts while introducing a system of test-based accountability in order to monitor the performance of individual districts. Second, it is unclear what a good comparison group might be since accountability systems are introduced to all schools in a state simultaneously and states differ from one another in many dimensions.

One approach to evaluating the impact of test-based accountability systems is to analyze whether student performance changes when a state introduces a particular accountability system. Hanushek and Raymond (2005), looking at state outcomes before the introduction of NCLB, find that consequential accountability (introducing rewards and sanctions based on student test performance) yields higher student performance (as measured by state NAEP scores). Simply reporting results has no effect. Thus, if an accountability system is to provide incentives for improving student outcomes, rewards or punishment should be attached to school performance.

Carnoy and Loeb (2002) categorized the strength of state accountability standards. They found that stronger accountability regimes are associated with significantly better student achievement. They also find that strong accountability does not result in lower graduation rates or more grade retention. Dee and Jacob (2011) looked directly at the effects of NCLB. They found strong impacts on math achievement but not necessarily on reading. They also found that, even though NCLB was aimed at bottom performers, gains were found at both the top and the bottom of the score distribution. Figlio and Loeb (2011) summarized the impact of accountability more broadly and concluded that the evidence strongly indicates positive impacts of accountability as generally found in the U.S. They also consider some of the unintended consequences to accountability, from which they conclude that indeed care must be taken to anticipate and allow for various other impacts. A less sanguine conclusion about test-based accountability was reached by a National Research Council report (Hout and Elliott (2011); for rebuttal, see Hanushek (2012).

The structure of NCLB-like incentives also warrants consideration. NCLB directed states to decide what education should be produced by setting up state outcome standards and then to test and monitor whether student performance meets these standards. States had to set a time path of performance that would lead to all students being proficient by 2014. In turn, the federal government set policies toward altered school operations if any school failed to meet its established performance goals. These policies included mandated remediation, expanded choice, and even closing schools. How states responded to these incentives, though, requires close attention to consequences of accountability systems: as Linn noted, “substantial differences between the accountability requirements of many state systems and NCLB still have resulted in mixed messages regarding the performance of schools…” (Linn (2005)).

In other words, the states specified *what* to produce while the federal government set *how* this was to be produced. But, this is 180 degrees off what one might think was the optimal answer. The states have difficulty in understanding both the demands for skills when many students will work outside of the state and when much of the competition comes from workers in other countries with different preparation and levels of performance. At the same time, the federal government, which is far

removed from knowledge of either the educational needs or the educational capacities at the school, is quite unprepared to prescribe how achievement is best produced.

Given these inverted roles of states and the federal government, it is notable that NCLB still appears to have led to student improvement. Clearly the details of any accountability system are important, and it remains speculative what might be achieved by a better designed system. The current version of federal accountability law (ESSA), reverts to a pre-NCLB position in which the states are primarily responsible for determining both the goals of their respective school systems and determining how best to achieve those goals. ESSA became law in 2015 but did not take full effect until 2018, so evidence on its impact has yet to become available.

International Evidence

Additional evidence on testing and accountability comes from looking across countries. Bergbauer, Hanushek, and Woessmann (2018) relate changes in testing and accountability policies to changes in PISA scores between 2000-2015. This work uses the country panel structure of the data to identify the impact of various testing and accountability measures. (Causal inferences from such research is of course a concern; see, e.g., Feuer (2012); Singer, Braun, and Chudowsky (2018); and Braun and Singer in this volume). They find that standardized tests used for external comparisons have a significant and positive impact on student performance. Moreover, testing both to evaluate schools and to evaluate individual students (through exit exams) have significant impacts, with the effects of school accountability being somewhat larger. On the other hand, nonstandardized testing and evaluations, including inspectorates, have no significant impact on student performance.

These results reinforce the argument that consequential accountability improves performance. Testing that is not comparable across schools or that just produces report card information has less if any impact.

Value-added for Teachers

Before there was widespread state accountability, student achievement began to be related to teacher performance – and here is where the most controversial use of student testing is found. Hanushek (1971), relying on data from a single large school system, showed that *learning growth* across classrooms within the district varied widely. This analysis also showed what proves to be another consistent finding: factors that are typically used in setting teacher pay levels – such as graduate education and prior experience – are not strongly related to student learning gains. This study was followed by analysis in another city, which found similar results (Murnane (1975)). These studies introduced the idea of statistically separating the impact of teachers on student performance from other factors such as families in order to assess the "value-added" of individual teachers (a terminology popularized by Sanders and Horn (1994)). Related studies of teacher value-added followed, as student performance data became more available. Hanushek and Rivkin (2010) reviewed estimates of variation in teacher value-added from different samples of students and teachers and showed considerable variation in effectiveness of teachers across U.S. schools. This overall value-added approach was linked to state accountability testing and commercialized in Tennessee (Sanders and Horn (1994)). They developed a different form of value-added estimation and provided teacher-by-teacher reports to

principals in Tennessee schools. Other school districts, such as Dallas, began developing teacher evaluation information from various forms of value-added modeling (Mendro et al. (1998)).

The estimation of teacher value-added in different circumstances may look like many other lines of scholarly investigation. The attention to and reaction to such estimation changed dramatically when estimates of teacher value-added – how much a teacher contributed to measureable changes in student achievement scores – began to be used for personnel decisions. As the idea of judging teachers on their effectiveness in the classroom grew increasingly interesting to policy makers, so too did the resistance by some teachers and school leaders, which led to calls for greater scrutiny and research (e.g., Braun, Chudowsky, and Koenig (2010), Hanushek and Rivkin (2012), Haertel (2013), Jackson, Rockoff, and Staiger (2014), and Koedel, Mihaly, and Rockoff (2015)).

Because of the intense on-going research in this area, the focus of attention and the research conclusions have varied. Koedel, Mihaly, and Rockoff (2015) provide a clear and reasonable summary of the current state of evidence. "The most important result for which consistent evidence has emerged in research is that students in K-12 schools stand to gain substantially from policies that incorporate information about value-added into personnel decisions for teachers." (p. 192) And, in addressing the technical issues, they conclude: "the research studies that have employed the strongest experimental and quasi-experimental designs to date indicate that the scope for bias in estimates of teacher value-added from standard models is quite small" (p. 192). Other work reaches basically similar conclusions (e.g.,Chetty, Friedman, and Rockoff (2016), Rothstein (2017).

Few people advocate using value-added measures exclusively in evaluations, but incorporating such measures seems like an obvious policy decision. Indeed, these concepts have been introduced quite broadly into state policy. In reviewing state policies, the National Council on Teacher Quality (2017) finds that by 2017 39 states require teacher evaluations that include objective measures of student achievement growth, although the exact form and weight placed on these varies widely.

Economic Value of Teacher Quality

The existing research provides very consistent evidence about the observed variation in teacher effectiveness (from value-added analyses) which yields direct information on how student achievement will vary across teachers of differing abilities. Matched with that, the estimation of the labor market returns to skills for individuals and of the impact of achievement on macroeconomic growth provides information on the future income gains to students that can be expected to follow any changes in achievement.

Hanushek (2011a, 2011b) put these two strands of analysis together to estimate the economic value of teacher quality. Figure 2 shows the expected gains across a class of students when compared to the outcomes expected from an average teacher. The economic gains from greater skills accrue throughout a student's lifetime, and the figure sums the earnings gain at each point through the lifetime. Gains that come early are weighted more than those in the more distant future, providing the present value of lifetime earnings for students in 2010 dollars (the estimates are discounted at three percent, which can be interpreted as the amount of money invested in a savings account with three percent interest that would allow for reproducing the entire future earnings gains. See Hanushek (2011a)). Finally, the economic gains from any teacher depend on how many students are affected by the teacher, and the horizontal axis shows the FTE's relevant to the teacher.

**********Figure 2 about here **********

The figure shows the remarkable impact a teacher can have on his or her class. A 75th percentile teacher – i.e., one who is very effective – with a class of 30 students produces future earnings gains of $400,000 compared to an average teacher *each year*. As shown, a 60th percentile teacher – a somewhat above average teacher – also produces noticeable gains in earnings for her class. Moreover, the figure shows that ineffective teachers – those who produce low learning gains among students – can do considerable harm. A 10th percentile teacher annually subtracts $800,000 from the future earnings of her class of 30 students when compared to an average teacher. These estimated economic impacts are large enough to warrant serious policy attention.

These estimates are confirmed by Chetty, Friedman, and Rockoff (2014), who use an entirely different approach to derive the economic value of high teacher quality. Their analysis matches the school experiences of individual students to their subsequent tax records. They relate the value-added of elementary and middle school teachers to the future income of the exact students in the class. While they observe students just for the early part of their working careers, their estimates align well with the narrative of Figure 2. Moreover, they show that effective teachers also have observable impacts on college attendance, early childbirth, and other important outcomes.

A different way of assessing the economic impact of teachers is to relate overall effects to aggregate achievement, thus permitting a linkage to both state and national economic growth. Think of taking all of the teachers in the U.S. and replacing the least effective with an average one, and so forth across the distribution of teacher classroom effectiveness. The closer teachers are to each other, the lesser will be the impact of replacing the least effective; indeed, if all teachers were equally effective in terms of student learning gains, this exercise would have no overall impact. Figure 3 displays the plausible range of impacts on overall U.S. performance based on a lower bound on the variation in teacher effectiveness (dashed line) and an upper bound (solid line) that is consistent with existing research on teacher value-added (Hanushek and Rivkin (2010)).

********** Figure 3 about here **********

This figure, similar to the prior estimates of individual earnings impacts, illustrates the large effect that ineffective teachers have on students and, by implication, on the nation. From Figure 3, replacing the bottom 6-9 percent of teachers with teachers performing at the average level would bring U.S. student performance up to the level of Canada. And, by the previous estimates of the impact of that on economic growth, this would translate on average into 20 percent higher incomes for every worker in the U.S. over the remainder of the century (Hanushek, Peterson, and Woessmann (2013). According to estimates at the upper bound of the estimated variation in teacher effectiveness, it might even get us to the level of Finland in terms of achievement. Again, such gains justify paying more attention to the effectiveness of teachers.

Of course knowing that teachers of varying effectiveness bring about different levels of student performance does not say anything about *how* one can improve the stock of teachers to get higher aggregate performance. It does not, for example, say that the best way to achieve better outcomes is to manage the schools and the teacher force in terms of test scores of students. Prior evidence on the impacts of test-based accountability systems indicated some potential for using student outcome information. But does this carry through to individual teacher policies?

Use of Teacher Accountability and Evaluation

Actually using information about teacher effectiveness is not the norm in U.S. schools (or those in other countries for that matter). Few school systems have significantly integrated teacher evaluation into their personnel systems. One is Washington, DC. Another is Dallas, Texas.

Washington, DC

After a very acrimonious contract negotiation, Washington, DC, moved to the IMPACT evaluation system in 2010. This system involves both very large economic rewards for top-rated teachers and dismissal for bottom-rated teachers. The evaluation system, which has changed somewhat since its inception, combines estimates of each teacher's value-added with a rigorous observational rating. Because of the limited testing by grade and subject, less than a quarter of DC teachers actually have a value-added component – implying that the majority of assessment involves outside raters using a structured rubric of teacher classroom performance.

Dee and Wyckoff (2015, 2017) provide a direct evaluation of the system. (See also Office of the District of Columbia Auditor (2014)). The salary recognition for highly effective teachers (about 14 percent of DC teachers) varied but could reach an increase in base pay of $25,000. At the other end of the spectrum, the least effective teachers were dismissed or induced to leave the system before any dismissal action. The resulting responsiveness of teachers to the incentives – both in terms of improvements in class room performance and in selective turnover – has led to a significant increase in overall teacher quality in Washington. After the introduction of IMPACT, gains in student performance by Washington students taking the NAEP tests outpaced those in all other large city districts that participate in NAEP.

Dallas, Texas

In 2015, after extensive study and development, the Dallas Independent School District (DISD) introduced a radically altered personnel system affecting both principals and teachers (see https://tei.dallasisd.org/ and https://www.dallasisd.org/Page/41972). Instead of relying on standard experience and education pay scales with the possibility of bonuses on top, the modified system linked pay directly to measured effectiveness.

Like many policies that are introduced for entire districts, states, and nations, it is difficult to find an adequate comparison that can be used to evaluate the overall effort. Nonetheless, there are signs that parts of the policies are indeed having clear effects. In the ACE (Accelerating Campus Excellence) program, a system of quality-based "combat pay" for teachers led the worst schools in Dallas to get very high quality principals and teachers. As a result, student scores improved dramatically (Morgan et al. (2018)).

Both DC and DISD systems have their detractors. My conclusion is that using measures of teacher effectiveness in personnel decisions appears to have significant potential impacts on student performance. It is not that these "demonstration" systems as currently designed are necessarily the best possible way to proceed, but it is remarkable on the other side that the vast majority of the school systems in the country set teacher compensation in ways unrelated to the effectiveness on teachers in the classroom.

Policy Options

In policy debates, considerable attention is rightfully given to alternative approaches to policies directly related to student test performance. This search is motivated by a desire to find approaches that might achieve the ends of improving the skills of American students (and workers) without disrupting the current functioning of schools. It is useful to touch on some of the issues in order to provide empirical perspective on the alternatives.

One solution would be getting superior personnel into the schools in the first place. This objective could be accomplished by improving the preparation and/or selection of teachers or by attracting highly qualified people away from other occupations and into teaching. There is international evidence suggesting that smarter teachers – teachers who themselves have higher measured cognitive skills – are more effective and that increases in alternative occupational choices for potential teachers have harmed schools (Hanushek, Piopiunik, and Wiederhold (forthcoming).

The evidence on teacher preparation programs is not very supportive of the potential for focusing on entry. Different sources of preparation do not seem to be systematically superior in terms of teacher effectiveness (Boyd et al. (2006); Kane, Rockoff, and Staiger (2008); also Feuer et al, 2013). Second, there is little existing evidence that just expanding the pool of potential teachers from which to choose is very effective.

An alternative might be to take the existing stock of teachers and make them better through professional development. Again, while some evaluations suggest positive impacts of specific programs, there is less reason to believe that we know how to implement effective programs at scale (Garet et al. (2008), Garet et al. (2011)).

Finally, the alternative of evaluation systems that do not use test-based information is not encouraging. Weisberg et al. (2009) show that traditional evaluations do not provide usable information. Further, the resulting personnel policies have not led to great results (TNTP (2012). Specifically, the standard evaluation/personnel policies do not result in retaining the best teachers.

Again the disappointment of these alternative policies does not establish the case for widespread test-based accountability for teachers. Suitable student test information (for evaluating value-added) is available for just a subset of teachers; and test information does not capture the range of factors that might appropriately enter into teacher evaluations. A composite evaluation and management approach seems reasonable (Kane et al. (2013)).

More than anything, however, the evidence establishes a strong case for judging the efficacy of any teacher policy on the basis of measured student performance. And this necessarily means that, if we are interested in student outcomes, there is no substitute for focusing on student outcomes as measured by tests.

Value-Added of Principals

Much of the research and policy attention has focused on teachers and has ignored the role of the principal. Research efforts in this direction have proven to be much more difficult than the estimation of teacher effectiveness. First, although there are many ways a principal might affect school performance – setting educational standards, mentoring teachers, selecting the teaching force – there are generally no data on these separate activities. Second, a principal will generally inherit a majority of the teacher force

on entry to the school, and changes will take time. Third, special circumstances might affect results surrounding the turnover of principals (key information used in assessing value-added). The research is not completely developed at this time; see, for example, Branch, Hanushek, and Rivkin (2012) and Grissom, Kalogrides, and Loeb (2015). Nonetheless, the evidence suggests an important role for principals, indicating then that understanding the impact of school management and leadership is a key but largely open issue.

Complementary Research

A side benefit of state accountability systems is the ready availability of information on student outcomes that can be used for research purposes. As was first demonstrated by various analyses done at the Texas Schools Project of the University of Texas at Dallas, it is possible to link student achievement data over time in order to track student learning (see https://www.utdallas.edu/research/tsp-erc/.) This development was followed by work in other states, starting with New York, Florida, and North Carolina, and expanding today to a significant number of states that work with researchers who are interested in analyzing student outcomes.

This complementary use of student achievement data has expanded quantitative research into the determinants of student outcomes. As these data are available over longer periods of time and are linked with other sources, a richer and more nuanced view of school outcomes and school performance is becoming available. Recent studies have linked primary and middle school performance to later life incomes, college attendance, and teen pregnancy (Chetty, Friedman, and Rockoff (2014)). Other work investigates the relationship between school quality and subsequent criminal behavior (Deming (2011).

It is worth noting that access to data has varied across individual states, partly dictated by each state's interpretation of the requirements for ensuring the confidentiality of student information as required by federal statute in Family Educational Rights and Privacy Act of (FERPA) and partly dictated by individual state views on permitting and supporting research on their schools.

Some Conclusions

The controversy surrounding testing and accountability has obfuscated the national importance of improving the skills of the population. The future economic well-being of the U.S. is, by historical evidence, highly dependent on having a skilled workforce. Many factors enter into the current achievement levels of students – mediocre by international standards – but the obvious place for bringing about improvement is the schools. Research also suggests that improving the stock of highly effective teachers can radically transform overall achievement.

The evidence indicates that test-based school accountability can lead to higher achievement. The experience with test-based teacher and principal accountability is much more limited but also suggests positive results.

The best design for school or personnel-based accountability is likely to vary, based on both the demands and capacities of different schools. This is not, however, an argument against moving forward,

because the stakes are simply too high to ignore. A sensible policy process would include local experimentation and evaluation, not waiting for the "one best system" to be implemented (Tyack (1974)).

References

Bergbauer, Annika B., Eric A. Hanushek, and Ludger Woessmann. 2018. "Testing." NBER Working Paper 24836. Cambridge, MA: National Bureau of Economic Research (July).

Boyd, Don, Pam Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff. 2006. "How changes in entry requirements alter the teacher workforce and affect student achievement." *Education Finance and Policy* 1, no. 2 (Spring): 176-216.

Branch, Gregory F., Eric A. Hanushek, and Steven G. Rivkin. 2012. "Estimating the Effect of Leaders on Public Sector Productivity: The Case of School Principals." NBER Working Paper W17803. Cambridge, MA: National Bureau of Economic Research (January).

Carnoy, Martin, and Susanna Loeb. 2002. "Does external accountability affect student outcomes? A cross-state analysis." *Educational Evaluation and Policy Analysis* 24, no. 4 (Winter): 305-331.

Chetty, Raj, John N. Friedman, and Jonah Rockoff. 2014. "Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood." *American Economic Review* 104, no. 9 (September): 2633-2679.

Coleman, James S., Ernest Q. Campbell, Carol J. Hobson, James McPartland, Alexander M. Mood, Frederic D. Weinfeld, and Robert L. York. 1966. *Equality of educational opportunity*. Washington, D.C.: U.S. Government Printing Office.

Dee, Thomas S., and Brian A. Jacob. 2011. "The impact of No Child Left Behind on student achievement." *Journal of Policy Analysis and Management* 30, no. 3: 418-446.

Dee, Thomas S., and James Wyckoff. 2015. "Incentives, selection, and teacher performance: Evidence from IMPACT." *Journal of Policy Analysis and Management* 34, no. 2 (Spring): 267-297.

Dee, Thomas S., and James Wyckoff. 2017. "A Lasting Impact: High-stakes teacher evaluations drive student success in Washington, D.C." *Education Next* 17, no. 4 (Fall): 58-66.

Deming, David J. 2011. "Better Schools, Less Crime?" *The Quarterly Journal of Economics* 126, no. 4 (November 1, 2011): 2063-2115.

Figlio, David, and Susanna Loeb. 2011. "School accountability." In *Handbook of the Economics of Education, Vol. 3*, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann. Amsterdam: North Holland: 383-421.

Garet, Michael S., Stephanie Cronen, Marian Eaton, Anja Kurki, Meredith Ludwig, Wehmah Jones, Kazuaki Uekawa, Audrey Falk, Howard S. Bloom, Fred Doolittle, Pei Zhu, and Laura Sztejnberg. 2008. *The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences. Washington, DC: U.S. Department of Education (September).

Garet, Michael S., Andrew J. Wayne, Fran Stancavage, James Taylor, Marian Eaton, Kirk Walters, Mengli Song, Seth Brown, Steven Hurlburt, Pei Zhu, Susan Sepanik, and Fred Doolittle. 2011. *Middle school mathematics professional development impact study: Findings after the second year of implementation*, NCEE 2011-4024. Washington, DC: Institute of Education Sciences (April).

Grissom, Jason A., Demetra Kalogrides, and Susanna Loeb. 2015. "Using Student Test Scores to Measure Principal Performance." *Educational Evaluation and Policy Analysis* 37, no. 1 (March): 3-28.

Hanushek, Eric A. 1971. "Teacher characteristics and gains in student achievement: Estimation using micro data." *American Economic Review* 60, no. 2 (May): 280-288.

Hanushek, Eric A. 2002. "Publicly provided education." In *Handbook of Public Economics, Vol. 4*, edited by Alan J. Auerbach and Martin Feldstein. Amsterdam: North Holland: 2045-2141.

Hanushek, Eric A. 2011a. "The economic value of higher teacher quality." *Economics of Education Review* 30, no. 3 (June): 466-479.

Hanushek, Eric A. 2011b. "Valuing teachers: How much is a good teacher worth?" *Education Next* 11, no. 3 (Summer).
Hanushek, Eric A., and Dennis D. Kimko. 2000. "Schooling, labor force quality, and the growth of nations." *American Economic Review* 90, no. 5 (December): 1184-1208.
Hanushek, Eric A., Paul E. Peterson, and Ludger Woessmann. 2013. *Endangering prosperity: A global view of the American school*. Washington, DC: Brookings Institution Press.
Hanushek, Eric A., Marc Piopiunik, and Simon Wiederhold. forthcoming. "The value of smarter teachers: International evidence on teacher cognitive skills and student performance." *Journal of Human Resources*.
Hanushek, Eric A., and Margaret E. Raymond. 2005. "Does school accountability lead to improved student performance?" *Journal of Policy Analysis and Management* 24, no. 2: 297-327.
Hanushek, Eric A., and Steven G. Rivkin. 2010. "Generalizations about using value-added measures of teacher quality." *American Economic Review* 100, no. 2 (May): 267-271.
Hanushek, Eric A., Jens Ruhose, and Ludger Woessmann. 2016. "It pays to improve school quality: States that boost student achievement could reap large economic gains." *Education Next* 16, no. 3 (Summer): 16-24.
Hanushek, Eric A., Jens Ruhose, and Ludger Woessmann. 2017a. "Economic gains from educational reform by US States." *Journal of Human Capital* 11, no. 4 (Winter): 447-486.
Hanushek, Eric A., Jens Ruhose, and Ludger Woessmann. 2017b. "Knowledge capital and aggregate income differences: Development accounting for U.S. states." *American Economic Journal: Macroeconomics* 9, no. 4 (October): 184-224.
Hanushek, Eric A., Guido Schwerdt, Simon Wiederhold, and Ludger Woessmann. 2015. "Returns to skills around the world: Evidence from PIAAC." *European Economic Review* 73: 103-130.
Hanushek, Eric A., Guido Schwerdt, Simon Wiederhold, and Ludger Woessmann. 2017. "Coping with change: International differences in the returns to skills." *Economic Letters* 153(April): 15-19.
Hanushek, Eric A., and Ludger Woessmann. 2011. "The economics of international differences in educational achievement." In *Handbook of the Economics of Education, Vol. 3*, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann. Amsterdam: North Holland: 89-200.
Hanushek, Eric A., and Ludger Woessmann. 2012. "Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation." *Journal of Economic Growth* 17, no. 4: 267-321.
Hanushek, Eric A., and Ludger Woessmann. 2015a. *The knowledge capital of nations: Education and the economics of growth*. Cambridge, MA: MIT Press.
Hanushek, Eric A., and Ludger Woessmann. 2015b. *Universal basic skills: What countries stand to gain*. Paris: Organisation for Economic Co-operation and Development.
Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. 2013. *Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment*. MET Project: Bill and Melinda Gates Foundation (January).
Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger. 2008. "What does certification tell us about teacher effectiveness? Evidence from New York City." *Economics of Education Review* 27, no. 6 (December): 615-631.
Koedel, Cory, Kata Mihaly, and Jonah E. Rockoff. 2015. "Value-added modeling: A review." *Economics of Education Review* 47: 180-195.
Levine, Ross, and David Renelt. 1992. "A sensitivity analysis of cross-country growth regressions." *American Economic Review* 82, no. 4 (September): 942-963.
Levine, Ross, and Sara J. Zervos. 1993. "What we have learned about policy and growth from cross-country regressions." *American Economic Review* 83, no. 2 (May): 426-430.

Linn, Robert L. 2005. "Conflicting Demands of No Child Left Behind and State Systems: Mixed Messages about School Performance." *Education Policy Analysis Archives* 13, no. 33 (June 28). [accessed Noveber 22, 2018]

Mendro, Robert L., Heather R. Jordan, Elvia Gomez, Mark C. Anderson, and Karen L. Bembry. 1998. "An Application of Multiple Linear Regression in Determining Longitudinal Teacher Effectiveness." Paper presented at *1998 Annual Meeting of the American Educational Research Association*, April 1998, at San Diego, CA.

Mincer, Jacob. 1970. "The distribution of labor incomes: a survey with special reference to the human capital approach." *Journal of Economic Literature* 8, no. 1 (March): 1-26.

Mincer, Jacob. 1974. *Schooling, experience, and earnings*. New York: NBER.

Morgan, Andrew, Minh Thac Nguyen, Eric A. Hanushek, Ben Ost, and Steve G. Rivkin. 2018 of Conference. "Getting effective educators in schools serving disadvantaged students." Paper presented at *Association for Public Policy Analysis and Management*, at Washington, DC.

Murnane, Richard J. 1975. *Impact of school resources on the learning of inner city children*. Cambridge, MA: Ballinger.

Murnane, Richard J., John B. Willett, and Frank Levy. 1995. "The growing importance of cognitive skills in wage determination." *Review of Economics and Statistics* 77, no. 2 (May): 251-266.

National Council on Teacher Quality. 2017. *State teacher policy yearbook, 20157*. Washington: National Council on Teacher Quality.

Office of the District of Columbia Auditor. 2014. *Trends in Teacher Effectiveness in the District of Columbia Public Schools (DCPS)*. DC Public Education Reform Amendment Act (PERAA) Report No. 3, Part I. Washington, DC: The Education Consortium for Research and Evaluation (EdCORE), George Washington University (June 30). [accessed Noveber 22, 2018]

Sala-i-Martin, Xavier, Gernot Doppelhofer, and Ronald I. Miller. 2004. "Determinants of long-term growth: A Bayesian Averaging of Classical Estimates (BACE) approach." *American Economic Review* 94, no. 4 (September): 813-835.

Sanders, William L., and Sandra P. Horn. 1994. "The Tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment." *Journal of Personnel Evaluation in Education* 8: 299-311.

TNTP. 2012. *The irreplaceables: Understanding the real retention crisis in America's urban schools*: The New Teachers Project.

Tyack, David B. 1974. *The One Best System: A history of American urban education*. Cambridge, MA: Harvard University Press.

Weisberg, Daniel, Susan Sexton, Jennifer Mulhern, and David Keeling. 2009. *The widget effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness* Second Edition ed. New York, NY: The New Teachers Project.

FIGURE 1

Time Pattern of the Introduction of State Accountability

FIGURE 2

Impact on Student Lifetime Incomes by FTE Students Taught (compared to average teacher)

FIGURE 3

Alternative Estimates of Teacher Deselection and Student Achievement

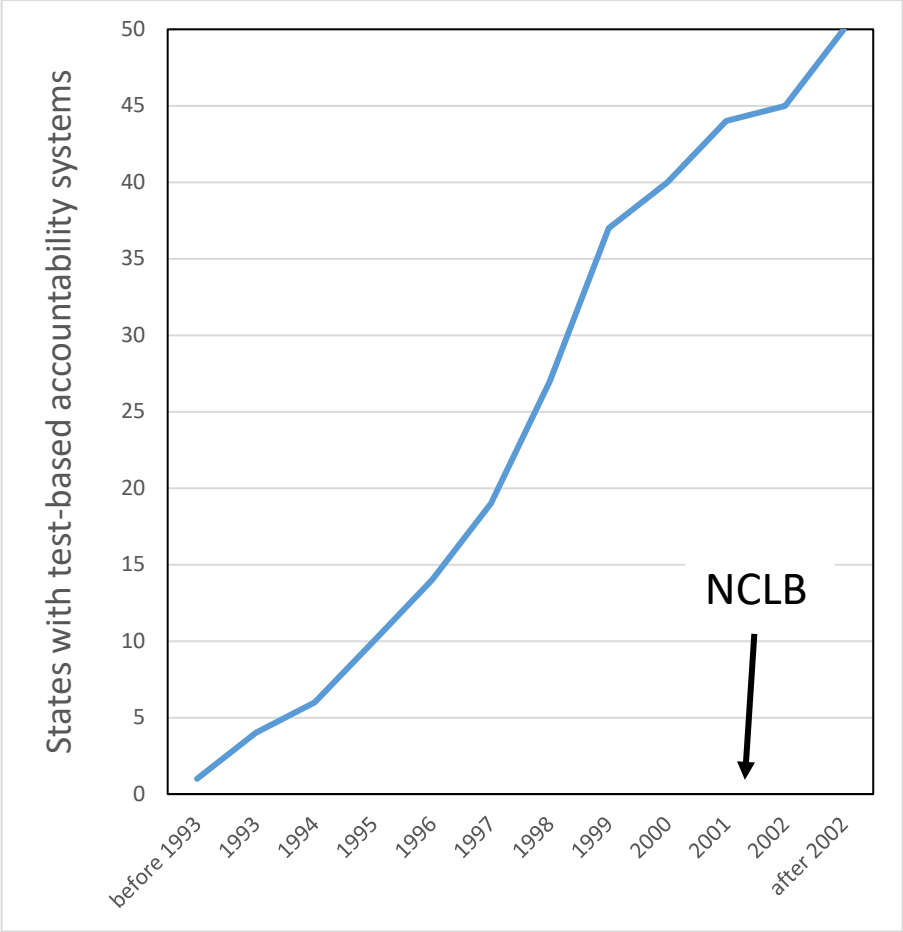

Figure 1. Time Pattern of the Introduction of State Accountability

Source: Author calculations based on Hanushek and Raymond (2005)

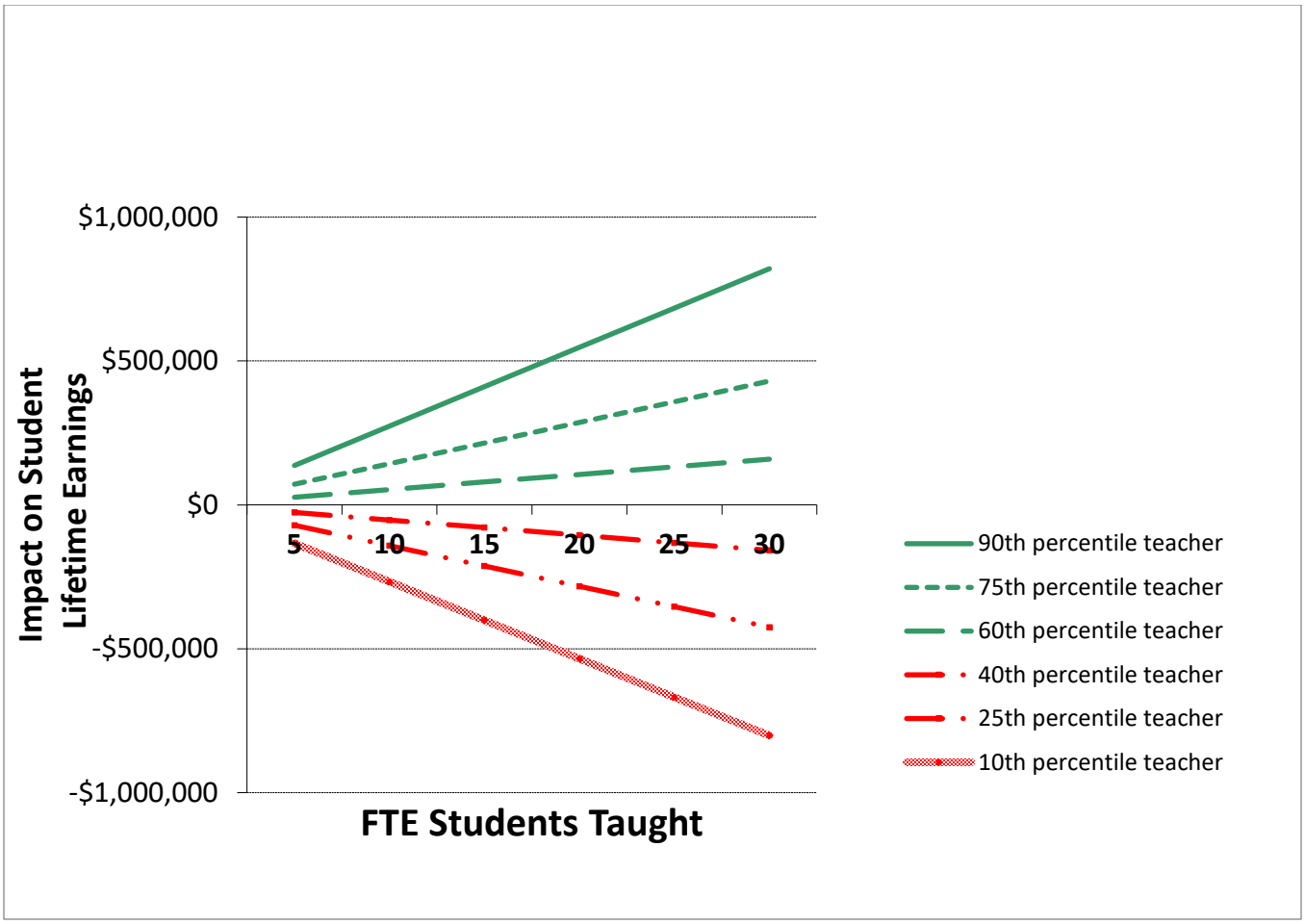

Figure 2. Impact on Student Lifetime Incomes by FTE Students Taught (compared to average teacher)

Note: Comparisons provide the present value of future incomes compared to an average teacher for teachers at different percentiles of effectiveness. For elementary schools with self-contained classrooms, the FTE counts are simple equivalent to class size. For specialist teachers who teacher multiple sections but just for a portion of the day, the FTE count comes from adding across the average students taught during the day.

Source: Hanushek (2011a)

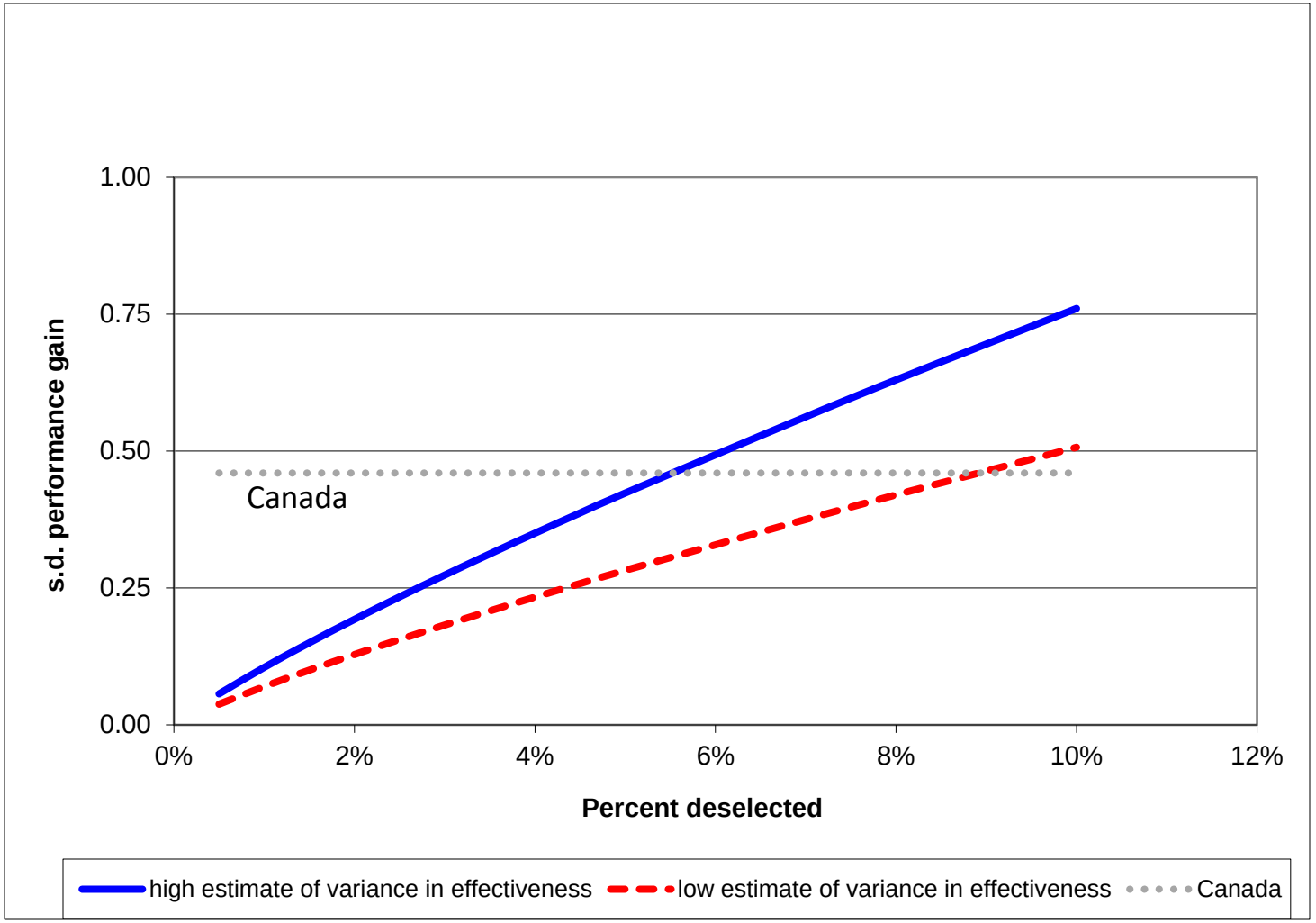

Figure 3. Alternative Estimates of Teacher Deselection and Student Achievement

Source: Hanushek (2011a)