The Distribution of Teacher Quality and Implications for Policy

Eric A. Hanushek[1,3–5] and Steven G. Rivkin[2–5]

[1]Hoover Institution, Stanford University, Stanford, California 94305; email: hanushek@stanford.edu

[2]Department of Economics, University of Illinois at Chicago, Chicago, Illinois 60607; email: sgrivkin@uic.edu

[3]Department of Economics, University of Texas at Dallas, Richardson, Texas 75080

[4]National Bureau of Economic Research, Cambridge, Massachusetts 02138

[5]CESifo, 81679 Munich, Germany

Annu. Rev. Econ. 2012. 4:131–57

First published online as a Review in Advance on March 20, 2012

The *Annual Review of Economics* is online at economics.annualreviews.org

This article's doi:
10.1146/annurev-economics-080511-111001

JEL codes: H4, I2, J4

1941-1383/12/0904-0131$20.00

Keywords

teacher value-added, persistence, test measurement error, LIFO policies, economic impacts

Abstract

It has become commonplace to measure teacher quality in terms of teacher value-added. Operationally, this means evaluating teachers according to the learning gains of students on various achievement tests. Existing research consistently shows large variations in teacher effectiveness, much of which is within schools as opposed to between schools. The policy implications of these variations are dramatic. But the underlying statistical modeling has become the subject of intense research, in part because of this direct use of value-added measures in policy discussions.

Value-added of teachers: the separate contribution of teachers to learning gains, independent of family, peer, and other influences

1. INTRODUCTION

Children, parents, administrators, and policy makers all tend to focus on the quality of teaching as the key determinant of school quality. However, the failure of observed teacher characteristics, including education and experience, to account for much of the variation in student test scores or other outcomes has presented researchers with an apparent conundrum. Although the absence of a strong relationship between outcomes and these characteristics is consistent with the idea that teacher quality is not an important determinant of learning, it is also consistent with the possibility that these quantifiable characteristics are simply poor measures of teacher effectiveness. In fact, recent outcome-based estimates find substantial variation in teacher contributions to achievement, supporting both the latter interpretation and the general emphasis on teacher quality.

Easily quantifiable characteristics explain little of the variation in teacher effectiveness, and this has important implications for the development of policies designed to raise the quality of instruction and to reduce unequal access to high-quality teachers. First, neither a graduate degree nor additional years of experience past the initial year or two translate into significantly higher instructional effectiveness, bringing into question a salary structure based almost entirely on these two variables. Second, descriptions of unequal access to quality teachers as measured by experience, education, or other quantifiable characteristics fail to portray accurately any actual differences in the quality of instruction by student demographics, community characteristics, and specific schools. Third, the failure of quantifiable characteristics to explain much of the variation in teacher effectiveness suggests that efforts to raise the quality of instruction through more stringent requirements for entering the teaching profession may be seriously misguided, particularly as they may discourage many from entering the profession by raising the cost of becoming a teacher. Fourth, the focus on student outcomes—highlighted by state and federal accountability systems, including No Child Left Behind—has led to attempts to legislate better teachers, although these efforts have not circumvented the problems of defining an effective teacher.[1]

The analysis of teacher effectiveness has largely turned away from attempts to identify specific characteristics of teachers. Instead attention has focused directly on the relationship between teachers and student outcomes. This outcome-based perspective, now commonly called value-added analysis, takes the perspective that a good teacher is simply one who consistently gets higher achievement from students (after controlling for other determinants of student achievement such as family influences or prior teachers). The underlying analysis has focused on statistical estimation that separates teacher influences from other factors, and most typically it has relied on administrative data from schools.

The current discussion of value-added measures of teacher effectiveness is not so much about the conceptual approach as it is about the details of the analysis and the interpretation of the results. The debate has been heightened by the direct introduction of value-added estimates into teacher evaluations and personnel decisions (see, e.g., Isenberg & Hock 2010) and by the public identification of teachers according to value-added in the newspapers (Song & Felch 2011). This interaction with actual policy decisions plus the larger focus of recent education reforms on student outcomes has contributed to growing

[1]No Child Left Behind introduced a requirement for "highly qualified teachers" within schools serving disadvantaged students. This requirement was phrased in terms of qualifications as opposed to effectiveness in the classroom, and the definitions of highly qualified were left up to the separate states. As a result, most states simply inserted variants of the existing requirements for teacher certification (see Chubb 2009, Hanushek 2009).

scholarly discussion of the validity of outcome-based, or value-added, estimates of teacher quality. Rothstein (2010) provides evidence that sorting both among and within schools may bias estimates of teacher effectiveness, and Meghir & Palme (2005) emphasize the possibility that school and family responses endogenous to realized teacher quality contribute to observed achievement. Other analyses have emphasized the importance of measurement error in using test outcome data (e.g., Kane & Staiger 2002, McCaffrey et al. 2009).

We consider the existing evidence on teacher quality, the methods used to produce those estimates, and implications of the findings for the development of teacher policies. Section 2 summarizes baseline estimates of the variance in teacher quality based on estimates of teacher fixed effects and other empirical approaches. Section 3 presents issues related to the measurement of outcomes. Section 4 discusses potential impediments to the identification of the teacher-quality variance caused by purposeful sorting of children and teachers into neighborhoods, schools, and classrooms and the endogenous responses by parents and school administrators to observed teacher effectiveness. Section 5 describes research that considers the persistence of teacher effects over time. Section 6 provides a discussion of selected implications for education policy, and Section 7 has some final thoughts about efforts to raise the quality of instruction.

2. OUTCOME-BASED ESTIMATES OF THE VARIANCE IN TEACHER EFFECTIVENESS

Research into the determinants of student outcomes and the role of teachers has changed dramatically over the past two decades. The earliest work, generally traced back to the "Coleman Report" (Coleman et al. 1966), was typically cross-sectional analysis that relied on measuring the various inputs to education—from families, peers, and schools. Additionally, much of the early analyses relied on data collected for other purposes and not specific to educational performance. More recently, the analysis has changed dramatically as data have improved, as researchers have extended the simple analyses in a variety of ways, and as replication of various analyses has become more feasible.

From the start, the impact of teachers was simply viewed as one of many possible inputs to achievement. The consistent finding over four decades of research—frequently called education production-function research in economics—has been that the most commonly used indicators of quality differences are not closely related to achievement gain, leading some to question whether teacher quality really matters (see the review in Hanushek & Rivkin 2006).[2] In reality, interpretations of research on teachers (and other school inputs) often confused the estimated effects of specific teacher characteristics with the overall contribution of teachers.

The focus on the measurement of teacher value-added to student achievement represents a shift from a research design focused on the link between student outcomes and specific teacher characteristics to a research framework using a less parametric approach to identify overall teacher contributions to learning. Using administrative databases (some covering all the teachers in a state), such research provides strong support for the existence

[2]The Coleman Report (Coleman et al. 1966) is generally acknowledged as the beginning of this entire line of research into the determinants of student achievement. It was commonly interpreted as finding that "schools don't matter;" instead, families and peers are the prime determinants of student achievement. Although this study was deeply flawed (Hanushek & Kain 1972), portions of this conclusion still are included in public policy discussions.

of substantial differences in teacher effectiveness, even within schools.[3] Although this approach circumvents the need to identify specific teacher characteristics related to quality, the less parametric approach introduces additional complications and has sparked an active debate on the measurement and subsequent policy use of estimated teacher value-added.

The precise method of attributing differences in classroom achievement to teachers is the subject of considerable analysis, as discussed below. We begin by briefly outlining the general analytical framework that forms the basis of much work in this area and then describe the range of results from recent efforts to measure the variance of teacher effectiveness.

As the available research discusses in great detail, the determinants of both student and teacher choices and the allocation of students among classrooms complicate efforts to isolate the contributions of teachers to learning. The discussion highlights the importance and difficulty of accounting for student and school differences that affect both own performance and the classroom environment, and the likely advantages of a value-added modeling approach that accounts explicitly for differences in the history of family, school, and community influences. It also recognizes, however, that the value-added framework does not address all potential impediments to consistent estimation of the quality of instruction and the variance in teacher quality. Specific methods mitigate some deficiencies and not others, although none may completely resolve the potential problems.

This section outlines the decision-making processes of families, teachers, and principals and the potential implications of these choices for the estimation of teacher quality; presents an empirical framework that captures potential biases from these choices and other factors; and ends with a discussion of empirical evidence on the validity of specific estimation methods.[4] A central issue in the empirical analysis is that the relevant participants—families, teachers, and administrators—are not randomly assigned to schools and classrooms. Families choose a community and school, possibly trading off school quality with other housing amenities (see Hanushek & Yilmaz 2007, Hanushek et al. 2011b). Teachers and administrators also make choices along a number of relevant dimensions, including geography, school, and individual effort. These choices must be taken into account to identify the variation in teacher value-added to student achievement.

The basic educational model implicitly underlying educational production functions is that

$$A_{gi} = f(S_i, X_i, \mu_i), \tag{1}$$

where A_g is the achievement of student i in grade g (the subscript i is suppressed when it is not central to the discussion), $\mathbf{S}$ is a vector of school and peer factors, $\mathbf{X}$ is a vector of family and neighborhood inputs, and μ is individual student ability.

One important facet of education is that it is a cumulative process. Achievement in grade g involves not only educational inputs in that grade, but also the whole history of inputs that provided the basic knowledge that enters into the summative achievement in grade g. To fix some basic ideas and to match much of the empirical analysis, assume a linear model such that

$$A_{Gi} = \sum_{g=0}^{G} S_{gi}\varphi_g + \sum_{g=0}^{G} X_{gi}\gamma_g + \sum_{g=0}^{G} \mu_i + \varepsilon_i, \tag{2}$$

[3]The earliest academic research includes Hanushek (1971), Murnane (1975), and Armor et al. (1976). Policy interest rose with the introduction of the ideas directly into state evaluations (Sanders & Horn 1994).

[4]Ballou et al. (2004) and McCaffrey et al. (2004) provide general discussions of value-added estimation of teacher effects.

where $\mathbf{S}_g$ and $\mathbf{X}_g$ are the vectors of school and family inputs during grade g, respectively; μ is assumed to be the constant input of ability; φ and γ are unknown parameters; ε is a stochastic term representing unmeasured influences; and we observe how inputs accumulate through grade G.[5] Then we can decompose the achievement determinants into current (grade G) and prior grade influences. If we put some structure on the parameters such that they follow a geometrically declining pattern for inputs in the more distant past (indicating that the impact of past inputs depreciates at a constant rate θ over time), we get an expression such as

$$\begin{aligned} A_{Gi} &= S_{Gi}\varphi + X_{Gi}\gamma + \mu_i + \varepsilon_{Gi} \\ &+ \sum_{g=0}^{G-1} S_{gi}\varphi(1-\theta)^{G-g} + \sum_{g=0}^{G-1} X_{gi}\gamma(1-\theta)^{G-g} + \sum_{g=0}^{G-1} \mu_i(1-\theta)^{G-g} + \sum_{g=0}^{G-1} \varepsilon_{gi}(1-\theta)^{G-g}. \end{aligned} \quad (3)$$

We note that, if this relationship holds across grades, the second line is simply $(1-\theta)A_{G-1,\,i}$ so that we can write current achievement as a function of (depreciated) past achievement plus the inputs during grade G:

$$A_{Gi} = S_{Gi}\varphi + X_{Gi}\gamma + (1-\theta)A_{G-1,i} + (\mu_i + \varepsilon_{Gi}). \quad (4)$$

This expression for achievement obviously puts considerable structure on the relationship, something that has been investigated in various forms, as indicated below.

This expression is particularly useful because few if any data sets actually track the entire past history of family and school inputs. Equation 4 suggests that it is possible to concentrate on the inputs in grade G, with A_{G-1} capturing the contributions of historical inputs.[6]

In addition, many data sets including administrative data have limited information on family income and other current family characteristics, including parental education, that are likely related to both commitment to schooling and home resources available to support education. Hence it is often difficult to control directly for family heterogeneity. The availability of multiple observations for students in panel data sets, however, makes possible alternative approaches to account for such heterogeneity. Specifically, many studies use measures of prior achievement, student fixed effects, or a combination of the two to account for stable differences among students not captured by the limited set of available variables (as suggested by Equation 4).

Empirical analyses of teacher value-added typically begin with a slightly modified version of the education production function in Equation 4 that breaks out a term for teacher inputs, τ_j:

$$A_G = (1-\theta)A_{G-1} + \tau_j + \mathbf{S\varphi} + \mathbf{X\gamma} + \varepsilon, \quad (5)$$

where τ_j is a teacher fixed effect that provides a measure of value-added for teacher j.[7] Most empirical analyses, although not explicit, also subsume ability effects (μ) from Equation 4 into a composite error, ε, although some of the discussion of sorting into

[5] Developments along these lines can be found in Boardman & Murnane (1979), Hanushek (1979), and, more recently, Todd & Wolpin (2003).

[6] Again, this interpretation involves a number of assumptions about the additive nature of inputs. For example, it does not allow for the "one great teacher" who has a lasting impact on the pattern of achievement (over and above the effect on achievement in the one year). It also assumes that knowledge from all sources depreciates at the same rate, something that comes up below in considering the persistence of teacher effects.

[7] Alternative estimation forms, largely restricting θ, have advantages and disadvantages but are currently less frequently employed (see Meghir & Rivkin 2011 on conceptual issues and Hanushek et al. 2009 on empirical implications).

Within-school variance: the variation in student achievement found inside schools as opposed to between schools; often the focus of estimates of teacher value-added

Test measurement error: the portion of test scores that does not reflect the true knowledge of a student in a given domain

classrooms can be thought of as focusing on these ability effects.[8] Furthermore, various analyses focus on the precise specification of τ_j, such as whether it is the same for all students or there are some heterogeneous treatment effects.[9]

Although we discuss further details of the estimation below, we begin simply with available estimates of the variation in teacher value-added (i.e., variation in τ_j).[10] The most conservative estimates try to guard against biases from unmeasured characteristics of families and teachers by concentrating just on the differences among teachers within a given school. This approach holds constant the aggregate motivation of families to be in a given school at the cost of ignoring any variation in average teacher effectiveness across schools.[11]

Table 1 summarizes existing estimates of the standard deviation of teacher effectiveness (σ_τ) expressed in units of student achievement (normalized to a standard deviation of one).[12] Although covering a range of schooling environments across the United States, these studies produce fairly similar estimates of the variance in teacher value-added: The average standard deviation is 0.13 for reading and 0.17 for math, and the distributions for both are fairly tight. We note also that these estimates rely on just within-school variation in value-added, ignoring the surprisingly small between-school component. The between-school component is not typically considered because of potential sorting, testing, and other interpretative problems.[13]

The magnitudes of these estimates support the belief that teacher quality is an important determinant of school quality and achievement. For example, the math results imply that having a teacher at the 25th percentile as compared to the 75th percentile of the quality distribution would mean a difference in learning gains of roughly 0.2 standard deviations in a single year. This would move a student at the middle of the achievement distribution to the 58th percentile. The magnitude of such an effect is large both relative to typical measures of black-white or income achievement gaps of 0.7–1 standard deviations and compared to methodologically compelling estimates of the effects of a ten-student reduction in class size of 0.1–0.3 standard deviations. We discuss the economic value of these differences in teacher effectiveness below.

3. OUTCOME MEASUREMENT ISSUES[14]

Test measurement error can complicate the outcome-based estimation of teacher quality. Four test measurement issues receive considerable attention in both the research literature

[8]For the most part, all estimates employ OLS estimation in a panel context.

[9]One issue that is not explicitly considered here is that of peer interactions and their effect on achievement. Although there is a substantial literature on this (see the review in Sacerdote 2011), the consideration of peer interactions and of teacher quality has been almost completely separate in the literature—and we follow that history in this discussion.

[10]Again, this may be interpreted as the average treatment effect, and (as discussed below) some attempts have been made to look at specific dimensions of heterogeneous effects. These differences appear relatively small, however, compared to the overall estimated variation in teacher value-added.

[11]Families have been seen to choose schools based on school value-added (see Hanushek et al. 2004a). The differences in demand show up in housing price variations, although the measurement of school quality varies widely across studies (see Black & Machin 2011).

[12]Some questions remain about whether performance on different tests can be compared through this normalization. At the very least, such comparisons require assumptions about the psychometric property of the tests such as normality of the distribution. Some analyses, as described below, consider such issues directly.

[13]The study by Kane & Staiger (2008) is the one exception that does not exclude the between-school component. Hanushek & Rivkin (2010a) provide estimates of the within- and between-school variation in value-added under different specifications.

[14]Ishii & Rivkin (2009) expand on many issues considered in this and the next section.

Table 1 The distribution of teacher effectiveness (standard deviations of student achievement)

Study	Location	Teacher effectiveness (σ_τ) Reading	Math
Rockoff (2004)	New Jersey	0.10	0.11
Nye et al. (2004)	Tennessee	0.07	0.13
Rivkin et al. (2005)	Texas	0.15	0.11
Aaronson et al. (2007)	Chicago		0.13
Kane et al. (2008)	New York City	0.08	0.11
Jacob & Lefgren (2008)	Midwest city?	0.12	0.26
Kane & Staiger (2008)	Los Angeles	0.18	0.22
Koedel & Betts (2011)	San Diego		0.23
Rothstein (2009)	North Carolina	0.11	
Hanushek & Rivkin (2010a)	Texas city?		0.11
Average		0.13	0.17

All estimates indicate the standard deviation of teacher effectiveness in terms of student achievement standardized to mean zero and variance one. All are corrected for test measurement error. All except Kane & Staiger (2008) use within-school estimators. Table taken from Hanushek & Rivkin (2010b).

and broader policy discussion: (*a*) random measurement error, (*b*) the focus of tests on particular portions of the achievement distribution, (*c*) cardinal versus ordinal comparisons of test scores, and (*d*) the multidimensionality of education outcomes. Not only do the test measurement issues introduce noise into the estimates of teacher effectiveness, but they also bias upward estimates of the variance in teacher quality.

The fixed-effect estimator is the conditional mean performance for a teacher's students, given the prior achievement levels of the students and other factors explicitly considered, i.e., the S and X. This estimated value of teacher quality, $\hat{\tau}_j$, is the true value of teacher quality, τ_j, plus estimation error, v_j:

$$\hat{\tau}_j = \tau_j + v_j. \tag{6}$$

Considerable attention has been devoted to the magnitude and character of test measurement error.[15] If the only error in Equation 5 comes from error in the individual measurement of achievement, the random test measurement error introduces sampling error into the estimates of the teacher fixed effects, v_j. When the test measurement errors are uncorrelated with true teacher quality and the expected value of v_j equals zero, the estimates of teacher effects will be unbiased, but there will be an upward bias in the estimated variance of teacher quality. In particular, the estimated variance based on raw

[15]This attention to test measurement error can be traced back to early work on school accountability by Kane & Staiger (2002).

test scores would equal $\mathrm{var}(\hat{\tau}) = \mathrm{var}(\tau) + \mathrm{var}(\nu)$. In this case, $\mathrm{var}(\nu)$ is a direct function of the variance of test measurement error.

All the estimates of the variance in teacher effectiveness in **Table 1** were adjusted in one way or another for test measurement error. One common approach is the empirical Bayes approach that adjusts the estimates by producing weighted averages of the actual estimates and the grand mean, in which the weights are a function of the estimated variance (see, e.g., Kane & Staiger 2002, Jacob & Lefgren 2008).[16] The impact of this adjustment is to pull each of the individual estimates toward the grand mean and, as a result, reduce the variance of the estimates. As Kane & Staiger illustrate in a slightly different context, sample size is the primary determinant of differences in the error variance; consequently, the grand mean tends to receive higher weight for teachers with fewer students.

Alternatively, assuming that the confounding measurement error is uncorrelated across years, the correlation of the fixed-effect estimates for the same teachers across years would equal the ratio of the variance of true teacher quality divided by the variance of estimated teacher quality, $r_{12} = \mathrm{var}(\tau)/\mathrm{var}(\hat{\tau})$. Multiplication of the estimated variance of $\hat{\tau}$ by the estimated year-to-year correlation would provide a consistent estimate of the overall variance in teacher quality corrected for the random contributions of test error and the other factors (see Hanushek et al. 2005, McCaffrey et al. 2009).

A potential problem with efforts to remove the non-teacher-quality components of the variance estimate relates to variation in teacher quality over time or across students. For the most part, studies have defined teacher quality as the stable component of teacher effects on students. But in a year when a teacher simply has a "bad year" for one reason or another, the teacher's impact on the students reflects this problem. The seriousness of this issue, perhaps including whether alternative definitions of teacher quality should be employed, is currently unknown—in part because it is difficult analytically to separate variations in teacher quality from random errors not associated with the teachers.[17] Similarly, heterogeneity in teacher quality by student demographic characteristics (perhaps owing to the benefits of a same-race or same-gender teacher) or skills (some teachers may be more effective with students in particular parts of the achievement distribution) will tend to increase the standard error in models that impose homogeneous effects (see Dee 2005, Hanushek et al. 2005, Lockwood & McCaffrey 2009). The standard focus on the "persistent" component of teacher quality ignores heterogeneity along these dimensions, which may be appropriate depending on the use of the estimates.[18]

The second and third issues concerning assessment differences across the knowledge distribution and the appropriate interpretation of test results are related in both substance and proposed solutions. First, there are complications introduced by the coverage of a test. If the degree of overlap between the curriculum and test varies, then estimates of value-added will not produce consistent rankings of teacher effectiveness. Consider first a test that focuses on basic skills and a class filled with students working above grade level. In this case, the efforts and efficacy of the current teacher can have virtually no impact on test

[16]Morris (1983) describes the empirical Bayes shrinkage estimator.

[17]In some instances, researchers have attempted to consider whether the impact of a teacher changes with a move to another school, thus allowing for variations in teacher quality (e.g., Jackson & Bruegmann 2009, Goldhaber et al. 2011). They have provide mixed evidence on any changes in teacher performance.

[18]When the approach to measurement error is an empirical Bayes method, ignoring these components tends to lead to greater shrinkage toward the grand mean for particular teachers.

results, and thus the estimate of value-added for that teacher would not be based on true performance. More generally, the less overlap there is between test coverage and the prescribed curriculum, the less informative are the estimates.

Second, consider assumptions related to test scaling and the meaning of test-score changes in various parts of the test distribution. There is an ongoing debate among psychometricians over the construction of vertically scaled tests for which a one unit change in score has similar meaning across the distribution in terms of actual knowledge gained (see, e.g., Koretz 2008, Braun et al. 2010). Many argue that even those tests assumed to be vertically scaled actually do not satisfy this criterion, and many tests administered in schools do not even claim to have this property. Such concerns have led some to focus solely on the ordinal properties of test results, for example, on percentile scores. Yet percentile score measures also do not avoid assumptions about scaling, as it takes a larger change in knowledge to move a given number of percentile points in the tails of the distribution. Therefore, two similar effective teachers may produce very different average changes in student percentile rankings with different distributions of students.

One approach designed to address both scaling and test-focus concerns is to compare each student with all students in the sample with similar prior test-score performance (see, e.g., Neal 2011). There are several different manifestations of this type of model, but essentially teachers would be judged by how their students perform relative to students at similar levels of achievement in other classrooms. For example, one could estimate a model with teacher fixed effects in which the percentile score would be regressed on dummy variables for each percentile of the prior-year score (and potentially scores in previous years or other subjects), and other controls. Alternatively, quantile treatment-effect methods could be used to calculate the median (or other) score percentile given the prior-year distribution of achievement. Percentile growth models, discussed here for the evaluation of teacher effectiveness, have been introduced into the accountability system of some states (e.g., Colorado).[19]

The final issue concerns the interpretation of estimates of teacher productivity. Education is a multidimensional enterprise in which schools seek to foster learning along a number of academic as well as nonacademic dimensions. Because teachers allocate time among various endeavors in pursuit of a number of objectives, estimates of teacher quality should ideally incorporate the time allocated to the type of instruction in question to generate meaningful measures of effectiveness. If it takes one teacher twice as long as another to produce a given amount of mathematics learning, that teacher should be judged as half as productive in this area. Yet in reality, a lack of information on teachers' time allocation prevents adjustments for time spent on a given task.[20]

4. OTHER IMPEDIMENTS TO THE ESTIMATION OF TEACHER QUALITY

Although the outcome measurement issues have been rather thoroughly addressed, another set of issues evokes considerable current discussion and disagreement. In particular, the

[19]Moreover, such comparisons are embedded into some state accountability systems such as Colorado (Betebenner 2007).

[20]Little analytical work has gone into identifying multiple outcomes, although it has entered directly into policy applications in which alternative teacher assessments enter even when value-added measures are available and when multiple value-added measures are combined [see, e.g., the IMPACT evaluation system in Washington, DC (**http://www.dc.gov/DCPS/impact**)].

myriad decisions of families, administrators, teachers, and students that determine the distributions of students and teachers in schools and classrooms raise concerns about the validity of estimates of teacher value-added. This section begins with a discussion of the various levels of choices and then turns to a discussion of the empirical evidence.

4.1. Purposeful Choices of Families, Teachers, and Schools

Residential location, school choice, classroom allocation, and academic support decisions complicate the estimation of teacher quality, and researchers have used a variety of methods to isolate the contributions of teachers. These include the use of student fixed effects and prior test scores to account for unobserved student heterogeneity and school fixed effects to account for unobserved differences in a range of school factors. Yet concerns remain that the endogenous sorting of children among classrooms biases estimates of both teacher quality and its variance. For example, Rothstein (2010) focuses on potential biases introduced by sorting on the basis of time-varying unobservables.[21]

The direction of any bias would depend on the character of that sorting, which would be determined largely by the principal's objective function. An egalitarian principal might place more disruptive children with a higher-quality teacher, whereas a principal that desires to please the senior staff might give experienced teachers the more compliant, intellectually engaged children. These alternative allocation mechanisms produce very different distributions of achievement within schools and different patterns of potential bias in the absence of adequate controls for classroom heterogeneity.

The direction of any bias also depends on the character of teacher sorting among communities and schools. Evidence on teacher preferences suggests that teachers tend to prefer schools with higher-achieving students and appear to have heterogeneous preferences regarding school location and characteristics related to student race and ethnicity.[22] Survey evidence also suggests that principal behavior influences the probability that a teacher remains in a school, which potentially introduces a link between teacher quality and principal behavior.[23] Together these findings suggest the existence of a positive correlation between teacher supply and factors that contribute to higher achievement (Hanushek & Rivkin 2007). To the extent that schools exploit superior positions in the teacher labor market by employing more effective teachers, unobserved student heterogeneity is likely to inflate the between-school component of the estimated variance in teacher quality.

The possibility of endogenous school and family responses to realized teacher quality complicates estimation even further. In the absence of information on the prevalence of the intervention of teachers or family support, the estimates of teacher effectiveness combine the actual teacher effect with these other influences. In the case of classroom support, it is likely that the allocation of resources is compensatory, meaning that it would bias estimates of teacher productivity toward the mean and estimates of the variance in teacher quality downward. With regards to family influences, Kim (2001), Todd & Wolpin (2003), and Meghir & Palme (2005) point out that estimates of school

[21]Note that this is not a story of peer interactions but one of unmeasured student attributes associated with teachers that lead to bias in Equation 6.

[22]Boyd et al. (2005) and Hanushek et al. (2004b) describe differences in teacher sorting by race.

[23]Loeb et al. (2005) discuss the relationship between teacher turnover and working conditions. Branch et al. (2012) analyze how principal quality affects teacher mobility and teacher quality.

and teacher effects incorporate family responses that vary systematically with input quality. If families tend to contribute more time and money to academic support during periods in which they perceive the teacher as inadequate or of low quality, this will tend to bias downward estimates of the variation in teacher quality and bias estimates of the quality of specific teachers toward the mean. However, if one component of high-quality teaching is raising the level of difficulty, higher-quality teachers may actually increase the likelihood that students seek help outside of school, and estimates of teacher fixed effects could be biased away from zero and the estimated variance could be biased upward. In either case, it is not adequate to represent families as fixed inputs into the education production process, meaning that even the inclusion of student fixed effects could fail to account for potentially important family influences.

4.2. Empirical Evidence on the Validity of Estimates

This section considers evidence on the validity of nonexperimental estimates of teacher value-added. We begin with a paper by Kane & Staiger that finds evidence in support of the validity of value-added estimates of teacher fixed effects from an analysis of data from a random assignment experiment. Next we discuss research by Rothstein (2010) that finds evidence that widely used value-added specifications are subject to considerable bias and research by Koedel & Betts (2011) that extends Rothstein's analysis by considering samples with multiple years, rather than a single year, of data for each teacher. This leads to a brief discussion of Guarino et al. (2011), which examines the performance of various value-added models under different data-generating processes. Finally, we consider the new test of bias based on teacher mobility introduced by Chetty et al. (2011b).

Recognizing the threat of student sorting both within and between schools to the estimation of teacher value-added, Kane & Staiger (2008) use experimental data generated by a random assignment study of the National Board for Professional Teaching Standards Certification Program to investigate the validity of nonexperimental estimates of teacher value-added. In the study, pairs of teachers are identified in each school, one with and the other without certification, and classrooms are randomly assigned to the pairs. The difference in the average test scores of the classrooms is regressed on the difference in empirical Bayes estimates of value-added for the same pair of teachers based on multiple years of data from years prior to the experiment to examine the validity of the estimation based on nonexperimental data. The authors develop a specification test of the hypothesis that the nonexperimental estimates provide consistent estimates of the difference in teacher effectiveness. Results of the empirical analysis are consistent with the hypothesis that sorting on unobservables does not confound the estimates of teacher value-added based on observational data in their sample.

It should be noted that the test does have some limitations. First, given the small sample size, the test lacks power in that the acceptance region includes values that would reflect significant sorting on unobservables that would bias the estimates. Second, the compensatory assignment of better teachers to more difficult students reduces the probability of not rejecting the null hypothesis of no sorting on unobservables when it is false. Finally, the small group of schools in which principals agreed to permit classes to be randomly assigned to teachers is unlikely to be representative, meaning that evidence of the validity of value-added estimates with this sample may not generalize beyond this sample.

The aforementioned analysis covers a small sample of teachers in schools that agreed to participate in an experiment, and this both weakens the power of the test and raises questions about the generalizability of the findings. By comparison, Rothstein (2009, 2010) evaluates some widely used models of teacher value-added using a large sample of North Carolina schools drawn from state administrative data. In contrast to the findings presented in Kane & Staiger (2008), Rothstein finds evidence that time-varying unobserved heterogeneity introduces bias into the estimates of teacher value-added for a number of widely used specifications. Importantly, this finding emerges even for school fixed-effect models that focus solely on within-school variation in teacher effectiveness. It appears that the models in question do not account fully for nonrandom sorting into classrooms on the basis of time-varying factors, as Rothstein finds that the strict exogeneity assumption is violated in specifications that include student fixed effects.

Rothstein (2010) develops falsification tests for three widely used value-added models that investigate the presence of nonrandom classroom sorting on the basis of unobservables. The first model regresses test-score gain on indicators for schools and teachers; the second model regresses test score on lagged test score and the same indicators; and the final model stacks the gain observations and regresses gain on teacher indicators and student fixed effects for a sample of students who do not switch schools.

Although the precise form of the test differs by specification, Rothstein essentially tests whether classroom teachers in grade *g* account for a significant portion of the variation in test-score gains in earlier grades. He finds strong evidence that future teachers predict variation in test-score gains in models that do and do not restrict the coefficient on lagged test score to equal 1, despite the fact that future teachers cannot be causally related to achievement. Even the inclusion of student fixed effects does not eliminate the explanatory power of future teachers, meaning that the strict exogeneity condition is violated by dynamic sorting on unobservables. Importantly, lagged teacher indicators are also significantly related to achievement in models that control for lagged achievement, indicating that the dynamic sorting on unobservables is not entirely based on achievement score.

Koedel & Betts (2011) replicate Rothstein's analysis for the San Diego school district and then extend the empirical analysis to investigate the sensitivity of Rothstein's findings to the particular character of his sample. Similar to Rothstein, they find evidence of substantial bias in estimates based on a single year of data for each teacher. However, they also find that restricting the sample to teachers with multiple years of data significantly attenuates the bias. This suggests that patterns of systematic sorting change over time, perhaps to compensate teachers who received a particularly difficult draw in the previous year.

One complication relevant for each of these analyses concerns the appropriate functional form for the value-added framework. As Rothstein (2010) points out, incorrect assumptions regarding the rate of knowledge depreciation, form of student heterogeneity, and other factors can introduce bias, and models differ along a number of important dimensions. Guarino et al. (2011) attempt to clarify the nature of any biases and sensitivity to the underlying sorting patterns by examining estimates of teacher value-added from samples produced by data-generating processes that differed according to various aspects of student sorting among classrooms.

Not surprisingly, the authors find that the magnitude and character of biases for different value-added models vary according to the underlying sorting mechanism. However, they also find that the simple lagged-achievement value-added model without student fixed effects tends to outperform other more highly structural models across a range of sorting types. This occurs despite the fact that this model was not the prescribed approach in any of the cases under the specified structural model of learning with cumulative effects and student heterogeneity. The authors conclude tentatively that this pattern of findings suggests that the flexibility of the simpler lagged-achievement model tends to improve performance.

In a novel analysis of the long-run impacts of teachers that we discuss below, Chetty et al. (2011b) begin with value-added measures for a large number of teachers over time in a large metropolitan area. They then exploit school-grade-cohort changes in achievement that are associated with the movement of teachers with varying estimated value-added. The predictions of how cohort achievement will change based on their estimates of teacher quality are closely related to the observed changes, leading them to conclude that any biases in the estimated teacher effects are small. They find no evidence of either selection on observables or selection on unobservables.

Persistence: the amount of learning in one year that carries over into the future

5. PERSISTENCE OF TEACHER EFFECTS

The degree to which teacher-quality effects persist into the future constitutes an important determinant of the importance of differences in teacher productivity. If a year with a poor teacher can be easily reversed by a higher-quality teacher in a future grade, the cost of having an ineffective teacher would be low. However, if the effects of teacher quality cumulate into the future in a way that is difficult to undo, the cost would be much higher. A small but growing number of papers either focus on or consider the issue of persistence. Prior to discussing the findings, we consider some of the main conceptual issues related to the question of persistence.

5.1. Conceptual Framework

Standard education production-function approaches impose specific assumptions on the rate of knowledge accumulation and therefore treatment-effect depreciation. One of the most common assumptions is that the rate of depreciation does not depend on the source of the learning, be it family, teacher, or another factor. Yet this framework does not capture some of the main ideas put forth in the work specific to teacher effects, including the potential trade-off between teacher effects on short-term knowledge that raises only current test scores and those on longer-term knowledge that foster learning in both the current and subsequent years. Therefore, we introduce a framework that builds on the notion that teachers can affect test scores by influencing either short-term or long-term knowledge. (Jacob et al. 2010 and Carrell & West 2010 use a similar framework.) Short-term knowledge refers to information that affects the current test score but has no effect on future achievement, such as information on the specific form in which questions are asked. Longer-term knowledge refers to information that adds to the long-run stock of human capital. Empirically, teacher effects on subsequent-year test scores are assumed to identify their effects on long-term knowledge, whereas those on current-year test scores reflect the effects on both long- and short-term knowledge.

Equation 7, an expanded version of Equation 5, represents achievement (A) for student i with teacher j in year Y as a function of achievement in year 1 (A_{iy_1}), a vector of student characteristics ($\mathbf{X}$), a vector of school fixed effects for the current and previous years (φ), a vector of long-term teacher effects for the current and previous years ($\boldsymbol{\tau}$), a short-term teacher effect for year y (λ), and a random error (ε):

$$A_{ijY} = (1-\theta)A_{iy_1} + X_{iY}\gamma + \sum_{y=2}^{Y}\varphi_y + \sum_{y=2}^{Y}\tau_y + \lambda_j + \varepsilon_{ijY}. \tag{7}$$

Because long- and short-term teacher effects cannot be observed, separate identification of these and the degree of persistence requires multiple years of data and additional assumptions on top of those discussed in the previous section. Underlying Equation 7 is the assumption that short-term effects disappear after the current year. Consequently, estimates of teacher effects in the current year capture both long- and short-term effects, whereas estimates of teacher effects on subsequent-year achievement capture only the long-term component. If estimated effects of the year $y-2$ teacher are available for years y, $y-1$, and $y-2$, then comparisons between y and $y-1$ provide information on the relative size of long- and short-term effects, and comparisons between $y-1$ and $y-2$ provide information on the depreciation of past teacher effects.

5.2. Empirical Findings

A small but growing number of papers estimate the persistence of teacher effects, and we now discuss three papers from that literature: Jacob et al. (2010), Konstantopoulos (2011), and McCaffrey et al. (2009). These papers take different approaches to the identification of the rate of persistence, in part because of data differences. Not surprisingly, complications introduced by omitted variables or sample attrition raise concerns in each analysis.

Jacob et al. (2010) develop a multistage empirical framework that isolates the long-term effect of teachers and find that only 20% of the initial effect persists beyond one year. First, estimates for the quality of teacher j for students in cohort c are derived by averaging the estimated teacher j fixed effects over all cohorts other than c; this removes unobserved cohort c specific classroom influences from the estimate of teacher quality. Next they use these estimates of teacher quality to instrument for grade g achievement in a regression of achievement in grade $g+1$ on grade g achievement, classroom fixed effects, and other controls. The coefficient on grade g achievement provides an estimate of the persistence of teacher effects, as it captures the extent to which teacher-induced changes in grade g achievement persist into the subsequent year. Then the authors substitute achievement in grade $g-1$ to estimate two-year persistence and find little change over time in the rate of persistence.

This approach provides an appealing method for identifying the persistence in teacher effects, but a number of questions emerge regarding the interpretation of the estimates. One question relates narrowly to the details of the method, whereas others relate to broader issues relevant to all the studies discussed. We consider first the method-specific concern and then turn to the broader estimation issues.

The exclusion of a student's own cohort from the calculation of the prior-year teacher's effectiveness mitigates biases introduced by unobserved peer or school influences, but it also imposes the strong assumption that the quality of a given teacher's work does not vary over time. If the quality of instruction fluctuates over time owing to on-the-job training, changes in personal circumstances, changes in curriculum, or some other factor, performance in other years will provide a noisy measure of instructional quality, and the estimates of persistence will be attenuated.

The remaining estimation issues generalize beyond the specific method used in this paper and relate to both bias and interpretation. First, estimates of persistence that include current-year teacher fixed effects in addition to prior-year teacher fixed effects or quality estimates rely on changes in classroom composition for identification. If classrooms progress through school intact, there is no way to identify the separate effects of teachers in different grades because of perfect multicollinearity. Schools with only one classroom per grade or one teacher per subject per grade therefore provide no information unless some students attend one of the grades in another school in the sample. In schools with multiple classrooms per grade, identification comes from the remixing of classrooms after each grade and school transfers. Time-varying unobservables that complicate the identification of current-year teacher effects can also bias persistence estimates.

Second, test measurement error that inflates estimates of the variance in teacher quality can attenuate persistence estimates if not appropriately addressed. Because fixed effects tend to exacerbate measurement error–induced attenuation bias, the problem is likely to be particularly problematic in this context given the presence of current classroom or teacher fixed effects. Empirical Bayes shrinkage estimators or another method that addresses this issue can mitigate the bias.

Third, as Jacob et al. (2010) point out, statistical properties and coverage of tests introduce uncertainty into the meaning of any persistence estimate. The absence of strict vertical scaling means that the amount of learning reflected in a given magnitude of test-score growth varies across the achievement distribution. Monotonic transformations of a given scale will alter the estimated degree of persistence, and there is no strong a priori argument in favor of a given scaling. Test coverage and curriculum structure also influence the interpretation of persistence estimates. If knowledge gained from a teacher in grade g relates only to a small portion of the material tested in grade $g + 1$, a finding of little or no persistence contains little information on the actual degree to which the student retains knowledge learned in the prior grade. This may occur for a number of reasons, including the focus of tests on rudimentary material that is barely covered in some schools or the lack of cumulative structure in certain subjects (e.g., biology followed by chemistry).

Fourth, within-classroom variation in current-teacher value-added could also affect estimates of persistence. If teachers tend to focus on lower achievers who received inferior instruction during the prior grade, classroom average value-added would overstate teacher value-added to higher achievers and understate current-teacher value-added to lower achievers. The failure to account for compensatory behavior by the current teacher would tend to bias downward estimates of persistence.

These general issues affect the interpretation of persistence estimates in the work by Jacob et al. (2010) and in the other papers referenced above to which we now turn. The one that stands in sharpest contrast to Jacob et al. (2010) is Konstantopoulos (2011), which finds a much higher level of persistence using data from the Tennessee STAR experiment. Although the estimated variances of contemporaneous-year teacher effects are not reported, the finding that a one-standard-deviation increase in teacher quality raises subsequent-year achievement by roughly 0.1 standard deviations suggests substantial persistence in teacher-quality effects.

The use of the Tennessee STAR experimental data imparts both advantages and challenges onto this study. Ideally, selection bias would not affect a study using random assignment data, and evidence suggests that the initial kindergarten randomization was done well. However, research by Hanushek (1999b) shows the existence of substantial nonrandom attrition in the subsequent years, and this can introduce bias in the estimation of the persistence of initial teacher quality. For example, if more educationally committed families were more sensitive to teacher quality in the decision to remain in the experimental classes, one would expect students taught by the less effective teacher to become more negatively selected over time. This would tend to increase the estimate of persistence above its true value.

In addition to the issue of selection bias, the structure of the STAR experiment complicates the estimation of persistence. Students initially assigned to a particular class type are scheduled to remain in that class-size category for the duration of the experiment. This means that in schools with only one teacher per class type, classes would remain intact from grade to grade in the absence of entry or exit. In this case, the effects of prior-year teachers could not be separate from those of current teachers, similar to the one-teacher-per-grade example discussed above. This requires movement and transfers among classrooms within the school or between schools, and such movements often introduce selection bias as they are often not part of the initial or even follow-up randomizations. The character of any such biases is complicated by the fact that the equations include both indicators for current-year teachers and estimated teacher quality for the prior grade.

McCaffrey et al. (2009) take a Bayesian approach and examine persistence using data for a large urban district. They find a rate of persistence significantly above zero but far closer to zero than to one (no depreciation). The teacher-effect estimates do not appear to be sensitive to the specification of the prior distribution, arbitrary rescaling of the data, or to the use of marginal models as opposed to joint models that pool mathematics and reading achievement. However, these estimates do not include any student background variables, and thus unobserved heterogeneity may contaminate the estimates. Moreover, although the estimates are not sensitive to the method used to deal with missing data, the absence of such controls raises the probability that nonrandom attrition does introduce substantial bias not accounted for by any of the missing data methods used in the paper.

Chetty et al. (2011b) provide a different perspective on the issue of persistence. They relate value-added measures for teachers in grades 4–8 to student achievement in subsequent years, finding that about 30% still shows up in achievement after three years. But, perhaps more importantly, they link teacher value-added estimates to students' subsequent outcomes, including earnings in the labor market. They find that on average a one-standard-deviation improvement in teacher value-added in a single grade raises earnings by about 1% at age 28. In other words, the impact of individual teachers carries through to the long-run success of students.

Clearly these estimates of persistence fail to find common ground, and this is not surprising given differences among the studies in both methods and material tested. It should be noted that Rothstein (2010) produces estimates similar to those of Jacob et al. (2010), whereas Kane & Staiger (2008) produce somewhat higher estimates of persistence but ones that are sensitive to the specification. Chetty et al. (2011b) show that depreciation in the impacts over time stabilizes after about three years and that the impact remains in life experiences after school. Many of the issues raised for the four papers discussed in this section are also relevant to these two studies. Finally, Carrell & West (2010) find that lasting teacher effects appear to be orthogonal to or even negatively correlated with short-term effects for a sample of Air Force Academy students. This pattern is consistent with the model introduced above that distinguishes between short- and longer-term effects.

6. SELECTED POLICY ISSUES

Mincer earnings function: the common statistical model, named after Jacob Mincer, that explains individual earnings in terms of school attainment and experience

As noted above, one reason for broad interest in the analyses of teacher quality is the direct relationship with policy deliberations. In this section we consider three areas of policy that have received considerable attention in recent years. The first is the desirability of test-based accountability as an important component of school regulations. Advocates of accountability implicitly assume that test scores are related to important long-term outcomes, including academic attainment and labor market performance, but some have questioned this assumption. Second, we discuss the use of information about the impact of achievement on economic outcomes to place a monetary value on differences in teacher quality. Finally, we discuss a method used to reduce the size of the teaching force in response to falling enrollment, declining budgets, or other factors. Regulations and practices in this area have long-term effects on the distribution of teacher quality, and we consider the implications of the widely used last in, first out policy.

6.1. Achievement and Longer-Term Outcomes

Economists' interest in human capital has largely centered on the quantity of schooling, but school quality has received increasing consideration over time. Mincer (1970, 1974) pioneered the study of the economic returns to additional years in school using readily available census data. A large number of researchers have subsequently estimated Mincer earnings functions that relate the log of individual earnings to years of schooling and labor market experience. The Mincer earnings work has been extended in a number of studies to incorporate the returns to achievement or school quality. This work is highly relevant to the current debate over test-based accountability and value-added because it provides evidence on the degree to which achievement differences translate into future differences in earnings.

The most common estimation begins with a standard Mincer earnings model with the addition of a measure of achievement (A_i), such as

$$\ln Y_i = \alpha_0 + rS_i + \alpha_1 Exper_i + \alpha_2 Exper_i^2 + \phi A_i + \varepsilon_i, \tag{8}$$

where Y_i is the earnings of individual i, S is school attainment, $Exper$ is potential labor market experience, and ε is a random error. When cognitive skills are standardized to mean zero and a standard deviation of one, ϕ is interpreted simply as the percentage increase in annual earnings that can be attributable to a one-standard-deviation increase in achievement.[24]

Table 2 summarizes different estimates of the impact of test performance on earnings (ϕ). The first three studies listed consider young workers (Mulligan 1999, Murnane et al. 2000, Lazear 2003). These studies employ different nationally representative data sets that follow students after they leave school and enter the labor force. When scores are standardized, they suggest that one-standard-deviation increase in mathematics performance

[24] Much of the labor economics consideration of earnings determination has focused on the estimates of r, interpreted as the rate of return on schooling (see Card 2001, but also the re-interpretation in Heckman et al. 2006). An underlying theme has been the influence of unmeasured ability on both the decision to attend further schooling and subsequent earnings, and one approach has been to add measured achievement to reflect ability. The interpretation that we give is that there is no conceptual difference between ability and achievement, that they can be measured with the same instruments, and that (borrowing from the educational production-function literature) achievement can be affected by schools.

Table 2 Estimates of the value of cognitive skills from Mincer earnings functions

Study	Earnings coefficient (ϕ)
Mulligan (1999)	0.11
Murnane et al. (2000)	0.10–0.15
Lazear (2003)	0.12
Hanushek & Zhang (2009)	0.20
Hanushek & Woessmann (2009)	0.14
Chetty et al. (2011a)	0.18

Table taken from Hanushek (2011).

at the end of high school translates into 10%–15% higher annual earnings.[25] The estimates do, however, come early in the worker's career, suggesting the impact may actually rise with experience.[26]

In a different set of estimates using data on a sample of workers for the United States, Hanushek & Zhang (2009) provide estimates of returns (ϕ) of 20% per standard deviation. One distinguishing feature of these estimates is that they come from a sample of workers throughout the career, as opposed to the prior estimates that all come from early-career earnings.

Using yet another methodology that relies on international test scores and immigrants into the United States, Hanushek & Woessmann (2009) obtain an estimate of 14% per standard deviation. These estimates come from a difference-in-differences formulation based on whether the immigrant was educated in the home country or in the United States. They find that skills measured by international math and science tests from each immigrant's home country are significant in explaining earnings within the United States.

Finally, Chetty et al. (2011a) look at how kindergarten test scores affect earnings at age 25–27 and find an increase of 18% per standard deviation. These estimates do not control for any intervening school attainment differences but do control for a rich set of parental characteristics.

The finding that moving one standard deviation in cognitive skills yields 10%–20% higher income may sound small, but these increments apply throughout the lifetime. In 2010, the average present value of income for full-time, full-year workers ages 25–70 is \$1.16 million.[27] Thus one-standard-deviation higher performance even at a low return of 13% per standard deviation amounts to over \$150,000.

[25]It is convenient to convert test scores into measures of the distribution of achievement across the population. A separate review of the normalized impact of measured cognitive skills on earnings by Bowles et al. (2001) finds that the mean estimate is only 0.07, or slightly over half of the specific studies here. More details on the individual studies discussed here can be found in Hanushek (2011).

[26]These estimates are derived from observations at a point in time. Over the past few decades, the returns to skill have risen. If these trends continue, the estimates may understate the lifetime value of skills to individuals. On the other hand, the trends themselves could move in the opposite direction (for an indication of the competing forces over a long period, see Goldin & Katz 2008).

[27]Calculations use average income by age for all full-time, full-year workers in the labor force in the first quarter of 2010. It is assumed that incomes rise 1% per year because of overall productivity improvements in the economy and that future incomes are discounted at 3%.

This approach will understate the full impact of achievement to the extent that higher achievement leads to higher levels of schooling, but that is generally not considered.[28] Thus these estimates can be considered lower bounds on the earnings impacts of higher achievement.

6.2. Economic Value of Increased Teacher Quality

The parameter estimates above provide the building blocks for assessing the implication of differences in teacher quality for lifetime income. Specifically, we can combine information about the distribution of teacher effectiveness and the labor market impact of higher achievement to estimate the economic impact of a given change in teacher value-added to achievement.

First, we use evidence on the distribution of teacher effectiveness to determine the magnitude of a one-standard-deviation difference in teacher quality, which constitutes a difference that is well within the range of the sample. **Table 1** provides a set of estimates based on within-school variations, but for this assessment we want the total variation that includes any between-school variation. Because of difficulties in identifying the between-school variance in quality, this analysis considers two magnitudes of total variance in teacher quality (σ_T): a lower estimate of 0.2 standard deviations and a larger estimate of 0.3 standard deviations. Thus a teacher who is one standard deviation above the mean to the distribution of teachers in terms of quality (i.e., roughly comparing the 84th-percentile teacher to the 50th-percentile teacher) is estimated to produce marginal learning gains of 0.2–0.3 standard deviations of student achievement above the average teacher.[29]

The implication for earnings depends on the persistence of this learning into the future and how this increased learning in any given year carries through into the labor market (i.e., the persistence discussed above). The baseline calculations presume that 70% of the added learning persists into the future, i.e., that θ in Equation 1 is 0.3. As a result of the imprecision of prior estimates, however, the impact of depreciation that is twice as large (i.e., $\theta = 0.6$) is also investigated.

It is now possible to calculate the value of an above-average teacher in terms of effectiveness. As a baseline case, consider the effect of having a teacher who is 0.5 standard deviations above average in the teacher-effectiveness distribution (i.e., at the 69th percentile) as opposed to a teacher at the mean of the distribution. She would, according to the above estimates, annually produce a 0.1-standard-deviations average improvement in cognitive skills of her students (assuming that the standard deviation of teacher effectiveness in units of student achievement is 0.2 standard deviations) above what would have been produced by the mean teacher.

Combining the improvement in scores for an individual with a conservative estimate of an impact on future individual earnings of 13% per standard deviation of achievement (from **Table 2**) and $\theta = 0.3$, we obtain a present value of \$10,600 over a lifetime of work for the average worker.

But this is not yet the full impact of the above-average teacher. The impact on one student is replicated across all the other students in the class. Thus the calculation of the

[28] Murnane et al. (2000) present an exception for tracing through the indirect effects (see also the discussion of the form of estimation in Hanushek & Zhang 2009).

[29] In terms of the student achievement distribution, this would move a student from the 50th percentile to the 58th to 62nd percentile (depending on the variation in teacher quality).

ECONOMIC GROWTH

An alternative estimate of the value of effective teachers focuses on the impact of student performance on economic growth. Recent analysis has demonstrated a close tie between international assessments of achievement and a country's economic growth rate (Hanushek & Woessmann 2008). Hanushek (2011) links teacher quality to potential implications for performance on international tests and, through that linkage, to effects on the aggregate US economy. Specifically, consider eliminating the bottom end of the teacher-quality distribution and replacing these teachers with average teachers. Using the prior bound on variations in teacher effectiveness as measured by achievement growth—specifically, 0.2–0.3 standard deviations—it is possible to see the impact of the least-effective teachers. Eliminating the least effective 5%–8% of teachers would bring student achievement up by 0.4 standard deviations or higher on the international tests. Using prior estimates of the achievement-growth linkage, Hanushek & Woessmann (2011) suggest that 0.4 standard deviations in gains could translate into a present value of added GDP of some \$70 trillion (compared with a US GDP of \$15 trillion in 2011).

impact of a teacher depends directly on class size, and the full annual economic value of teachers at different points in the distribution can be found simply by multiplying the individual student impact times the class size. **Figure 1** (see color insert) displays the impact of different-quality teachers according to class sizes at varying percentiles of the distribution. A teacher who is at the 60th percentile (0.25 standard deviations above average) raises individual earnings by \$5,292, and this translates into a present value of \$105,830 for a class size of 20 students. A teacher who is one standard deviation above the mean (84th percentile) produces over \$400,000 in added earnings for her class of 20.[30]

The first thing to note is that this is an annual increment by the teacher. Any teacher who stays at the given level of performance produces such an amount each year.

Secondly, as seen in the bottom half of **Figure 1**, a below-average teacher leads to a similar decrease in lifetime earnings.[31] Thus, having an effective teacher followed by an equally ineffective teacher will cancel out the gains.

The precise marginal economic value depends crucially on the three parameters of the teacher distribution and of how achievement evolves over time and affects earnings: $\sigma_T = 0.2$, $\phi = 0.13$, and $\theta = 0.3$. The impact of the different parameters is straightforward. A lower depreciation rate (higher persistence of achievement), a wider teacher-effectiveness distribution, and a larger labor market payoff to skill lead to a larger economic value of teacher effectiveness. Additionally, a return to skill of $\phi = 0.13$ most closely mirrors the labor market estimates for young workers and for time periods in the past when the demand for skill was less. More recent estimates and consideration of the full age range of workers yield larger estimates, suggesting that $\phi = 0.2$ is a plausible upper bound on the estimates. The baseline estimates do use a depreciation rate of 0.3, whereas other studies described above suggest a larger depreciation, particularly of

[30]Chetty et al. (2011a), extrapolating from their data on early career earnings, estimate the impact of a high-quality teacher at about \$214,000 per class of 20 for a teacher one standard deviation above the mean. This is very close to the lower bound estimate in **Table 3**.

[31]The decrease is slightly different because the estimates come from Mincer earnings functions, which relate the logarithm of earnings to the level of cognitive skills and thus to a slight different percentage change when evaluated at a different place in the distribution.

achievement gains induced by the teacher. We thus also look at $\theta = 0.6$, or a depreciation rate that is twice as large.

Table 3 presents alternative estimates of marginal impacts evaluated at one point in the teacher distribution—one standard deviation above the mean, or the 84th percentile. Even at the lower bound in column 1 of the table, defined by the previous quality and earnings parameters ($\sigma_T = 0.2$ and $\phi = 0.13$) but higher depreciation ($\theta = 0.6$), a good teacher with a class of 15 annually produces \$182,000 more in present value than the average teacher. The marginal annual economic value of a good teacher (compared with the average) evaluated at a given class size—say 20 students per class—does make a large difference in the estimated impact. The annual economic value with a class size of 20 ranges from a quarter of a million dollars to a million dollars at the top of the range for the three parameters together. (The final column of **Table 3** is an upper bound on estimates based on current empirical work.)

Although the difference in estimates across the parameters is large, the more striking feature of the table is the magnitude of the lower bound. A teacher in the top 15% with a class of 20 or more students yields at least \$240,000 in economic surplus each and every year compared with an average teacher.

A host of unknown factors—including the compensatory behavior of parents and schools, the cumulative nature of skills, the specific attributes valued in the labor market, and the nature of peer-classroom interactions—come into play in determining the long-run impact of specific teachers. But even twice the depreciation of achievement that was used in the baseline yields very large estimates of the value of an effective teacher—say, \$150,000 per year present value for a 75th-percentile teacher with a class of 20 students.

A final caveat is that none of the prior estimates in **Table 3** addresses any possible offset from direct class-size increases that would result if teachers were simply laid off instead of being replaced by average teachers (see the next section, however). The magnitude of any such class size effects has been controversial and does not readily permit explicit analysis (see, e.g., Hanushek 1999a, Krueger 1999, Mishel & Rothstein 2002). At large

Table 3 Sensitivity of demand based on earnings to key parameters (marginal annual economic value of a teacher one standard deviation above the mean)

	$\theta = 0.6$				$\theta = 0.3$			
	$\sigma_T = 0.2$		$\sigma_T = 0.3$		$\sigma_T = 0.2$		$\sigma_T = 0.3$	
Class size	$\phi = 0.13$	$\phi = 0.2$	$\phi = 0.13$	$\phi = 0.2$	$\phi = 0.13$	$\phi = 0.2$	$\phi = 0.13$	$\phi = 0.2$
5	\$60,652	\$93,573	\$91,215	\$140,923	\$106,556	\$164,741	\$160,566	\$248,858
10	\$121,303	\$187,145	\$182,430	\$281,847	\$213,113	\$329,482	\$321,132	\$497,715
15	\$181,955	\$280,718	\$273,645	\$422,770	\$319,669	\$494,223	\$481,698	\$746,573
20	\$242,607	\$374,290	\$364,860	\$563,693	\$426,225	\$658,964	\$642,264	\$995,431
25	\$303,259	\$467,863	\$456,075	\$704,617	\$532,781	\$823,706	\$802,831	\$1,244,288
30	\$363,910	\$561,435	\$547,290	\$845,540	\$639,338	\$988,447	\$963,397	\$1,493,146

θ is the depreciation rate, σ_T is the standard deviation of teacher quality, and ϕ is the labor market return to one-standard-deviation higher achievement.

LIFO (last in, first out): in education, refers to the common practice of laying off the teachers with least seniority (i.e., the last in) if reductions are necessary

class sizes, the estimates might be interpreted as an upper bound.[32] Of course, that would also imply that those at lower class sizes are lower bounds.

These estimates of the economic impact of teachers—which come from chaining together parameters estimated across different studies—are validated by the direct estimates of Chetty et al. (2011a). They consider how the least-effective teachers affect the earnings of students, and they also find present values of teacher impacts for the bottom 5% of teachers that are greater than $250,000 per class. Although it is difficult to compare their estimates and those here directly because of varying discount rates, class sizes, and the like, it is clear that they confirm the orders of magnitude described here.

6.3. LIFO (Last In, First Out)

Many union contracts for teachers specify that any layoffs should be done according to inverse seniority rules. The least-senior teachers must be dismissed before any tenured teachers are eliminated. But, as noted above, there is little relationship between experience and effectiveness in the classroom, except perhaps for teachers in their first few years. A layoff policy based on seniority thus would be expected to bear little relationship to the impact on student achievement. Until recently, few teachers were laid off, so the rules were not binding. However, the fiscal distress of states and localities in the aftermath of the 2008 recession led to some layoffs of teachers. Recent analyses have linked information about teacher value-added to layoff policies, and they indicate that the impact of reductions depends importantly on how the layoffs are implemented (Glazerman et al. 2010, Boyd et al. 2011, Goldhaber & Theobald 2011).

Goldhaber & Theobald (2011) evaluate the pattern of layoff notices sent to teachers in Washington State over two school years (2008–2009 and 2009–2010).[33] They found that, although teachers with master's degrees and some specialties had slightly lower probabilities of getting a notice, the overwhelmingly most important factor in the layoff decision is simply seniority. Because the less-senior teachers have lower salaries, to obtain the same fiscal savings, districts would have to lay off 20% fewer teachers if those laid off fell at the district average salary. More interesting, however, is the impact of layoffs on student outcomes that is estimated using the value-added of teachers.

Their analysis incorporates teacher-effectiveness measures in the consideration of who gets layoff notices (for the subset of teachers for which value-added estimation is possible). The teacher value-added measure is never close to being a statistically significant determinant of receiving a layoff notice. Moreover, the differential in average effectiveness (value-added) between teachers separated on the basis of effectiveness as opposed to the largely seniority basis employed in the notices is 0.2 standard deviations of student achievement.

Boyd et al. (2011) pursue a similar simulation in trying to investigate the implications of a hypothetical 5% reduction in teacher salaries for New York City. Their estimates are similar but somewhat larger than those in Goldhaber & Theobald (2011). In their analysis, an average teacher who is laid off under a value-added system is 26% of one standard

[32]Among the various class-size estimates, those in Krueger (1999) are among the largest, suggesting that a decrease in class size of eight students for four years might yield an achievement gain of 0.2 standard deviations.

[33]Note that, because of advance-notice provisions in law and contracts, school districts send out many more layoff notices than they ultimately exercise.

deviation in student achievement less effective than the typical seniority-based layoff. As they note, this is more than twice the size of the "first-year teaching" effect.

Part of the importance of these findings, however, depends on decision makers being able to identify differences in teacher effectiveness, as only a minority of teachers can participate in formal value-added analysis. For example, in the evaluation system used in Washington, DC, less than one-fifth of the teachers were assessed by value-added measures, implying that other elements of evaluation must necessarily be incorporated.

Several studies have related estimates of teacher value-added to the evaluations of principals (e.g., Murnane 1975, Armor et al. 1976, Jacob & Lefgren 2008). The recent study by Jacob & Lefgren (2008) finds that principals can do quite well at identifying which teachers are in the extreme of the distribution, although they do worse at differentiating teachers in the middle of the distribution. Identifying the extremes, however, is probably most relevant for teacher personnel and evaluation systems, particularly as related to any layoff policies. Therefore, greater flexibility in layoff decisions could lead to substantial changes in the distribution of teacher quality following a reduction in the size of the teaching force.

Of course, major changes in the terms of the standard teacher contract could have implications (positive or negative) for the teacher labor market. They could affect not only the career decisions of existing teachers, but also the supply of potential teachers into the profession. The first-order effects on achievement from selective decision making about layoffs are, nonetheless, large.

7. SOME CONCLUDING THOUGHTS

This discussion of course falls far short of describing all the relevant policy issues related to teacher quality. Experience on the impact of employing value-added measures in personnel decisions, contracts, and overall policy is currently quite limited, implying that some of the gains (and losses) from changed policies remain speculative at this time. At the same time, policies are moving rapidly, suggesting that they will soon be subject to empirical analysis.

A big issue in all the policy considerations is simply what drives the observed differences in teacher quality. Without knowing what leads to better or worse performance, it is hard to know what should be done to train teachers. It is hard to know how to hire teachers who have no observed performance. And it is hard to decide on such issues as mentoring new teachers or providing professional development.

Clearly research is continuing to fill in these missing elements. At the same time, it seems sensible to develop systems that recognize the large differences and act on them rather than delaying decisions until all the answers are available.

SUMMARY POINTS

1. The differences in the effectiveness of teachers are not captured by common observable characteristics such as teacher experience or teacher degrees.
2. The identification of teacher effectiveness on the basis of teacher value-added has shown that teacher quality varies dramatically, even within individual schools.

3. Teacher value-added has entered into policy discussions, leading to considerable current research to understand how estimation approaches affect the statistical estimates.
4. Test measurement error influences the estimates of teacher value-added, but its impact can be ameliorated with a variety of techniques.
5. It is important to consider how families (in their choices of schools) and schools (in their decisions about classroom assignments) influence estimates of teacher value-added.
6. One topic of current investigation is the persistence of learning across grades because this determines the ultimate influence of good or bad teachers on student outcomes.
7. Elements of teacher value-added have entered directly into a variety of policy discussions, such as which teachers should be laid off when there is a reduction in the teaching force.

FUTURE ISSUES

1. A variety of technical issues continue to be topics for additional research: the best ways to deal with test measurement error, the amount of persistence in learning, and the interaction of teachers with the classroom composition of students.
2. On the policy side, further consideration of how estimates of individual teacher value-added should be incorporated into personnel policies is needed, including the retention of teachers and their pay.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors benefitted from helpful comments and suggestions by Charles Manski.

LITERATURE CITED

Aaronson D, Barrow L, Sander W. 2007. Teachers and student achievement in the Chicago public high schools. *J. Labor Econ.* 25:95–135

Armor DJ, Conry-Oseguera P, Cox M, King N, McDonnell L, et al. 1976. Analysis of the school preferred reading program in selected Los Angeles minority schools. *Rep. R-2007*, RAND Corp., Santa Monica, Calif.

Ballou D, Sanders W, Wright P. 2004. Controlling for student background in value-added assessment of teachers. *J. Educ. Behav. Stat.* 29:37–65

Betebenner BW. 2007. Estimation of student growth percentiles for the Colorado Student Assessment Program. *Tech. Rep.*, Natl. Cent. Improv. Educ. Asess., Dover, NH. **http://www.cde.state.co.us/cdedocs/Research/PDF/technicalsgppaper_betebenner.pdf**

Black SE, Machin S. 2011. Housing valuations of school performance. See Hanushek et al. 2011a, pp. 485–519

Boardman AE, Murnane RJ. 1979. Using panel data to improve estimates of the determinants of educational achievement. *Sociol. Educ.* 52:113–21

Bowles S, Gintis H, Osborne M. 2001. The determinants of earnings: a behavioral approach. *J. Econ. Lit.* 39:1137–76

Boyd D, Lankford H, Loeb S, Wyckoff J. 2005. The draw of home: how teachers' preferences for proximity disadvantage urban schools. *J. Policy Anal. Manag.* 24:113–32

Boyd D, Lankford H, Loeb S, Wyckoff J. 2011. Teacher layoffs: an empirical illustration of seniority versus measures of effectiveness. *Educ. Finance Policy* 6:439–54

Branch GF, Hanushek EA, Rivkin SG. 2012. *Estimating the effect of leaders on public sector productivity: the case of school principals*. NBER Work. Pap. W17803

Braun H, Chudowsky N, Koenig J, eds. 2010. *Getting Value out of Value-Added: Report of a Workshop*. Washington, DC: Natl. Acad.

Card D. 2001. Estimating the return to schooling: progress on some persistent econometric problems. *Econometrica* 69:1127–60

Carrell SE, West JE. 2010. Does professor quality matter? Evidence from random assignment of students to professors. *J. Polit. Econ.* 118:409–32

Chetty R, Friedman JN, Hilger N, Saez E, Schanzenbach DW, Yagan D. 2011a. How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *Q. J. Econ.* 126:1593–660

Chetty R, Friedman JN, Rockoff JE. 2011b. *The long-term impacts of teachers: teacher value-added and student outcomes in adulthood*. NBER Work. Pap. 17699

Chubb JE. 2009. *Learning from No Child Left Behind: How and Why the Nation's Most Important but Controversial Education Law Should Be Renewed*. Stanford, CA: Hoover Inst. Press

Coleman JS, Campbell EQ, Hobson CJ, McPartland J, Mood AM, et al. 1966. *Equality of Educational Opportunity*. Washington, DC: U.S. Gov. Print. Off.

Dee TS. 2005. A teacher like me: Does race, ethnicity, or gender matter? *Am. Econ. Rev.* 95:158–65

Glazerman S, Loeb S, Goldhaber D, Staiger D, Raudenbush S, Whitehurst G. 2010. *Evaluating Teachers: The Important Role of Value-Added*. Washington, DC: Brookings Inst.

Goldhaber D, Gross B, Player D. 2011. Teacher career paths, teacher quality, and persistence in the classroom: Are public schools keeping their best? *J. Policy Anal. Manag.* 30:57–87

Goldhaber D, Theobald R. 2011. *Managing the teacher workforce in austere times: the implications of teacher layoffs*. Work. Pap. 2011-1.2, Cent. Educ. Data Res., Seattle

Goldin C, Katz LF. 2008. *The Race Between Education and Technology*. Cambridge, MA: Harvard Univ. Press

Guarino CM, Reckase MD, Wooldridge JM. 2011. *Can value-added measures of teacher performance be trusted?* Work. Pap. 18, Educ. Policy Cent., Michigan State Univ.

Hanushek EA. 1971. Teacher characteristics and gains in student achievement: estimation using micro data. *Am. Econ. Rev.* 60:280–88

Hanushek EA. 1979. Conceptual and empirical issues in the estimation of educational production functions. *J. Hum. Resour.* 14:351–88

Hanushek EA. 1999a. The evidence on class size. In *Earning and Learning: How Schools Matter*, ed. SE Mayer, PE Peterson, pp. 131–68. Washington, DC: Brookings Inst.

Hanushek EA. 1999b. Some findings from an independent investigation of the Tennessee STAR experiment and from other investigations of class size effects. *Educ. Eval. Policy Anal.* 21:143–63

Hanushek EA. 2009. Building on No Child Left Behind. *Science* 326:802–3

Hanushek EA. 2011. The economic value of higher teacher quality. *Econ. Educ. Rev.* 30:466–79

Hanushek EA, Kain JF. 1972. On the value of "equality of educational opportunity" as a guide to public policy. In *On Equality of Educational Opportunity*, ed. F Mosteller, DP Moynihan, pp. 116–45. New York: Random House

Hanushek EA, Kain JF, O'Brien DM, Rivkin SG. 2005. *The market for teacher quality.* NBER Work. Pap. 11154

Hanushek EA, Kain JF, Rivkin SG. 2004a. Disruption versus Tiebout improvement: the costs and benefits of switching schools. *J. Public Econ.* 88:1721–46

Hanushek EA, Kain JF, Rivkin SG. 2004b. Why public schools lose teachers. *J. Hum. Resour.* 39:326–54

Hanushek EA, Kain JF, Rivkin SG. 2009. New evidence about *Brown v. Board of Education*: the complex effects of school racial composition on achievement. *J. Labor Econ.* 27:349–83

Hanushek EA, Machin S, Woessmann L, eds. 2011a. *Handbook of the Economics of Education*, Vol. 3. Amsterdam: North Holland

Hanushek EA, Rivkin SG. 2006. Teacher quality. In *Handbook of the Economics of Education*, Vol. 2, ed. EA Hanushek, F Welch, pp. 1051–78. Amsterdam: North Holland

Hanushek EA, Rivkin SG. 2007. Pay, working conditions, and teacher quality. *Future Child*. 17:69–86

Hanushek EA, Rivkin SG. 2010a. *Constrained job matching: Does teacher job search harm disadvantaged urban schools?* NBER Work. Pap. 15816

Hanushek EA, Rivkin SG. 2010b. Generalizations about using value-added measures of teacher quality. *Am. Econ. Rev.* 100:267–71

Hanushek EA, Sarpça S, Yilmaz K. 2011b. Private schools and residential choices: accessibility, mobility, and welfare. *B.E. J. Econ. Anal. Policy* 11:44

Hanushek EA, Woessmann L. 2008. The role of cognitive skills in economic development. *J. Econ. Lit.* 46:607–68

Hanushek EA, Woessmann L. 2009. *Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation*. NBER Work. Pap. 14633

Hanushek EA, Woessmann L. 2011. How much do educational outcomes matter in OECD countries? *Econ. Policy* 26:427–91

Hanushek EA, Yilmaz K. 2007. The complementarity of Tiebout and Alonso. *J. Hous. Econ.* 16:243–61

Hanushek EA, Zhang L. 2009. Quality-consistent estimates of international schooling and skill gradients. *J. Hum. Cap.* 3:107–43

Heckman JJ, Lochner LJ, Todd PE. 2006. Earnings functions, rates of return and treatment effects: the Mincer equation and beyond. In *Handbook of the Economics of Education*, Vol. 1, ed. EA Hanushek, F Welch, pp. 307–458. Amsterdam: North Holland

Isenberg E, Hock H. 2010. Measuring school and teacher value added for IMPACT and TEAM in DC public schools: final report. *Rep.*, Math. Policy Res., Princeton, NJ

Ishii J, Rivkin SG. 2009. Impediments to the estimation of teacher value added. *Educ. Finance Policy* 4:520–36

Jackson CK, Bruegmann E. 2009. Teaching students and teaching each other: the importance of peer learning for teachers. *Am. Econ. J. Appl. Econ.* 1:85–108

Jacob BA, Lefgren L. 2008. Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *J. Labor Econ.* 26:101–36

Jacob BA, Lefgren L, Sims D. 2010. The persistence of teacher-induced learning. *J. Hum. Resour.* 45:915–43

Kane TJ, Rockoff JE, Staiger DO. 2008. What does certification tell us about teacher effectiveness? Evidence from New York City. *Econ. Educ. Rev.* 27:615–31

Kane TJ, Staiger DO. 2002. Volatility in school test scores: implications for test-based accountability systems. In *Brookings Papers on Education Policy 2002*, ed. D Ravitch, pp. 235–69. Washington, DC: Brookings Inst.

Kane TJ, Staiger DO. 2008. *Estimating teacher impacts on student achievement: an experimental evaluation*. NBER Work. Pap. 14607

Kim H-K. 2001. Is there a crowding-out effect between school expenditure and mother's child care time? *Econ. Educ. Rev.* 20:71–80

Koedel C, Betts J. 2011. Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Educ. Finance Policy* 6:18–42

Konstantopoulos S. 2011. Teacher effects in early grades: evidence from a randomized study. *Teach. Coll. Rec.* 113:1541–65

Koretz D. 2008. *Measuring Up: What Educational Testing Really Tells Us*. Cambridge, MA: Harvard Univ. Press

Krueger AB. 1999. Experimental estimates of education production functions. *Q. J. Econ.* 114:497–532

Lazear EP. 2003. Teacher incentives. *Swed. Econ. Policy Rev.* 10:179–214

Lockwood JR, McCaffrey DF. 2009. Exploring student-teacher interactions in longitudinal achievement data. *Educ. Finance Policy* 4:439–67

Loeb S, Darling-Hammond L, Luczak J. 2005. How teaching conditions predict teacher turnover in California schools. *Peabody J. Educ.* 80:44–70

McCaffrey DF, Koretz D, Lockwood JR, Hamilton LS. 2004. *Evaluating Value-Added Models for Teacher Accountability.* Santa Monica, CA: RAND

McCaffrey DF, Sass TR, Lockwood JR, Mihaly K. 2009. The intertemporal variability of teacher effect estimates. *Educ. Finance Policy* 4:572–606

Meghir C, Palme M. 2005. Educational reform, ability, and family background. *Am. Econ. Rev.* 95:414–23

Meghir C, Rivkin SG. 2011. Econometric methods for research in education. See Hanushek et al. 2011a, pp. 1–87

Mincer J. 1970. The distribution of labor incomes: a survey with special reference to the human capital approach. *J. Econ. Lit.* 8:1–26

Mincer J. 1974. *Schooling, Experience, and Earnings*. New York: Columbia Univ. Press

Mishel L, Rothstein R, eds. 2002. *The Class Size Debate*. Washington, DC: Econ. Policy Inst.

Morris CN. 1983. Parametric empirical Bayes inference: theory and applications. *J. Am. Stat. Assoc.* 78:47–55

Mulligan CB. 1999. Galton versus the human capital approach to inheritance. *J. Polit. Econ.* 107:S184–224

Murnane RJ. 1975. *Impact of School Resources on the Learning of Inner City Children*. Cambridge, MA: Ballinger

Murnane RJ, Willett JB, Duhaldeborde Y, Tyler JH. 2000. How important are the cognitive skills of teenagers in predicting subsequent earnings? *J. Policy Anal. Manag.* 19:547–68

Neal D. 2011. The design of performance pay in education. In *Handbook of the Economics of Education*, Vol. 4, ed. EA Hanushek, S Machin, L Woessmann, pp. 495–550. Amsterdam: North Holland

Nye B, Konstantopoulos S, Hedges LV. 2004. How large are teacher effects? *Educ. Eval. Policy Anal.* 26:237–57

Rivkin SG, Hanushek EA, Kain JF. 2005. Teachers, schools, and academic achievement. *Econometrica* 73:417–58

Rockoff JE. 2004. The impact of individual teachers on student achievement: evidence from panel data. *Am. Econ. Rev.* 94:247–52

Rothstein J. 2009. Student sorting and bias in value-added estimation: selection on observables and unobservables. *Ed. Finance Policy* 4:537–71

Rothstein J. 2010. Teacher quality in educational production: tracking, decay, and student achievement. *Q. J. Econ.* 125:175–214

Sacerdote B. 2011. Peer effects in education: how might they work, how big are they and how much do we know thus far? See Hanushek et al. 2011a, pp. 249–77

Sanders WL, Horn SP. 1994. The Tennessee value-added assessment system (TVAAS): mixed-model methodology in educational assessment. *J. Pers. Eval. Educ.* 8:299–311

Song J, Felch J. 2011. Times updates and expands value-added ratings for Los Angeles elementary school teachers. *Los Angeles Times*, May 7. **http://articles.latimes.com/2011/may/07/local/la-me-value-added-20110508**

Todd PE, Wolpin KI. 2003. On the specification and estimation of the production function for cognitive achievement. *Econ. J.* 113:F3–33

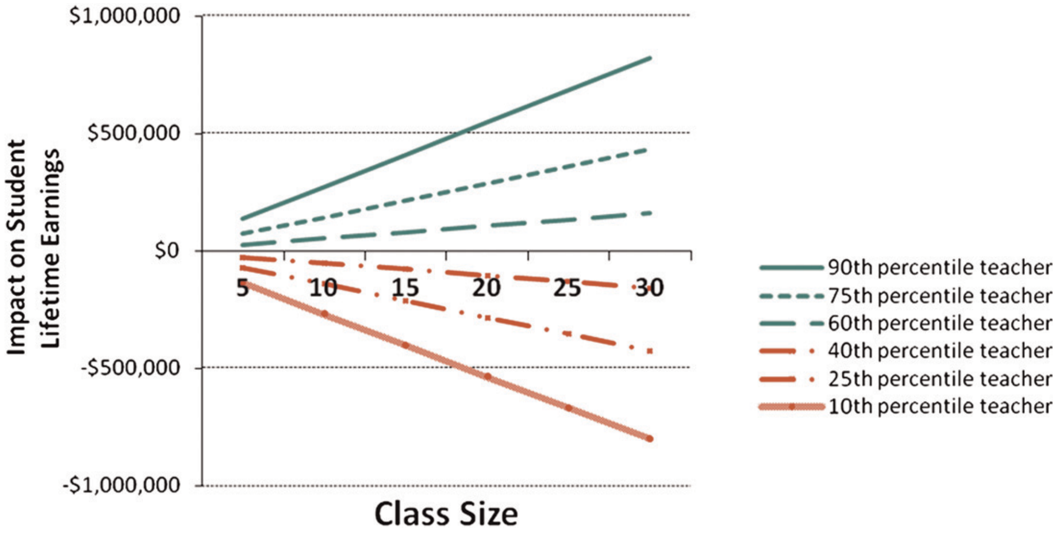

Figure 1

Impact of different-quality teachers according to class sizes compared to an average teacher.

Annual Review of Economics

Volume 4, 2012

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Economics* articles may be found at http://econ.annualreviews.org