

Measuring Effect Sizes, the Effect of Measurement Error

Don Boyd*, Pam Grossman**, Hamp Lankford*,
Susanna Loeb** & Jim Wyckoff***

*University at Albany, **Stanford University, ***University of Virginia

DRAFT
PLEASE DO NOT QUOTE OR CITE

paper prepared for the
National Conference on Value-Added Modeling
University of Wisconsin-Madison
April 22-24, 2008

We gratefully acknowledge support from the National Science Foundation and the Center for Analysis of Longitudinal Data in Education Research (CALDER). The authors are solely responsible for the content of this paper.

With the increasing availability of administrative databases that include student-level achievement, the use of value-added models in education research has expanded rapidly. These models allow us to explore how a wide variety of policies and measured school inputs affect the academic performance of students. An important question is whether such effects are sufficiently large to achieve various policy goals. For example, would hiring teachers having stronger academic backgrounds sufficiently increase test scores for traditionally low-performing students to warrant the increased cost of doing so?

The *effect sizes* of independent variables commonly are measured as the estimated effect of a one standard deviation change in the variable divided by the standard deviation of test scores in the relevant population of students. Effect size estimates derived from value-added models (VAM) employing administrative databases are typically quite small. For example, in several recent papers the average effect size of being in the second year of teaching relative to the first year, *ceteris paribus*, is about 0.04 standard deviations for math achievement and 0.025 standard deviations for reading achievement, with variation no more than 0.02. Additional research examines the effect sizes of a variety of other teacher attributes: alternative certification compared to traditional certification (Boyd et. al., 2006; Kane et. al., 2007); passing state certification exams (Boyd et. al., 2007; Clotfelter et. al., 2007; Goldhaber, 2007); National Board Certification (Clotfelter et. al., 2007; Goldhaber and Anthony, 2007; Harris and Sass, 2007); ranking of undergraduate college (Boyd et. al., 2007; Clotfelter et. al., 2007). The effect size of any single individual teacher attribute rarely exceeds the first-year experience effect, and never does so consistently across studies.

Most policymakers and researchers would judge these effect sizes to be of little policy relevance, and would rightly continue the search for the policy grail that can transform student achievement. Indeed, these estimates appear small in comparison to effect sizes obtained for other interventions. Hill, Bloom, Black and Lipsey (2007) summarize effect sizes for a variety of elementary school educational interventions from 61 random assignment studies, where the mean effect size was 0.33 standard deviations.

While specific attributes of teachers are estimated to have small effects, researchers and policymakers agree that teachers can make an important difference in student outcomes (Sanders and Rivers, 1996; Aaronson, Barrow and Sander, 2003; Rockoff, 2004; Rivkin, Hanushek and Kain, 2005; Kane, Rockoff and Staiger, in press). This finding, alongside the result that observable individual attributes of teachers appear to be of little importance, has led some to conclude that policy should focus on removing ineffective teachers rather than attempting to predict those who will be more effective or attempting to identify policies that improve teacher effectiveness.

Why might the effect sizes of teacher attributes computed from administrative databases appear so small? It is easy to imagine a variety of factors that could cause estimates of the effects of teacher attributes to appear to have little or no effect on student achievement gains, even when in reality they do. These include: measures of teacher attributes are probably very weak proxies for the underlying teacher characteristics that influence student achievement; measures of teacher attributes often are made many years before we measure the link between teachers and student achievement gains; high-stakes achievement tests may not be sensitive to differences in student learning resulting from teacher attributes;¹ multicollinearity resulting from the similarity of many of the commonly employed teacher attributes; and measurement error in student assessments. We believe that each of the preceding likely contributes to a diminished perceived importance of measured teacher attributes on student learning. In this paper, we focus on two issues pertaining to how effect sizes are measured that we believe are especially important.

First, we argue that model coefficients should be compared to the standard deviation of gain scores, not the standard deviation of scores, in calculating most effect sizes. The second issue concerns the need to account for test measurement error in reported effect sizes. The standard deviation of observed scores in the denominator of the effect-size measure reflects such measurement error as well as the dispersion in the true academic achievement of students, thus overstating variability in achievement. It is the size of an estimated effect relative to the dispersion in the gain in true achievement that is of interest. Because gain scores have measurement error in pre-tests and post-tests, netting out measurement error is especially important in this context.

Adjusting estimates of effect-size to account for these considerations is straightforward if one knows the extent of test measurement error. Technical reports provided by test vendors typically only provide information regarding the measurement error associated with the test instrument (e.g., a particular set of questions being selected). However, there are a number of other factors (e.g., variation in scores associated with students having particularly good or bad days) which can result in a particular test score not accurately reflecting true academic achievement. Using test scores of students in New York City during the 1999-2007 period, we estimate the overall extent of test measurement error and how measurement error varies across students. We apply these estimates in an analysis of how various attributes of teachers affect the test-score gains of their students and find that estimated effect sizes that include the two adjustments are four times larger than estimates that do not.

¹ Hill et. al. (2007) find that the mean effect sizes when measured by broad standardized tests is 0.07, while that for tests designed for a special topic is 0.44. So, similar interventions when calibrated by different assessments produce varying effect sizes.

In the following section we briefly introduce generalizability theory, the framework for characterizing multiple sources of test measurement error that we employ. Information regarding the test measurement error associated with the test instruments employed in New York is also discussed. This is followed by a discussion of alternative auto-covariance structures for test scores that allow us to estimate the overall extent of test measurement error, as well as how test measurement error from all sources varies across the population of students. To make tangible the implications of accounting for test measurement error in the computation of effect sizes, we consider the findings of Boyd, Lankford, Loeb, Rockoff and Wyckoff (2007) regarding how the achievement gains of students in mathematics are affected by the qualifications of their teachers. We conclude with a brief summary.

Defining Test Measurement Error

From the perspective of classical test theory, an individual's observed test score is the sum of two components, the first being the *true score* representing the expected value of test scores over some set of test replications. The second component is the residual difference, or random error, associated with test measurement error.² Generalizability theory, which we draw upon here, extends test theory to explicitly account for multiple sources of measurement error.³

Consider the case where a student takes a test consisting of a set of tasks (e.g., questions) administered at a particular point in time. Each task, t , is assumed to be drawn from some universe of similar conditions of measurement (i.e., measures that ...) with the student doing that task at some point in time. The universe of possible occurrences is such that the student's knowledge/skills/ability is the same for all feasible times. Here students are the object of measurement and are assumed to be drawn from some population. (As is typical, it is assumed the numbers of students, tasks and occurrences that could be observed are infinite.) The case where each pupil, i , might be asked to complete each task at each of the possible occurrences is represented by $i \times t \times o$ where the symbol "x" is read "crossed with".

Let S_{io} represent the i th student's score on task t carried out at occurrence o , which can be decomposed using the random-effects specification shown in (1).

$$S_{io} = \tau + \nu_i + \nu_t + \nu_o + \nu_{it} + \nu_{io} + \nu_{io} + \epsilon_{io} \quad (1)$$

$\tau_i \equiv \tau + \nu_i$, the *universal score* for the student, equals the expected value of S_{io} over the universe of generalization, here the universes of possible tasks and occurrences. The universal score is comparable to the true score as defined in classical test theory. In our case, τ_i measures the student's underlying

² Classical test theory is the focus of many books and articles. For example, see Haertel (2006).

³ See Brennan (2001) for a detailed development of Generalizability Theory. The basic structure of the framework is outlined in Conbach, Linn, Brennan and Haertel (1997) as well as Feldt and Brennan (1988).

academic achievement, e.g., ability, knowledge and skills. The v 's represent a set of uncorrelated random effects which, along with ε_{it} and the student's universal score, sum to S_{it} . Here v_t (v_o) reflect the random effect, common to all test-takers, associated with scores for a particular task (occurrence) differing from the population mean, τ . v_{it} reflects the fact that a student might do especially well or poorly on a particular task. v_{io} is the measurement error associated with a student's performance not being temporally stable even when his or her underlying ability is unchanged (e.g., a student having a particularly good or bad day, possibly due to illness or fatigue). v_{to} reflects the possibility that the performance of all students on a particular task might vary across occurrences. ε_{it} reflects the three-way interaction and other random effects. Even though there are other potential sources of measurement error, we limit the number here to simplify the exposition.⁴

The observed score for a particular individual completing a task will differ from the individual's universal score because of the components of measurement error shown in (2). In turn, this implies the measurement error variance decomposition for a particular student and a single task shown in (3).

$$\eta_{it} \equiv (S_{it} - \tau_i) = v_t + v_o + v_{it} + v_{po} + v_{to} + \varepsilon_{it} \quad (2)$$

$$\sigma^2(\eta_{it}) = \sigma^2(t) + \sigma^2(o) + \sigma^2(it) + \sigma^2(io) + \sigma^2(to) + \sigma^2(\varepsilon_{it}) \quad (3)$$

Now consider a test defined in terms of its timing (occurrence) and the N_t tasks making up the examination. The student's actual score, S_{iT} , will equal $\tau_i + \eta_{iT}$ shown in (4) where η_{iT} is a composite measure reflecting the errors in test measurement from all sources.

$$S_{iT} = \sum_t S_{it} / N_t = \tau + v_t + v_o + v_{io} + \sum_t (v_t + v_{it} + v_{to} + \varepsilon_{it}) / N_t = \tau_i + \eta_{iT} \quad (4)$$

The variance of η_{iT} for student i equals $\sigma_{\eta_{iT}}^2 = \sigma^2(o) + \sigma^2(io) + [\sigma^2(t) + \sigma^2(it) + \sigma^2(to) + \sigma^2(\varepsilon_{it})] / N_t$.

$$S_{i,g} = \tau_{i,g} + \eta_{i,g} \quad (5)$$

Equation (5) generalizes the notation in (4) to allow for tests in multiple grades. $S_{i,g}$ is the i th student's score on a test for a particular subject taken in grade g . $\tau_{i,g}$ is the i^{th} student's true academic achievement in that subject and grade. We drop subscript "T" to simplify notation, but maintain that a different test in a single occurrence is given in each grade and year. $\eta_{i,g}$ is the corresponding test measurement error from all sources, where $E\eta_{i,g} = 0$. Allowing for the possibility of heteroskedasticity,

$E\eta_{i,g}^2 = \sigma_{\eta_{i,g}}^2$ To simplify the analysis, we maintain that the measurement error variance for each student

⁴ Thorndike (1951, p. 568) provides a taxonomy characterizing different sources of measurement error. The framework also can be generalized to reflect students being grouped within schools and there being common random components of measurement error at that level.

is constant across grades; $\sigma_{\eta_{i,g}}^2 = \sigma_{\eta_i}^2, \forall g$. Let $\sigma_{\eta_i}^2$ equal $\sigma_{\eta_i}^2$ for all pupils in the homoskedastic case or, more generally, the mean value of $\sigma_{\eta_i}^2$ in the population of students. The v in (1) being uncorrelated implies that $E\eta_{i,g}\eta_{i,g'} = 0, \forall g \neq g'$ and $E\eta_{i,g}\tau_{i,g'} = 0, \forall g, g'$.

For a variety of reasons, researchers and policymakers are interested in the distribution of test scores across students. In such cases it is possible to decompose the overall variance of observed scores for a particular grade, $\sigma_{S_g}^2$, into the variance in universal scores across the student population, $\sigma_{\tau_g}^2$, and the measurement-error variance; $\sigma_{S_g}^2 = \sigma_{\tau_g}^2 + \sigma_{\eta_i}^2$. Here $K_g = \sigma_{\tau_g}^2 / \sigma_{S_g}^2$ is the *generalizability coefficient* measuring the portion of the total variation in observed scores that is explained by the variance of universal scores. The reliability coefficient is the comparable measure in classical test theory. As discussed below, we normalize test scores to have zero means and unit standard deviations, so that $\sigma_{S_g}^2 = 1 = \sigma_{\tau_g}^2 + \sigma_{\eta_i}^2$. In this case, the generalizability coefficient equals $K_g = 1 - \sigma_{\eta_i}^2$.

The distribution of observed scores from a test of student achievement will differ from the distribution of true student learning because of the errors in measurement inherent in testing. Psychometricians long have worried how such measurement error impedes the ability of educators to assess the academic achievement, or growth in achievement, of individual students and groups of students. This measurement error is less problematic for researchers carrying out analyses where test scores, or gain scores, are the dependent variable, as the measurement error will only affect the precision of parameter estimates, a loss in precision (but not consistency) which can be overcome with sufficiently large numbers of observations.⁵

Even though test measurement error does not complicate the estimation of how a range of factors affect student learning, such errors in measurement do have important implications when judging the sizes of those estimated effects. A standard approach in empirical analyses is to judge the sizes of estimated effects relative to either the standard deviation of the distribution of observed scores, $\sigma_{S_g}^2$, or the standard deviation of observed gain scores. From the perspective that the estimated effects shed light

⁵ Measurement error in lagged test scores entering as right-hand-side controls will result in biased estimates of the coefficients for lagged score. In particular, the estimated coefficients typically will be less than one, consistent with observed scores being subject to regression to the mean. However, measurement error in the lagged scores will affect the consistency of the estimates of other explanatory variables only if the measurement error is correlated with these variables. See Sullivan (2001) for a useful discussion of estimating regression models with heteroskedastic measurement error and Jacob and Lefgren (2005) for an application focusing on measurement error in the estimation of teacher effects. The method we employ to estimate the overall extent of test measurement error and how the measurement error variance differs across students could be used along with the methods discussed in these papers to deal with the errors in variables associated with included test scores as right-hand-side variables in regression equation.

on the extent to which various factors can explain systematic differences in student learning, not test measurement error, the sizes of those effects should be judged relative to the standard deviation of universal scores or the standard deviation of universal score gains. In most cases, it is the latter that is pertinent.

At a point in time, a student's universal score will reflect the history of all those factors affecting the student's cumulative, retained learning. This includes early childhood events, the history of family and other environmental factors, the historical flow of school inputs, etc. . The standard deviation of the universal score at a point in time reflects the causal linkages between all such factors and the dispersion in these varied and long-run factors. From this perspective, almost any short-run intervention – say a particular feature of a child's education during one grade – is likely to move a student by only a modest amount up or down in the overall distribution of universal scores. Of course, this in part depends upon the extent to which the test focuses on current topics covered, or draws upon prior knowledge and skills.⁶ The nature of the relevant comparison depends upon the question. For example, if policymakers want to invest in policies that provide at least a minimum year-to-year student achievement growth, for example to comply with NCLB in a growth context, then the relevant metric is the standard deviation in the gain in universal scores. However, if policymakers are interested in the extent to which an intervention may close the achievement gap, then comparing the effect of that intervention to the standard deviation of the universal score provides a better metric of improvement. Even in the latter case, it is important to keep in mind that interventions often are short lived when compared to the period over which the full set of factors affect cumulative achievement.

We now turn to the issue of distinguishing between the gain in universal scores reflecting the underlying achievement growth and the measured test score gain. Equation (6) shows that a student's observed test score gain in a subject between grades $g - 1$ and g , $\Delta S_{i,g}$, differs from the student's underlying achievement gain, $\Delta \tau_{i,g} = \tau_{i,g} - \tau_{i,g-1}$, because of the measurement error associated with

$$\Delta S_{i,g} = S_{i,g} - S_{i,g-1} = (\tau_{i,g} - \tau_{i,g-1}) + (\eta_{i,g} - \eta_{i,g-1}) = \Delta \tau_{i,g} + \Delta \eta_{i,g} \quad (6)$$

both tests, $\Delta \eta_{i,g} = \eta_{i,g} - \eta_{i,g-1}$. Here the variance of the gain-score measurement error for a pupil is

$\sigma_{\Delta \eta_{i,g}}^2 = 2\sigma_{\eta}^2$ when the measurement error is uncorrelated, and has constant variance, across grades.

Going from an individual student to the distribution of test score gains for the population of students, it is possible to decompose the distribution's overall variance; $\sigma_{\Delta S_g}^2 = \sigma_{\Delta \tau_g}^2 + \sigma_{\Delta \eta_g}^2$ where $\sigma_{\Delta \tau_g}^2$ is

⁶ This might help explain the result noted in footnote 1; standardized tests often measure cumulative learning whereas tests designed for a specific topic may measure the growth in learning targeted by a particular intervention.

the variance of the universal score growth in the population of students and $\sigma_{\Delta\eta}^2$ is the mean value of $\sigma_{\Delta\eta_i}^2$. Here $K_g^\Delta = \sigma_{\Delta\tau_g}^2 / \sigma_{\Delta S_g}^2$ is the proportion of the overall variance in gain scores that actually reflects variation in students' underlying growth in educational achievement. In general, $K_g^\Delta = \sigma_{\Delta\tau_g}^2 / \sigma_{\Delta S_g}^2$ will be smaller than $K_g = \sigma_{\tau_g}^2 / \sigma_{S_g}^2$ so that test measurement error is especially problematic when analyzing achievement growth.⁷

New York State Tests

We analyze math test scores of New York City students in grades three through eight for the years 1999 through 2007. Prior to 2006, New York State administered examinations in mathematics and English language arts for grades four and eight. In addition, the New York City Department of Education tested 3rd, 5th, 6th and 7th graders in these subjects. All the exams are aligned to the New York State learning standards and IRT methods were used to convert raw scores (e.g., number or percent of questions correctly answered) into scale scores. New York State began administering all the tests in 2006, with a two-step procedure used to obtain scale scores that year. First, for each grade, a temporary raw score to scale score conversion table was determined and the cut score was set for Level III (i.e., “the minimum scale score needed to demonstrate proficiency”). The temporary scale scores were then transformed to have a common scale across grades, with a state-wide standard deviation of 40 and a scale score of 650 reflecting the Level III cut score for each grade.⁸ Scale scores in 2007 were “anchored” using IRT methods so as to be comparable to the scale-score metric used for each grade in 2006.⁹ Even though efforts were made to anchor cut points prior to 2006, there appears to be some variation in how reported scale-scores were centered. However, the dispersion in scale scores varies little across grades and years. For example, the grade-by-year standard deviations for the years prior to 2006 have an average of 40.3, almost identical to that in 2006 and 2007, and a coefficient of dispersion of only 0.0445; the average absolute differences from the mean standard deviation is less than five percent of the mean. Given these properties, we normalize the test scores by grade and year, with little, if any, loss in useful information.¹⁰

⁷ This point has been made in numerous publications. See, for example, Ballou (2002). Rogosa () points out that there are circumstances in which the reliability of gain scores is not substantially smaller than that for the scores upon which the measure of gains is based.

⁸ CTB/McGraw-Hill (2006).

⁹ CTB/McGraw-Hill (2007).

¹⁰ Rothstein (2007, p. 12) makes the point that when scores are measured on an interval scale, normalizing those scores by grade “can destroy any interval scale unless the variance of achievement is indeed constant across grades.” Even though the variance in the underlying achievement may well vary (e.g., increase) as students move through grades, the reality is that the New York tests employ test scales having roughly constant variance. Thus, our normalizing scores are of little, if any, consequence.

Technical reports produced by test vendors provide information regarding test measurement error as defined in classical test theory and the IRT framework. For both, the focus is on the measurement error associated with the test instrument (e.g., the selection of test items and the scale-score conversion). The documents for the New York State tests report reliability coefficients that range from 0.88 to 0.95 and average 0.92, indicating that 0.08 percent of the variation in the scores for a test reflect measurement error associated with features of the test. However, in addition to only reflecting one aspect of measurement error, other factors limit the usefulness of these reliability estimates for our purpose. First, reported statistics are for the population of students statewide. Differences in student composition will mean that measures of reliability will differ to an unknown degree for New York City. More importantly, the reliability measures are with respect to raw scores, not the scaled-test scores typically employed in VA analyses. As a result of the nonlinear mapping between raw and scaled scores, a given raw-score increase yields quite different increases in scaled scores, depending upon the score level. For example, consider a one point increase in the raw score (e.g., one additional question being answered correctly) on the 2006 fourth grade math exam. At raw scores of 8, 38 and 68, respectively, a one point increase translates into scale-score increases of 12, 2 and 22 points. Even if the variance or standard error of measurement is constant across the range of raw scores, as assumed in classical test theory used to produce reliability coefficients in the technical reports, this would not be the case for scaled scores.

The technical reports provide estimates of the standard errors of measurement (SEM) for the scaled scores. These estimates have a conceptual foundation that differs from classical test theory in that an IRT framework is employed. Even so, the results may well be of general interest, as SEM estimates for a given test, based upon IRT, test theory and generalizability theory, have been found to have similar values.¹¹ The technical documents for New York report IRT standard errors of measurement for every scaled-score value. Reflecting our normalizations of scale-scores discussed above, we normalize the SEM and average over the grades and years. The dashed line in Figure 1 shows how the corresponding variances (i.e., SEM^2) differ across the range of true score values. We estimate the weighted mean value of the variance value to be 0.102 where the weights are the relative frequencies of students having the various scores.

Even though this estimate is a lower bound for the measurement error variance when all aspects of measurement error are considered, it is instructive to use this information to infer upper-bound estimates of $\sigma_{\tau}^2 = \sigma_S^2 - \sigma_{\eta_{\bullet}}^2$ and $\sigma_{\Delta\tau}^2 = \sigma_{\Delta S}^2 - \sigma_{\Delta\eta_{\bullet}}^2 = \sigma_{\Delta S}^2 - 2\sigma_{\eta_{\bullet}}^2$. With $\sigma_S^2 = 1$, $\hat{\sigma}_{\Delta S}^2 = 0.397$ and 0.102 being a lower-bound estimate of $\sigma_{\eta_{\bullet}}^2$, 0.898 and 0.193 are upper-bound estimates of σ_{τ}^2 and $\sigma_{\Delta\tau}^2$,

¹¹ Lee, Brennan and Kolen (2000).

respectively. Thus, effect sizes measured in relation to $\sigma_{\Delta\tau}$ are more than twice as large as effect sizes measured in relation to σ_S^2 . (Our estimate of $\sigma_{\Delta\tau}$ is $0.439 = \sqrt{0.193}$.) By contrast, σ_S is 1.0, which is 2.27 times as large as $\sigma_{\Delta\tau}$.)

The above estimate of the measurement error variances associated with the test instrument may well be substantially below the overall measurement error variance, σ_n^2 , not a reasonable point estimate. As noted in footnote 4, Thorndike (1951) provides a useful, detailed classification of factors that contribute to test measurement error. To a large degree, these fall within the framework outlined above where the measurement error is associated with the selection of test items included in a test, the timing (occurrence) of the test and these factors *crossed with* students. Reliability coefficients based on the test-retest approach using parallel test forms is recognized in the psychometric literature as being the gold standard for quantifying the measurement error from all these sources. Students take alternative, but parallel (i.e., interchangeable), tests on two or more occurrences sufficiently separated in time so as to allow for the “random variation within each individual in health, motivation, mental efficiency, concentration, forgetfulness, carelessness, subjectivity or impulsiveness in response and luck in random guessing”¹² but sufficiently close in time that individuals’ knowledge, skills and abilities being tested are unchanged. However, we know of only one application of this method in the case of achievement tests like those considered here.¹³

Rather than analyzing the consistency of student test scores over time, the standard approach used by test vendors is to divide the test taken at a single point in time into what is hoped to be parallel parts. Reliability is then measured with respect to the consistency (i.e., correlation) of students’ scores across these parts. Psychometricians have developed reliability measures that reflect the number of test parts and the types of questions included on the test. As Feldt and Brennan note (1989), such approaches “frequently present a biased picture” in that “reported reliability coefficients tend to overstate the trustworthiness of educational measurement, and standard errors underestimate within-person variability,” the problem being that measures based on a single test occurrence ignore potentially important day-to-day differences in student performance.

In the following section, we describe a method for obtaining what we believe is a credible point estimate of σ_n^2 and, in turn, a point estimate of the standard deviation of gain scores net of measurement error needed to compute effect sizes. The method accounts for test measurement error from all sources.

¹² Feldt and Brennan (1989).

¹³ Rothstein (2007) discusses results from a test-retest reliability analysis based upon 70 students in North Carolina.

Analyzing the Overall Measurement-Error Variance.

Using vector notation, $S_i = \tau_i + \eta_i$ where $S_i' = [S_{i,3} \ S_{i,4} \ \cdots \ S_{i,8}]$, $\tau_i' = [\tau_{i,3} \ \tau_{i,4} \ \cdots \ \tau_{i,8}]$, and $\eta_i' = [\eta_{i,3} \ \eta_{i,4} \ \cdots \ \eta_{i,8}]$. The entries in each vector reflect us having test scores for grades three through eight. Let $\Omega(i)$ represent the auto-covariance matrix for the i^{th} student's observed test scores;

$$\Omega(i) = E(S_i S_i') = E(\tau_i \tau_i') + E(\eta_i \eta_i') = \Gamma + \sigma_{\eta_i}^2 I \quad (7)$$

where Γ is the auto-covariance matrix for the universal scores and I is a 6x6 identity matrix. For the population of all students, $\Omega_{\bullet} = E\Omega(i) = \Gamma + \sigma_{\eta_{\bullet}}^2 I$ where $\sigma_{\eta_{\bullet}}^2 = E\sigma_{\eta_i}^2$ is the mean measurement error variance in the population. Here $\Omega(i)$ is assumed to differ from $\Omega(i')$ only because of possible heteroskedasticity in the measurement error across students; Γ and, therefore, the off diagonal elements of $\Omega(i)$ are assumed to be constant across students.

We employ test-score data from New York City to estimate the empirical counterpart of Ω_{\bullet} , $\tilde{\Omega}_{\bullet} = \sum_i S_i S_i'$. Even though auto-covariance matrices typically reflect settings having equal-distant time intervals (e.g., annual measures), here we consider test scores of students across grades. Whereas this distinction is without consequence for students making normal grade progressions, this is not true when students repeat grades. Multiple test scores for repeated grades complicate the computation of $\tilde{\Omega}_{\bullet}$ since only one score per grade is included in our formulation of S_i . We deal with this complication employing three different approaches, computing $\tilde{\Omega}_{\bullet}$ using: (1) the scores of students on their first taking of each exam, (2) the scores on their last taking of each exam or (3) pair-wise comparisons of the score on the last taking in grade g and the score on the first taking in grade $g+1$, $g = 3, 4, \dots, 8$. Because the three methods yield almost identical results, we only present estimates based on the first approach, using the first score of students in each grade.

A second complication arises because of missing test scores. The extent to which this is a problem depends upon the reasons for the missing data. If scores are missing completely at random, there is little problem.¹⁴ However, this does not appear to be the case. In particular, we find evidence that lower-scoring and, to a lesser degree, very high scoring students are more likely to have missing exam scores. For example, the dashed line in Figure 2 shows the distribution of fifth-grade math scores of students for whom we also have sixth grade scores. In contrast, the solid line shows the distribution of fifth-grade scores for those students for whom grade-six scores are missing. The higher right tail in the latter distribution is explained by some high-scoring students skipping the next grade. Consistent with

¹⁴ Reference

this explanation, many of these students took the fifth-grade exam one year and the seventh-grade exam the following year. However, it is more common that those with missing scores scored relatively lower in the grades where scores are present. To avoid statistical problems associated with this systematic pattern of missing scores, we impute values of missing scores using SAS Proc MI, first using the Markov Chain Monte Carlo procedure to

Describe the method used to impute missing scores.

Table 1 shows the estimated auto-covariance matrix, $\tilde{\Omega}_\bullet$, for students in the cohorts entering the third grade in years 1999 through 2005. With the exception of third grade scores, the estimates are consistent with stationarity in the auto-covariances. For example, consider the auto-covariance measures for scores in adjacent grades, $Cov(S_{i,g}, S_{i,g+1})$, starting in grade four (i.e., 0.7975, 0.7813, 0.7958, and 0.7884). The range of these values is only two percent of the mean value (0.7908). A similar pattern hold for two- and, to a lesser degree, three-grade lags in scores. This stationarity meaningfully reduces the number of parameters needed to characterize Ω_\bullet . In particular, let $\omega^s \equiv Cov(S_{i,g}, S_{i,g+s})$, $s = 1, 2, \dots, 4$, starting with grade four. Estimates of these measures are shown in Table 3, along with the estimate of $\omega^0 = V(S_{i,g}) = \gamma^0 + \sigma_{\eta_\bullet}^2$. In the following section, we describe the approach used to estimate both $\sigma_{\eta_\bullet}^2$ and γ^0 , which yields an estimate of the variance in the gain in universal scores;

$$V(\tau_{i,g+1} - \tau_{i,g}) = 2(\gamma^0 - \gamma^1) = 2(\gamma^0 - \omega^1). \quad \text{Alternatively, } \sigma_{\Delta\tau}^2 = \sigma_{\Delta S}^2 - 2\sigma_{\eta_\bullet}^2$$

Our estimation strategy draws upon an approach commonly used to study the covariance structure of individual- and household-level earnings, hours worked and other panel-data time-series.¹⁵ However, our application differs in one important way; here we focus on test scores of students across grades, whereas other applications typically consider equal-distant time intervals (e.g., annual measures).

Our Approach We assume the time-series pattern of universal scores for each student is as shown in equation (8).

$$\tau_{i,g} = \beta\tau_{i,g-1} + \theta_{i,g} \quad (8)$$

This first-order autoregressive (AR(1)) structure models student attainment in grade g as being a cumulative process with the prior level of knowledge and skills subject to decay if $\beta < 1$. Repeated

substitution yields $\tau_{i,g} = \beta^g \tau_{i0} + \sum_{s=1}^g \beta^{g-s} \theta_{i,s}$ where τ_{i0} is the initial condition. In the special case where

¹⁵ The approach developed by Abowd and Card (1989) has been applied and extended in numerous papers.

$\beta = 1$, $\theta_{i,g}$ is the student's gain in achievement while in grade g .¹⁶ This special case is the basic structure maintained in many value-added analyses, including the layered model employed by Sanders. Models allowing for decay are discussed by McCaffrey et. al. (2004) as well as Rothstein (2007).

Equation (8) and the statistical structure of the $\theta_{i,g}$ (i.e., $\theta_{i,1}, \theta_{i,2}, \dots$) together determine the dynamic pattern of the universal scores as reflected in the parameterization of $\Gamma = E(\tau_i \tau_i')$ which, given stationarity, is completely characterized by $\gamma^0, \gamma^1, \dots, \gamma^4$ where $\gamma^s = E\tau_{i,g} \tau_{i,g+s}$. Before considering a specific specification of the $\theta_{i,g}$ and the corresponding structure of Γ , several general implications of stationarity are relevant. First, stationarity in $E\tau_{i,g}^2 = \gamma^0$ and $E\tau_{i,g} \tau_{i,g+1} = \gamma^1$ implies that $\psi^1 \equiv E\tau_{i,g} \theta_{i,g+1} = \gamma^1 - \beta\gamma^0$ is also stationary. The same is true for $\psi^s \equiv E\tau_{i,g} \theta_{i,g+s} = \gamma^s - \beta\gamma^{s-1}$, $s > 0$. This stationarity and equation (8) imply the structure of the unique elements of Ω_\bullet shown in (9)

$$\begin{aligned}
\omega^0 &\equiv E(S_{i,g}^2) \\
&= \gamma^0 + \sigma_{\eta}^2 \\
\omega^1 &\equiv E(S_{i,g} S_{i,g+1}) = E(\tau_{i,g} + \eta_{i,g})(\tau_{i,g+1} + \eta_{i,g+1}) = E(\tau_{i,g} + \eta_{i,g})(\beta\tau_{i,g} + \theta_{i,g+1} + \eta_{i,g+1}) \\
&= \beta\gamma^0 + \psi^1 \\
\omega^2 &\equiv E(S_{i,g} S_{i,g+2}) = E(\tau_{i,g} + \eta_{i,g})(\beta^2\tau_{i,g} + \beta\theta_{i,g+1} + \theta_{i,g+2} + \eta_{i,g+2}) \\
&= \beta^2\gamma^0 + \beta\psi^1 + \psi^2 \\
\omega^3 &\equiv E(S_{i,g} S_{i,g+3}) = E(\tau_{i,g} + \eta_{i,g})(\beta^3\tau_{i,g} + \beta^2\theta_{i,g+1} + \beta\theta_{i,g+2} + \theta_{i,g+3} + \eta_{i,g+3}) \\
&= \beta^3\gamma^0 + \beta^2\psi^1 + \beta\psi^2 + \psi^3 \\
\omega^4 &\equiv E(S_{i,g} S_{i,g+4}) = E(\tau_{i,g} + \eta_{i,g})(\beta^4\tau_{i,g} + \beta^3\theta_{i,g+1} + \beta^2\theta_{i,g+2} + \beta\theta_{i,g+3} + \theta_{i,g+4} + \eta_{i,g+4}) \\
&= \beta^4\gamma^0 + \beta^3\psi^1 + \beta^2\psi^2 + \psi^3 + \psi^4
\end{aligned} \tag{9}$$

We consider alternative specifications of the $\theta_{i,g}$, the ψ^s and, in turn, the structure of the ω^s .

Model 1: Consider an individual-effects specification for the $\theta_{i,g}$; $\theta_{i,g} = \mu_i + \varepsilon_{i,g}$ where μ_i is an random student effect with $E\mu_i = 0$ and $E\mu_i^2 = \sigma_{\mu}^2$. $\varepsilon_{i,g}$ is a white-noise random error; $E\varepsilon_{i,g} = 0$, $E\mu_i \varepsilon_{i,g} = 0$, and $E\tau_{i,0} \varepsilon_{i,g} = 0$. Also, $E\varepsilon_{i,g} \varepsilon_{i,g'} = 0 \quad \forall g \neq g'$. This structure implies that $\psi^s = E\tau_{i,g} \theta_{i,g+s} = E\tau_{i,g} (\mu_i + \varepsilon_{i,g+s}) = E\tau_{i,g} \mu_i \equiv \lambda$ for all $s > 0$ as well as the test-score auto-covariances shown in (10).

¹⁶ We will generally refer to $\theta_{i,g}$ as the student's achievement gain. However, when prior achievement is subject to decay ($\beta < 1$), $\theta_{i,g}$ is the gain in achievement gross of that decay; $\theta_{i,g} = S_{i,g+1} - S_{i,g} + (1 - \beta)\tau_{i,g}$.

$$\begin{aligned}
\omega^0 &= \gamma^0 + \sigma_{\eta}^2 \\
\omega^1 &= \beta\gamma^0 + \lambda \\
\omega^2 &= \beta^2\gamma^0 + (\beta+1)\lambda \\
\omega^3 &= \beta^3\gamma^0 + (\beta^2 + \beta + 1)\lambda \\
\omega^4 &= \beta^4\gamma^0 + (\beta^3 + \beta^2 + \beta + 1)\lambda
\end{aligned} \tag{10}$$

This model includes two pertinent special cases. First, if $\mu_i = 0, \forall i$, then $\theta_{i,g} = \varepsilon_{i,g}$; the grade-level gains of students are independent across grades. This implies that $\lambda = 0$ and that the equations in (10) reduce to $\omega^s = \beta^s \gamma^0, s = 1, 2, 3, 4$. Second, if $\theta_{i,g} = \mu_i + \varepsilon_{i,g}$ but there is no decay in prior achievement ($\beta = 1$), the test-score auto-covariances are of the form $\omega^s = \gamma^0 + s\lambda$.

Model 2: A complication arises when we specify more general statistical structures of the $\theta_{i,g}$ and infer the implied structures of ψ^s and $\omega^0, \omega^1, \dots$; stationarity along with the recursive nature of (2) result in the elements of the test-score auto-covariance matrix (e.g., ω^s) being quite complex functions of the models' underlying parameters. This leads us to consider the alternative approach of specifying a reduced-form parameterization of the ψ^s in (9). In particular, we consider the case where $\psi^s = \frac{\Psi}{s^\alpha}$, which implies the test-score auto-covariance structure shown in (11). This specification includes Model 1 as a special case where $\alpha = 0$. In the case of $\alpha \rightarrow \infty$, (11) reduces to the case where $\omega^s = \beta^s \gamma^0 + \beta^{s-1}\psi$, which is the structure resulting from $\theta_{i,g}$ being a first-order moving average (i.e., $\theta_{i,g} = \varepsilon_{i,g} + \pi \varepsilon_{i,g-1}$).

$$\begin{aligned}
\omega^0 &= \gamma^0 + \sigma_{\eta}^2 \\
\omega^1 &= \beta\gamma^0 + \psi \\
\omega^2 &= \beta^2\gamma^0 + \beta\psi + \frac{\Psi}{2^\alpha} \\
\omega^3 &= \beta^3\gamma^0 + \beta^2\psi + \beta\frac{\Psi}{2^\alpha} + \frac{\Psi}{3^\alpha} \\
\omega^4 &= \beta^4\gamma^0 + \beta^3\psi + \beta^2\frac{\Psi}{2^\alpha} + \beta\frac{\Psi}{3^\alpha} + \frac{\Psi}{4^\alpha}
\end{aligned} \tag{11}$$

Let χ represent the vector of unknown parameters for a model we wish to estimate, where $\omega(\chi) \equiv \left[\omega^0(\chi) \ \omega^1(\chi) \ \omega^2(\chi) \ \omega^3(\chi) \ \omega^4(\chi) \right]$. For example, $\chi \equiv \left[\sigma_{\eta}^2 \ \gamma^0 \ \beta \ \lambda \right]$ in (10) for Model 1. Let $\hat{\omega} \equiv \left[\hat{\omega}^0 \ \hat{\omega}^1 \ \hat{\omega}^2 \ \hat{\omega}^3 \ \hat{\omega}^4 \right]$ represent the empirical counterpart of the unique elements of auto-

covariance matrix Ω_{\bullet} , i.e., $\tilde{\Omega}_{\bullet}$, shown in Table 3. We estimate χ using a minimum distance estimator; $\hat{\chi}$ is the value of χ that minimizes the distance between $\omega(\chi)$, and $\hat{\omega}$ as measured by

$Q = (\hat{\omega} - \omega(\chi))(\hat{\omega} - \omega(\chi))' = \sum_j (\hat{\omega}^j - \omega^j(\chi))^2$. This *equally weighted minimum distance estimator* is widely used, including empirical analyses of the auto-covariance structure of earnings.¹⁷

The intuition behind our approach for isolating the extent of measurement error is relatively straightforward. We illustrate this using the test-score covariance structure of Model 1 in (10).

Substituting the computed values of $\hat{\omega}^s$ for the ω^s in (10) yields a system of five equations in four unknowns. If the last equation were dropped from the analysis, it would be possible to solve the system of four equations to obtain estimates of $\chi \equiv [\sigma_{\eta_{\bullet}}^2 \ \gamma^0 \ \beta \ \lambda]$.¹⁸ The equations for ω^1 , ω^2 , and ω^3 , would imply estimates of β , λ and γ^0 . In turn, the estimate of γ^0 along with the first equation would yield an estimate of $\sigma_{\eta_{\bullet}}^2$. This illustrates the importance of two key assumptions. First, identification requires the universal test scores to reflect a cumulative process in which there is some degree of persistence (i.e., $\beta > 0$). Second, there is no persistent (correlation) in the test measurement error across grades.

Together, these assumptions allow us to isolate the overall extent of test measurement error.

Note that an alternative estimation strategy would be to directly estimate student growth models with measurement error using a hierarchical model estimation strategy. Compared to this strategy, our approach has several advantages. First, having well in excess of a million student records, estimating an HLM would be a computational challenge. Instead, we simply compute $\hat{\omega}$ and then need only minimize Q . Using this approach, estimating the alternative specifications is quite easy. Second, estimating models that allow for decay (i.e., $\beta < 1$) is straightforward using the minimum-distance estimator, which would not be the case using standard HLM software. Finally, other than the assumptions regarding first and second moments discussed above, the minimum-distance estimator does not require us to assume the distributions from which the various random components are drawn. Such explicit assumptions are integral to the hierarchical approach.

Results Parameter estimates for the alternative models discussed above are shown in Table 4. The first column corresponds to Model 1 and the specification shown in (10). Estimates in the second column (Model 1a) are for the case where the grade-level gains for each student are assumed to be

¹⁷ See Cameron and Trivedi (2005, pp. 202-203) for a general discussion of minimum distance estimators. The appendix in Abowd and Card (1989) discuss these estimators in the context of estimating the auto-covariance of earnings.

¹⁸ It is the over-identification of parameters in the full set of equations that leads us to estimate the parameters by minimizing Q .

independent across grades, implying that $\psi^s = \lambda = 0$. Model 1b employs the student-effect specification $\theta_{i,g} = \mu_i + \varepsilon_{i,g}$ as in Model 1 but maintains that there is no decay in prior achievement (i.e., $\beta = 1$). Finally, estimates in the last column of Table 3 are for the specification in (11), which includes the other three models as special cases.¹⁹

Note the meaningful difference in the estimates of β and ψ across the four sets of estimates. The qualitative differences are the result of the stationarity in test-score variances across grades. Given the time-series pattern of test scores maintained in equation (8), it follows that

$$E\tau_{i,g}^2 = \beta^2 E\tau_{i,g-1}^2 + E\theta_{i,g}^2 + 2\beta E\tau_{i,g-1}\theta_{i,g}. \text{ Stationarity of } E\tau_{i,g}^2 = E\tau_{i,g-1}^2 = \gamma^0 \text{ implies that } (1 - \beta^2)\gamma^0 = \sigma_\theta^2 + 2\beta\psi,$$

which implies a relationship between β and ψ . For example, when $\beta = 1$, $\psi = -\sigma_\theta^2/2 \leq 0$. Thus, $\hat{\psi} \leq 0$ in

Model 1b is to be expected and is consistent with the estimate $\hat{\beta} = 1.009$ in Model 2. Model 2 includes the other specifications as special cases. However, the parameter estimates in Model 2 differ in important ways from estimates in the other models and yield a better fit between $\omega(\chi)$, and $\hat{\omega}$ as measured by

$$Q = \sum_j \left(\hat{\omega}^j - \omega^j(\chi) \right)^2. \text{ Thus, in the following analysis, we will employ parameter estimates for Model 2.}$$

We estimate the parameters of the alternative auto-covariance structures to quantify the overall test measurement error measured by $\sigma_{\eta_\bullet}^2$. In this regard, the results in Table 4 are quite robust. In the four specifications, as well as other specifications estimated but not reported here, estimates of $\sigma_{\eta_\bullet}^2$ all fell within the 0.16 to 0.18 range, increasing our confidence that approximately 17 percent of the overall dispersion in the NYS tests is attributable to various forms of test measurement error.

Our estimates of $V(\tau_{i,g})$ and $V(S_{i,g})$ -- that is $\hat{\gamma}^0 = 0.827$ and $\hat{\omega}^0 = 0.992$ -- imply the overall generalizability coefficient is estimated to be $\hat{K}_g = \frac{\hat{\sigma}_\tau^2}{\hat{\sigma}_S^2} = \frac{\hat{\gamma}^0}{\hat{\omega}^0} = 0.834$. This is meaningfully smaller than the reliability coefficients, approximately equal to 0.90, reported in the test technical reports and implied by the reported (IRT) standard errors of measurement discussed above. A technical report for North

¹⁹ The standard errors reported in Table 4 are the square roots of the diagonal elements of the estimated covariance matrix of $\hat{\chi}$, $V(\hat{\chi}) = [D'D]^{-1} [D'V(\hat{\omega})D] [D'D]^{-1}$. Here D is the first derivative of $\omega(\chi)$ with respect to χ evaluated at $\hat{\chi}$. Standard deviations for the parameter estimates in model 2 are not reported because of computational problems associated with $D'D$ being close to singular; in this case the determinant $|D'D|$ equals 5.31E-15. In contrast, $|D'D|$ equals 8.55E-3, 10.3 and 20.0 for models 1, 1a and 1b, respectively.

Carolina's reading test reports a test-retest reliability equal to 0.86,²⁰ only slightly larger than our estimate. Even so, it is important to note that the North Carolina estimate was based on an analysis of 70 students and is for a test that may well differ in important ways for the New York tests.

Our primary goal here is to obtain credible estimates of the overall measurement-error variance, so that we can infer an estimate of the standard deviation of students' universal-score gains measuring growth in skills and knowledge. Utilizing Model 2 estimates, we calculate the variance of gain scores net of measurement error to be 0.067; $\hat{\sigma}_{\Delta\tau}^2 = \hat{\sigma}_{\Delta S}^2 - 2\hat{\sigma}_{\eta_s}^2 = 0.398 - 2 \cdot 0.165 = 0.067$. Thus, we estimate the standard deviation of universal gain scores to be 0.259, indicating that effect sizes based on the dispersion in the gains in actual student achievement are roughly four times as large as those typically reported. Here it is useful to summarize how we come to this conclusion. Comparing the magnitudes of effects relative to the standard deviation of observed score gains, $\hat{\sigma}_{\Delta S} = 0.63$, rather than the standard deviation of observed scores, $\sigma_s \approx 1.0$, would result in effect size estimates being roughly 50 percent larger. Thus, most of the four-fold increase results from accounting for the test measurement error, i.e., employing $\hat{\sigma}_{\Delta\tau}^2 = 0.067$ rather than $\hat{\sigma}_{\Delta S} = 0.63$ as the measure of gain score dispersion. This large difference reflects that less than 20 percent of the dispersion in gain scores is actually attributable to the dispersion of academic achievement gains.²¹

We have focused on the mean measurement error variance for the population of students, $\sigma_{\eta_s}^2$, because of its importance in calculating effect sizes. However, we are also interested in the extent to which measurement error varies across students. This can be estimated in a relatively straightforward manner. Equation (8) implies that the variance of $S_{i,g+1} - \beta S_{i,g} = \theta_{i,g+1} + \eta_{i,g+1} - \beta\eta_{i,g}$ equals the expression shown in (13).

$$V(S_{i,g+1} - \beta S_{i,g}) = \sigma_{\theta}^2 + \sigma_{\eta_i}^2 + \beta^2 \sigma_{\eta_i}^2 = \sigma_{\theta}^2 + (1 + \beta^2) \sigma_{\eta_i}^2 \quad (13)$$

This, along with the formula $\sigma_{\theta}^2 = (1 - \beta^2)\gamma^0 - 2\beta\psi$ ²² and our estimates of $\sigma_{\eta_s}^2$, γ^0 , β , and ψ , imply the estimator of the measurement error variance for each student, $\hat{\sigma}_{\eta_i}^2$, shown in (14).

²⁰ At the same time, Sanford (1996) reports Coefficient alpha reliability coefficients for the reading comprehension exams in grades three through eight as ranging from 0.92 to 0.94. Thus, we see a large difference between the type of measure typically reported and the actual extent of measurement error.

²¹ $\hat{\sigma}_{\Delta\tau}^2 = 0.068$ implies that the generalizability coefficient for student gain scores, $\hat{K}^{\Delta} = (\hat{\sigma}_{\Delta\tau}^2 / \hat{\sigma}_{\Delta S}^2) = (0.068 / 0.397) = 0.169$, is much smaller than that for scores.

²² This follows from the formula $(1 - \beta^2)\gamma^0 = \sigma_{\theta}^2 + 2\beta\psi$ derived above.

$$\hat{\sigma}_{\eta_i}^2 = \frac{\left[\frac{1}{G-1} \sum_{g=1}^{G-1} (S_{i,g+1} - \hat{\beta} S_{i,g})^2 - \hat{\sigma}_{\theta}^2 \right]}{(1 + \hat{\beta}^2)} = \frac{\left[\frac{1}{G-1} \sum_{g=1}^{G-1} (S_{i,g+1} - \hat{\beta} S_{i,g})^2 - \left((1 - \hat{\beta}^2) \hat{\gamma}^0 - 2 \hat{\beta} \hat{\psi} \right) \right]}{(1 + \hat{\beta}^2)} \quad (14)$$

To explore how the measurement error varies across students, we assume that $\sigma_{\eta_i}^2$ for the i th student is a function of that student's mean universal score across grades, which we estimate using the student's mean test score, $\bar{S}_i = \frac{1}{G} \sum_g S_{i,g}$.

The solid line in Figures 1 shows the estimated relationship between the estimates $\hat{\sigma}_{\eta_i}^2$ and \bar{S}_i . Here the values of \bar{S}_i for all students are grouped into intervals of length 0.10 (e.g., values of \bar{S}_i between 0.05 and 0.15). The graph shows the mean value of $\hat{\sigma}_{\eta_i}^2$ for the students falling in each interval. In this way, the solid line in the graph is a simple non-parametric characterization of how the overall measurement error varies across the range of universal scores. As discussed above, the dashed line shows the average measurement error variance associated with the test instrument, as reported in the technical reports provided by the test vendors.

We find the similarity between the two curves in Figure 1 quite striking. In particular, our estimates of how the overall measurement error variance varies over the range of universal scores follows a pattern almost identical to that implied by the measurement error variances associated with test construction, as reported by the test vendors. The overall variance estimates are larger, consistent with there being multiple sources of measurement error, in addition to that associated with the test construction. It appears that the measurement error variance associated with these other factors is roughly constant across the range of achievement levels. The consistency of results from quite different strategies for estimating the level and pattern of the measurement error, increases our confidence in the method we have used to estimate the variance in universal score gains and, in turn, effect sizes.

Going beyond increasing our confidence in the statistical approach we have employed to estimate the extent of measurement error for the overall population of students, the relationship between the aggregate measurement error variance for individual students and their universal scores shown in Figure 2 allows us to estimate the distributions of universal scores and universal gain scores. For example, the more dispersed line in Figure 3 (short dashes) shows the distribution of gains in normalized scaled-scores between grades four and five. Because of the measurement error embedded in these gain scores, this distribution overstates the dispersion in the universal gain scores, $\Delta\tau_{i,5}$. Such gain scores can be “shrunk” using the empirical Bayes estimator, to account for the measurement error. The line with long dashes is the distribution of empirical Bayes estimates of universal gain scores, computed using the

formula $\Delta S_{i,5}^{EB} = G_i \Delta S_{i,5} + (1 - G_i) \overline{\Delta S}_5$ where $G_i^\Delta \equiv \sigma_{\Delta\tau}^2 / (\sigma_{\Delta\tau}^2 + \sigma_{\Delta\eta_i}^2)$ and $\overline{\Delta S}_5$ is the mean value of $\Delta S_{i,5}$. Even though the empirical Bayes estimator ($\Delta S_{i,5}^{EB}$) is the best linear unbiased estimator of the underlying parameters for individual students ($\Delta\tau_{i,5}$)²³, the empirical distribution of the empirical Bayes estimates understates the actual dispersion in the distribution of the parameters estimated. Thus, the empirical distribution of the $\Delta S_{i,5}^{EB}$ shown in Figure 3 understates the dispersion in the empirical distribution of universal gain scores, $F_N(z) = \sum_i I(\Delta\tau_{i,g} \leq z) / N$. As discussed by Carlin and Louis (1996), Shen and Louis (1998), and others, it is possible to more accurately estimate the distribution of $\Delta\tau_{i,5}$ by employing an estimator that minimizes the expected distance defined in terms of that distribution and some estimator \hat{F}_N . If $\Delta\tau_{i,5}$ and $\eta_{i,5}$ are normally distributed,

$$E[F_N(z)|S] = \sum_i \Phi\left(\frac{z - \Delta S_{i,5}^{EB}}{\sigma_{\eta_i} \sqrt{G_i^\Delta}}\right) / N. \text{ This motivates our use of the formula}$$

$\hat{F}_N(z)|S = \sum_i \Phi\left(\frac{z - \Delta S_{i,5}^{EB}}{\hat{\sigma}_{\eta_i} \sqrt{\hat{G}_i^\Delta}}\right) / N$ to estimate the empirical density of universal gain scores shown by the solid line in Figure 3.²⁴

To this point, our discussion of the importance of accounting for measurement error in the calculation of effect sizes has been in general terms. We apply the methods described above to estimates of the effects of teacher attributes to make the implications of these methods clear and to suggest that the growing perception among researchers and policymakers that observable attributes of teachers make little difference in true student achievement gains may be misleading.

An Analysis of Teacher Attribute Effect Sizes

In a recent paper, Boyd, Lankford, Loeb, Rockoff and Wyckoff (2007) use data for fourth and fifth grade students in New York City over the 2000 to 2005 period to estimate how the achievement gains of students in mathematics are affected by the qualifications of their teachers. The effect of teacher attributes were estimated using the specification shown in equation (15).

$$S_{ikgty} - S_{ik'g(g-1)t'(y-1)} = \gamma_0 + \gamma_1 Z_{iy} + \gamma_3 C_{ty} + \gamma_4 T_{ty} + \pi_i + \pi_g + \pi_y + \epsilon_{ikgty} \quad (15)$$

²³ $\Delta S_{i,5}^{EB}$ is the value of $\square\tau_{i,g}$ which minimizes the loss function $\sum_i (\Delta\tau_{i,g} - \square\tau_{i,g})^2$.

²⁴ An alternative would be to utilize the distribution of constrained empirical Bayes estimators. Louis (1984) and Ghosh (1992).

Here the standardized achievement gain score of student i in school \mathbf{k} in grade \mathbf{g} with teacher \mathbf{t} in year \mathbf{y} is a linear function of time-varying characteristics of the student \mathbf{Z} , characteristics of the other students in the same grade having the same teacher in that year \mathbf{C} , and the teacher's qualifications \mathbf{T} . The model also includes student, grade and time fixed effects and a random error term. The time-varying student characteristic is whether the student changed schools between years. Class variables include the proportion of students who are black or Latino, the proportion who receive free- or reduced-price school lunch, class size, the average number of student absences in the prior year, the average number of student suspensions in the prior year, the average achievement scores of students in the prior year, and the standard deviation of student test scores in the prior year. Teaching experience is measured by separate dummy variables for each year of teaching experience up to a category of 21 or more years. Other teacher qualifications include whether the teacher passed the general knowledge certification exam on the first attempt, certification test scores, whether and in what area the teacher was certified, the Barron's ranking of the teacher's undergraduate college, math and verbal SAT scores, the initial path through which the teacher entered teaching (e.g., a traditional college-recommended program or the New York City Teaching Fellows program) and an interaction term of the teacher's certification exam score and the portion of the class eligible for free lunch. The standard errors are clustered at the teacher level to account for multiple student observations per teacher.

Like much of the research briefly summarized in the introduction, Boyd et. al. (2007) find that while teacher experience is statistically significant and appears important, few of the other measures of teacher qualifications are statistically significant or of even modest effect sizes as traditionally measured. (See Table 6 for a full set of parameter estimates.) We reproduce the parameter estimates for selected measures of teacher attributes from Table 6 in the first column of Table 7. These estimated effects, measured relative to the standard deviation of observed student achievement scores, seem to indicate that none of the estimated effect sizes are large by standards often employed by educational researchers in other contexts (see, Hill et. al., 2007). However, at least the effect of not being certified, and a one standard deviation increase in math SAT scores, is comparable to about two-thirds of the gain that accrues to the first year of teaching experience, an effect which most observers believe is meaningful.

The second column of Table 7 shows the estimated effects as a ratio to the standard deviation of observed *gain* scores. As argued above, we believe that in many contexts the sizes of effects should be measured relative to the standard deviation of year-to-year gains, not the standard deviation of achievement. In the context of our analysis, estimated effect sizes measured relative to the standard deviation of observed gains are 59 percent larger than those based on the standard deviation of observed scores. The additional effect of accounting for measurement error in gain scores is shown in column 3 where we employ the estimates of the standard deviation of universal gain scores corresponding to Model

2 in Tables 4 and 5; $\sigma_{\eta_e} = 0.2608$. Netting out test measurement error, we see the effect sizes estimates for teacher attributes are substantially larger. For example, the effect of a student having a second year teacher, rather than teacher having no prior experience, is estimated to be about a quarter of a standard deviation in the (universal) achievement gain experienced by students. Although somewhat smaller, the effect of having an uncertified teacher, or a teacher with a one standard deviation lower math SAT, is about 16 percent of the standard deviation of the gain in achievement net of measurement error.

Finally, Boyd et. al. (2007) examine the joint effect of all observable attributes of teachers as described in the first paragraph of this section by using the estimated model to predict the value-added for each student based only on these observable attributes, holding all of the other variables in Table 6 constant. The estimates for teachers in the poorest quartile of schools are divided into quintiles based on their predicted value-added. The difference between the highest and lowest quintiles is 0.16 (0.11 when experience is held constant). Recall that this estimate is relative to the standard deviation of observed scores. When the estimated effect is adjusted to account for test measurement error, the effect size is much greater than a half a standard deviation of the universal score gain.

Summary

Understanding the implications of VAM estimates is important as they are increasingly being employed to inform policy decisions. Estimates of the effects of teacher attributes commonly employed in VAM estimates using state and district student achievement tests are frequently small or statistically insignificant. In this paper we explore the role that measurement error plays in creating the perception that observable attributes of teachers matter little. We believe this paper makes two important contributions. First, we have laid out a relatively simple approach for estimating test measurement error from all sources and calculating the standard deviations of universal scores and universal gain scores. Second, when applied to estimates of the effect of teacher attributes commonly observed in the literature, we find that many of these attributes are responsible for important gains in student achievement.

Our approach for estimating the test measurement error variance for the student population of interest, as well as how the variance varies across students, is possible to the extent that (1) the random components in test scores for each student associated with test measurement error are not correlated across grades and (2) the grade-to-grade gains in student achievement are to some extent persistent (i.e., $\beta > 0$). In such settings, it is possible to specify relatively general structures for the auto-covariance of observed test scores, for which the underlying parameters can be estimated in a relatively straightforward manner, yielding estimates of the overall extent of test measurement error. In turn, this allows us to

quantify the dispersion (e.g., standard deviation) in student achievement as measured by universal scores as well as the dispersion in universal gain scores.

We apply these methods to a recent paper that reports VAM estimates of various teacher attributes (Boyd et. al., 2007). Many of these estimates appear small when compared to the standard deviation of student achievement, that is effect sizes of less than .05. However, when measurement error is taken into account the associated effect sizes are often about 0.16. Furthermore, when teacher attributes are considered jointly, based on the attributes of teachers commonly observed, the overall effect of teacher attributes is much greater than half a standard deviation of universal score gains. We judge these effects to be important from a policy perspective, and believe they should inform the recruitment of teachers.

We have much to learn about how to employ VAMs to help improve education policy and student achievement. Much of this modeling is at a formative stage in terms of conceptual models of student learning and statistical models that account for this complex learning process and the vagaries of observational data in education.

Figure 1
 Estimated Total Measurement Error Variance and Average
 Variance of Measurement Error Associated with Test Instrument (IRT Analysis)
 Grades 4-8 and Years 1999-2007

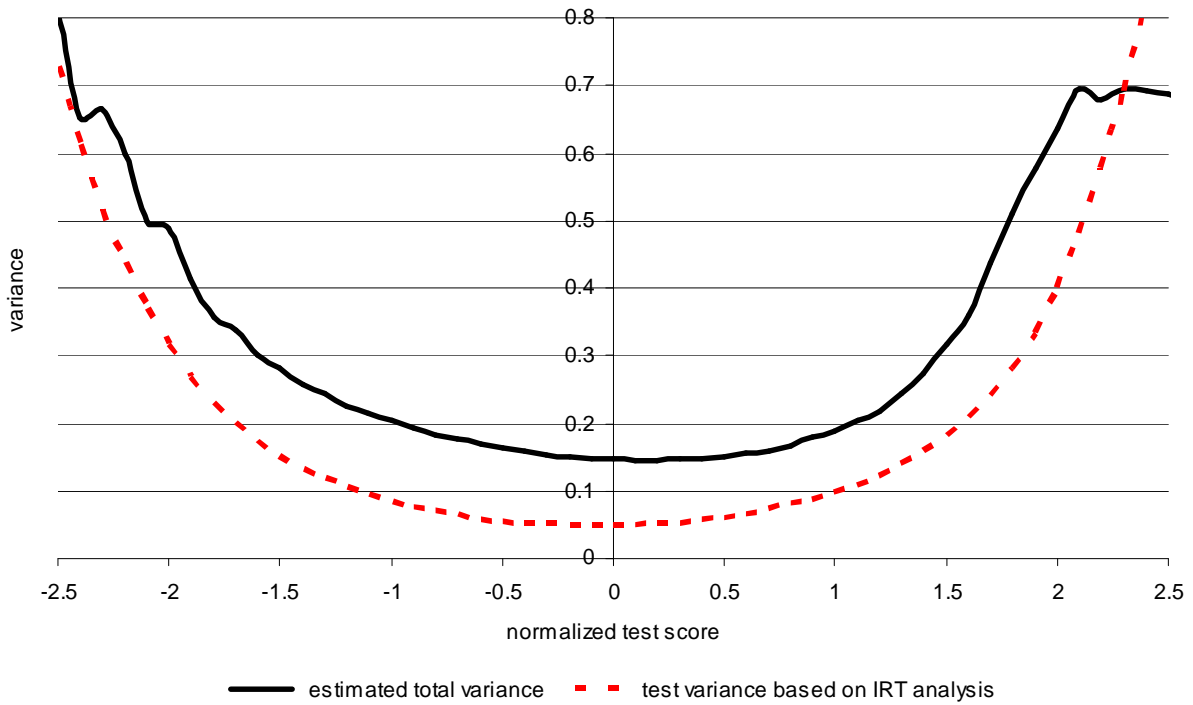


Figure 2
 Distributions of Grade Five Test Scores by Whether Records Include Scores for Grade Six

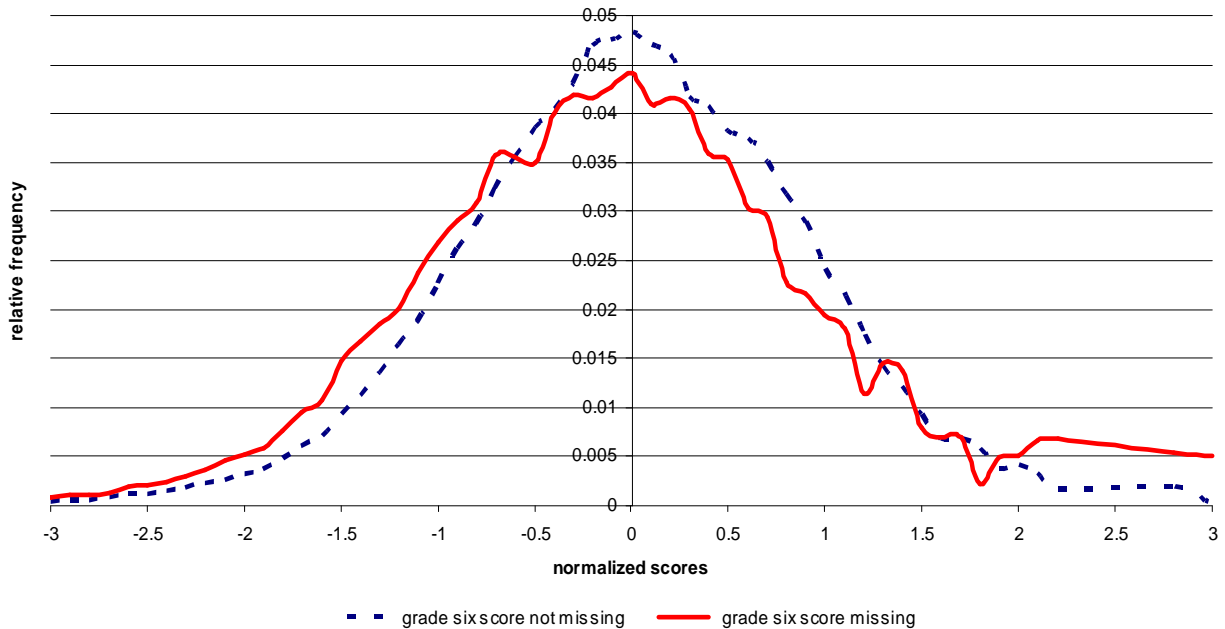


Figure 3
Distributions of Gain Scores, the Empirical Bayes Estimates of Universal Gain Scores
and the Estimated Empirical Distribution of Universal Gain Scores, Grade 5

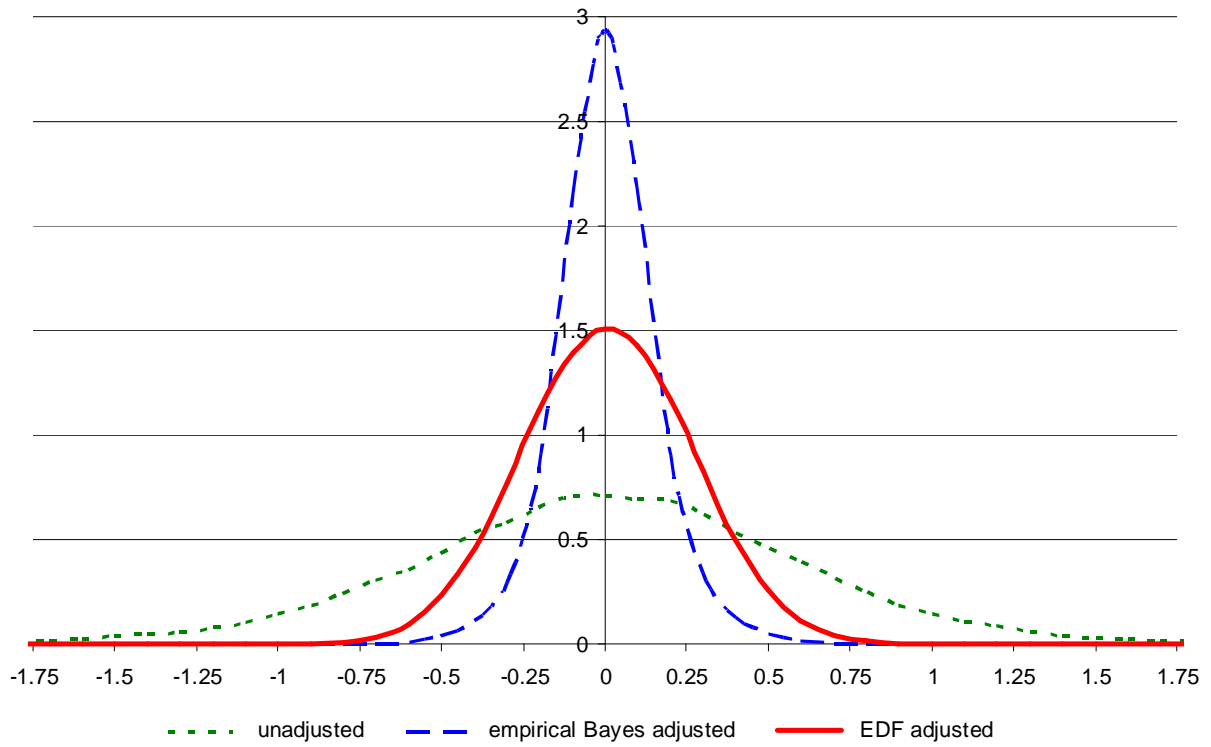


Table 1 Auto-Covariance Matrix of Test Scores, $\tilde{\Sigma}$.

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Grade 3	1.0000	0.7598	0.7199	0.6940	0.6869	0.6432
Grade 4	0.7598	1.004	0.7975	0.7675	0.7574	0.7189
Grade 5	0.7198	0.7975	0.9933	0.7813	0.7639	0.7218
Grade 6	0.6940	0.7675	0.7813	0.9899	0.7958	0.7579
Grade 7	0.6869	0.7574	0.7639	0.7958	0.9820	0.7884
Grade 8	0.6432	0.7189	0.7218	0.7579	0.7884	0.9826

Table 2
Auto-Covariance Estimates
Assuming Stationarity

parameters	estimates	S.D.
$\hat{\omega}^0$	0.9924	0.0022
$\hat{\omega}^1$	0.7907	0.0018
$\hat{\omega}^2$	0.7631	0.0018
$\hat{\omega}^3$	0.7396	0.0018
$\hat{\omega}^4$	0.7189	0.0017

	Model 1	Model 1a	Model 1b	Model 2
σ_{η}^2	0.1699 (0.044)	0.1775 (0.026)	0.1795 (0.025)	0.1650
γ^0	0.8225 (0.058)	0.8149 (0.038)	0.8129 (0.038)	0.8274
β	0.8647 (0.432)	0.9687 (0.008)		1.0092
λ or ψ	0.0795 (0.330)		-0.0239 (0.006)	-0.0443
α				0.3410
Q	4.059E-08	7.344E-06	1.202E-05	2.674E-9

	Model 1	Model 1a	Model 1b	Model 2
Variance in scores for a particular grade	0.9924	0.9924	0.9924	0.9924
Variance in universal scores for a particular grade	0.8225	0.8149	0.8129	0.8274
Variance in gain scores	0.3980	0.3980	0.3980	0.3980
Variance of the gain in universal scores	0.0582	0.0430	0.0390	0.0680
Standard deviation of universal score gain	0.2412	0.2074	0.1975	0.2608

Table 5: Base Model for Math Grades 4 & 5 with Student Fixed Effects, 2000-2005

Constant	0.17147 [1.51]	SD ELA score t-1	-0.02332 [1.91]	14	0.1263 [8.21]**	Not certified	-0.04235 [5.72]**
Student changed schools	-0.03712 [6.60]**	SD math score t-1	-0.11722 [8.27]**	15	0.1252 [6.82]**	Barrons undergrad college Most competitive	0.01498 [1.48]
Class Variables		Teacher Variables		16	0.12464 [6.36]**	Competitive	0.01426 [2.24]*
Proportion Hispanic	-0.4576 [12.89]**	Experience	0.06549 [10.61]**	17	0.08298 [3.10]**	Least Competitive	0.00686 [1.25]
Proportion Black	-0.57974 [16.16]**	3	0.1105 [16.56]**	18	0.14161 [4.02]**	Imputed Math SAT	0.00043 [9.05]**
Proportion Asian	-0.07711 [1.75]	4	0.13408 [17.91]**	19	0.13686 [2.62]**	Imputed Verbal SAT	-0.00034 [6.06]**
Proportion other	-0.56887 [3.95]**	5	0.117 [14.24]**	20	0.24658 [2.50]*	SAT missing	-0.01535 [2.94]**
Class size	0.002 [3.36]**	6	0.13365 [14.58]**	21 or more	0.38977 [3.89]**	Initial path into teaching College Recommended	0.03108 [4.95]**
Proportion Eng Lang Learn	-0.42941 [14.16]**	7	0.12307 [12.27]**	Cert pass first	0.00657 [0.94]	NYC Teaching Fellows	0.01173 [1.10]
Proportion home lang Eng	-0.02902 [1.16]	8	0.11898 [10.81]**	Imputed LAST score	0.00025 [0.57]	Teach for America	0.02364 [1.20]
Proportion free lunch	-0.00181 [0.01]	9	0.12433 [10.04]**	LAST missing	0.00188 [0.26]	Individual evaluation	0.00866 [1.00]
Proportion reduced lunch	0.10521 [3.40]**	10	0.13693 [9.85]**	Certified Math	0.07086 [1.30]	Other	-0.00138 [-0.09]
Mean absences t-1	-0.01367 [15.10]**	11	0.12592 [9.41]**	Certified Science	-0.04852 [0.95]	Teacher LAST* class proportion free lunch	-0.00024 [0.49]
Mean suspensions t-1	0.14069 [2.78]**	12	0.10209 [7.66]**	Certified special ed	0.01086 [1.05]	Observations	578,630
Mean ELA score t-1	0.33811 [31.29]**	13	0.11831 [8.23]**	Certified other	-0.00521 [0.62]		
Mean math score t-1	-0.88479 [58.78]**						

Table 6
 Estimated Effect Sizes for Teacher Attributes Model for
 Math Grades 4 & 5, NYC 2000-2005

	Effect Sizes: Estimated effects relative to		
	S.D. of observed score	S.D. of observed gain score	S.D. of universal score gain
First year of experience	0.065**	0.103	0.253
Not certified	-0.042**	-0.067	-0.162
Attended competitive college	0.014*	0.022	0.054
One S.D. increase in math SAT score	0.041**	0.065	0.158
All observable attributes of teachers	0.162	0.256	0.631

** 1% statistical significance * 5% statistical significance.

References

- Abowd, J.M. and D. Card (1989) "On the Covariance Structure of Earnings and Hours Changes," *Econometrica* 57(2), 411-445.
- Aaronson, D., L. Barrow and W. Sander (2003) "Does Teacher Testing Raise Teacher Quality? Evidence from State Certification Measurements," Working Paper. Research Department Federal Reserve Bank of Chicago.
- Ballou, D. (2002) "Sizing Up Test Scores," *Education Next* 2(2), 10-15.
- Boyd, D., H. Lankford, S. Loeb, J. Rockoff and J. Wyckoff (2007) "The Narrowing Gap in New York City Teacher Qualifications and its Implications for Student Achievement in High-Poverty Schools," working paper.
- Boyd, D., P. Grossman, H. Lankford, S. Loeb and J. Wyckoff (2007) "Who Leaves? Teacher Attrition and Student Achievement," working paper.
- Brennan, R. L. (2001) *Generalizability Theory*, New York: Springer-Verlag.
- Cameron, A.C. and P.K. Trivedi (2005) *Microeconometrics: Methods and Applications*, New York: Cambridge University Press.
- Carlin, B. P. and T. A. Louis (1996) *Bayes and Empirical Bayes Methods for Data Analysis*, Boca Raton: Chapman & Hall/CRC.
- Clotfelter, C., H. Ladd, and J. Vigdor (2006) "Teacher-Student Matching and the Assessment of Teacher Effectiveness" *Journal of Human Resources* 41(4): 778–820.
- Clotfelter, C., H. Ladd, and J. Vigdor (2007) "How and Why do Teacher Credentials Matter for Student Achievement?" CALDER working paper.
- Cohen, P. C. () "Small Area Estimation for the Distribution of Parameters with Covariates," working paper.
- Conbach, Linn, Brennan and Haertel (1997)
- CTB/McGraw-Hill (2006) "New York State Testing Program 2006: Mathematics, Grades 3-8: Technical Report", Monterey, CA.
- CTB/McGraw-Hill (2007) "New York State Testing Program 2007: Mathematics, Grades 3-8: Technical Report", Monterey, CA.
- Feldt, L. S. and R. L. Brennan (1989) "Reliability," in *Educational Measurement* 3rd ed., New York: American Council on Education.
- Goldhaber, 2007
- Goldhaber, D. and E. Anthony (2007) "Can Teacher Quality Be Effectively Assessed? National Board Certification as Signal of Effective Teaching," *The Review of Economics and Statistics* 89(1), 134-150.

- Haertel, E. H. (2006) "Reliability," in Educational Measurement, fourth edition, R. L. Brennan, ed., Praeger.
- Harris, D. and T. Sass (2007) "The Effects of NBPTS-Certified Teachers on Student Achievement," working paper.
- Hill, C., H. Bloom, A. Black, M. Lipsey, "Empirical Benchmarks for Interpreting Effect Sizes in Research" MDRC Working Paper, July 2007.
- Jacob, B. A. and L. Lefgren (2005) "Principals as Agents: Subjective Performance Measurement in Education," working paper.
- Kane, T., J. Rockoff and D. Staiger (in press) "What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City" *Economics of Education Review*.
- Lee, W. C., R. L. Brennan and M. J. Kolen (2000) "Estimators of Conditional Scale-Score Standard Errors of Measurement: A Simulation Study," *Journal of Educational Measurement* 37(1), 1-20.
- McCaffrey, D. F., J. R. Lockwood, D. Koretz, T. A. Louis and L Hamilton (2004) "Models for Value-Added Modeling of Teacher Effects," *Journal of Educational and Behavioral Statistics* 29(1), 67-101.
- Nye, B., S. Konstantopoulos, L. Hedges, (2004) "How Large Are Teacher Effects?" *Educational Evaluation and Policy Analysis*, 26(3), 237-257.
- Rivkin, S., E. Hanushek and J. Kain (2005) "Teachers, Schools, and Academic Achievement," *Econometrica* 73(2), 417-458.
- Rockoff, J. (2004) "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," *American Economic Review* 94(2), 247-252.
- Rogosa, D.R.
- Rothstein, J. (2007) "Do Value-Added Models Add Value? Tracking, Fixed Effects, and Causal Inference," working paper.
- Sanders, W. and J. Rivers (1996) "Cumulative and Residual Effects of Teachers on Future Student Academic Achievement," working paper, University of Tennessee Value-Added Research and Assessment Center.
- Sanford, E. E. (1996) "North Carolina End-of-Grade Tests: Reading Comprehension, Mathematics," Technical Report #1. Division of Accountability/Testing, Office of Instruction and Accountability Services, North Carolina Department of Public Instruction.
- Shen, W. and T. A. Louis (1998) "Triple-goal Estimates in Two-Stage Hierarchical Models," *Journal of the Royal Statistical Society* 60(2), 455-471.
- Sullivan, D. G. (2001) "A Note on the Estimation of Linear Regression Models with Heteroskedastic Measurement Errors," Federal Reserve Bank of Chicago working paper WP 2001-23.

Thorndike, R. L. (1951) "Reliability," in Educational Measurement, E.F. Lindquist, ed., Washington, D.C.: American Council on Education.