CALDER

Urban Institute

NATIONAL
CENTER *for* ANALYSIS *of* LONGITUDINAL DATA *in* EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

*A program of research by the Urban Institute with Duke University, Stanford University, University of Florida, University of Missouri-Columbia, University of Texas at Dallas, and University of Washington*

# *Measuring Effect Sizes*

## The Effect of Measurement Error

DONALD BOYD, PAMELA GROSSMAN, HAMILTON LANKFORD, SUSANNA LOEB, AND JAMES WYCKOFF

WORKING PAPER 19 • JUNE 2008

# Measuring Effect Sizes: the Effect of Measurement Error

**Donald Boyd*, Pamela Grossman**, Hamilton Lankford*, Susanna Loeb**, and James Wyckoff*****

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

We gratefully acknowledge support from the National Science Foundation and the National Center for the Analysis of Longitudinal Data in Education Research (CALDER), supported by Grant R305A060018 to the Urban Institute from the Institute of Education Sciences, U.S. Department of Education. Thanks also to Brian Jacob, Dale Ballou and participants of the National Conference on Value-Added Modeling for their helpful comments. The authors are solely responsible for the content of this paper and the views expressed in the paper may not reflect those of the Urban Institute, the Institute of Education Sciences, the U.S. Department of Education, or any other sponsors of the research.

CALDER working papers have not gone through final formal review and should be cited as working papers. They are intended to encourage discussion and suggestions for revision before final publication.

Measuring Effect Sizes: the Effect of Measurement Error
Donald Boyd, Pamela Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff
CALDER Working Paper No. 19

## Abstract

Value-added models in education research allow researchers to explore how a wide variety of policies and measured school inputs affect the academic performance of students. Researchers typically quantify the impacts of such interventions in terms of *effect sizes*, i.e., 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 based on administrative databases typically are quite small.

Research has shown that high quality teachers have large effects on student learning but that measures of teacher qualifications seem to matter little, leading some observers to conclude that, even though effectively choosing teachers can make an important difference in student outcomes, attempting to differentiate teacher candidates based on pre-employment credentials is of little value. This illustrates how the perception that many educational interventions have small effect sizes, as traditionally measured, are having important consequences for policy.

In this paper we focus on two issues pertaining to how effect sizes are measured. 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. 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 true achievement or the gain in true achievement that is of interest.

Adjusting effect-size estimates 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. However, there are a number of other factors, including variation in scores associated with students having particularly good or bad days, which can result in test scores not accurately reflecting true academic achievement. Using the covariance structure of student test scores across grades in New York City from 1999 to 2007, we estimate the overall extent of test measurement error and how measurement error varies across students. Our estimation strategy follows from two key assumptions: (1) there is no persistence (correlation) in each student's test measurement error across grades; (2) there is at least some persistence in learning across grades with the degree of persistence constant across grades. Employing the covariance structure of test scores for NYC students and alternative models characterizing the growth in academic achievement, we find estimates of the overall extent of test measurement error to be quite robust.

Returning to the analysis of effect sizes, our effect-size estimates based on the dispersion in gain scores net of test measurement error are four times larger than effect sizes typically measured. To illustrate the importance of this difference, we consider results from a recent paper analyzing how various attributes of teachers affect the test-score gains of their students (Boyd et al., in press). Many of the estimated effects appear small when compared to the standard deviation of student achievement – that is effect sizes of less than 0.05. However, when measurement error is taken into account, the associated effect sizes often are about 0.16. Furthermore, when teacher attributes are considered jointly, based on the teacher attribute combinations commonly observed, the overall effect of teacher attributes is roughly half a standard deviation of universe score gains – even larger when teaching experience is also allowed to vary. The bottom line is that there are important differences in teacher effectiveness that are systematically related to observed teacher attributes. Such effects are important from a policy perspective, and should be taken into account in the formulation and implementation of personnel policies.

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 researchers 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? Judging whether a change in student achievement is important requires some meaningful point of reference. In certain cases a grade equivalence scale or some other intuitive and policy relevant metric of educational achievement can be used. However, this is not the case with item response theory (IRT) scale-score measures common to the tests usually employed in value-added analyses. In such cases, researchers typically describe the impacts of various interventions in terms of *effect sizes*, although conveying the intuition of such a measure to policymakers often is a challenge.

The *effect size* of an independent variable is 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 typically are 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., in press); 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., in press; Clotfelter et al., 2007). In most studies the effect size of any single individual teacher attribute is smaller than the first-year experience effect.

Most researchers 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 high quality teachers have large effects on student learning so that effectively choosing 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). The findings that teachers greatly influence student outcomes but that measures of teacher qualifications seem to matter little, taken together have led some observers to conclude that attempting to differentiate teachers on their pre-employment credentials is of little value. Rather, they argue, education policymakers would be better served by reducing educational and credential barriers to enter teaching in favor of more rigorous performance-based evaluations of teachers.[1] Indeed, this perspective appears to be gaining some momentum. Thus, the perception that many educational interventions have small effect sizes, as traditionally measured, are having important consequences for policy.

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 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[2]; and multicolinearity resulting from the similarity of many of the commonly employed teacher attributes. 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 additional issues pertaining to how effect sizes are measured, which we believe are especially important.

First, we argue that estimated 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. Netting out measurement error is especially important in this context. Because gain scores have measurement error in pre-tests and post-tests, the measurement error in gains is even greater than that in levels. The noise-to-signal ratio is also larger as a result of the gain in actual achievement being smaller than the level of achievement.

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

[1] See, for example, R. Gordon, T. Kane and D. Staiger (2006).

[2] 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.

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, including variation in scores resulting from 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.

Measuring effect sizes relative to the dispersion in gain scores net of test measurement error will result in all the estimated effect sizes being larger by the same multiplicative factor, so that the relative sizes of effects will not change. Such relative comparisons are important in cost-effectiveness comparisons where the effect of one intervention is judged relative to some other. However, as noted above, the absolute magnitudes of effect sizes for measurable attributes of teachers are relevant in the formulation of optimal personnel (e.g., hiring) policies. More generally, the absolute magnitudes of effect sizes are relevant in cost-benefit analyses and when making comparisons across different outcome measures (e.g., different tests).

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 (in press) 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.[3] Generalizability theory, which we draw upon here, extends test theory to explicitly account for multiple sources of measurement error.[4]

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 (e.g., questions) with the student doing that task at some point in time. The universe of possible occurrences is such that the student's knowledge, skills, and 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, we assume 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 " $\times$ " is read "crossed with".

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

$$S_{ito} = \tau + \upsilon_i + \upsilon_t + \upsilon_o + \upsilon_{it} + \upsilon_{io} + \upsilon_{to} + \varepsilon_{ito} \quad (1)$$

$\tau_i \equiv \tau + \upsilon_i$ , the *universe score* for the student, equals the expected value of $S_{ito}$ over the universe of generalization, here the universes of possible tasks and occurrences. The universe score is comparable to the true score as defined in classical test theory. In our case, $\tau_i$ measures the student's underlying academic achievement, e.g., ability, knowledge and skills. The $\upsilon$ 's represent a set of uncorrelated random effects which, along with $\varepsilon_{ito}$ and the student's universe score, sum to $S_{ito}$ . Here $\upsilon_t$ ( $\upsilon_o$ ) reflect the random effect, common to all test-takers, associated with scores for a particular task (occurrence) differing from the population mean, $\tau$ . $\upsilon_{it}$ reflects the fact that a student might do especially well or poorly on a particular task. $\upsilon_{io}$ 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). $\upsilon_{to}$ reflects the possibility that the performance of all students on a particular task might vary across occurrences. $\varepsilon_{ito}$ 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.[5]

[3] Classical test theory is the focus of many books and articles. For example, see Haertel (2006).
[4] See Brennan (2001) for a detailed development of Generalizability Theory. The basic structure of the framework is outlined in Cronbach, Linn, Brennan and Haertel (1997) as well as Feldt and Brennan (1988).
[5] Thorndike (1951, p. 568) provides a taxonomy characterizing different sources of measurement error. The above framework also can be generalized to reflect students being grouped within schools and there being common random components of measurement error at that level.

The observed score for a particular individual completing a task will differ from the individual's universe 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).[6]

$$\eta_{ito} \equiv \left(S_{ito} - \tau_i\right) = \upsilon_t + \upsilon_o + \upsilon_{it} + \upsilon_{po} + \upsilon_{to} + \varepsilon_{ito} \quad (2)$$

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

Now consider a test (T) 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 $\eta_{iT}$ is a composite measure reflecting the errors in test measurement from all sources.[7]

$$S_{iT} = \sum\nolimits_t S_{it} \Big/ N_T = \tau + \upsilon_i + \upsilon_o + \upsilon_{io} + \sum\nolimits_t \left(\upsilon_t + \upsilon_{it} + \upsilon_{to} + \varepsilon_{ito}\right) \Big/ N_T = \tau_i + \eta_{iT}. \quad (4)$$

The variance of $\eta_{iT}$ for student i equals $\sigma^2_{\eta_{iT}} = \sigma^2\left(o\right) + \sigma^2\left(io\right) + \left[\sigma^2\left(t\right) + \sigma^2\left(it\right) + \sigma^2\left(to\right) + \sigma^2\left(\varepsilon_{ito}\right)\right] / N_T$.

Equation (5) generalizes the notation in (4) to allow for tests in multiple grades.

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

$S_{i,g}$ is the $i^{\text{th}}$ student's score on a test for a particular subject taken in grade g. $\tau_{i,g}$ is the $i^{\text{th}}$ 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, $E\eta^2_{i,g} = \sigma^2_{\eta_{i,g}}$ To simplify the analysis, we maintain that the measurement error variance for each student is constant across grades; $\sigma^2_{\eta_{i,g}} = \sigma^2_{\eta_i}, \forall g$. Let $\sigma^2_{\eta_\bullet}$ equal $\sigma^2_{\eta_i}$ for all pupils in the homoskedastic case or, more generally, the mean value of $\sigma^2_{\eta_i}$ in the population of students. The $\upsilon$ 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, $\sigma^2_{S_g}$, into the variance in universe scores across the student population, $\sigma^2_{\tau_g}$, and the measurement-error variance, $\sigma^2_{\eta_\bullet}$; $\sigma^2_{S_g} = \sigma^2_{\tau_g} + \sigma^2_{\eta_\bullet}$. Here $K_g = \sigma^2_{\tau_g} \big/ \sigma^2_{S_g}$ is the *generalizability coefficient* measuring the portion of the total variation in observed scores that is explained by the variance

[6] By construction, $\varepsilon_{ito}$ and the $\upsilon$ are independent.

[7] Here we represent the score as the mean over the set of test items. An alternative would be to employ $S_{iT} = \sum\nolimits_t S_{it}$, e.g., the number of correct items.

of universe scores. The reliability coefficient is the comparable measure in classical test theory. As discussed below, we standardize test scores to have zero means and unit standard deviations; $\sigma^2_{S_g} = 1 = \sigma^2_{\tau_g} + \sigma^2_{\eta\bullet}$ . In this case, the generalizability coefficient equals $K_g = 1 - \sigma^2_{\eta\bullet}$ .

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.[8]

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, $\sigma_{S_g}$ , or the standard deviation of observed gain scores. From the perspective that the estimated effects shed light 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 universe scores or the standard deviation of gains in the universe score. In most cases, it is the latter that is pertinent.

At a point in time, a student’s universe 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 universe score at a point in time reflects the causal linkages between all such factors and the dispersion in these varied and long-run factors across students. 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 universe 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.[9] 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

[8] Measurement error in lagged test scores entering as right-hand-side controls in regression models is discussed below.

[9] This might help explain the result noted in footnote 2; standardized tests often measure cumulative learning whereas tests designed for a specific topic may measure the growth in learning targeted by a particular intervention.

growth, for example to comply with NCLB in a growth context, then the relevant metric is the standard deviation in the gain in universe 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 universe 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 measured test score gain and the gain in universe scores reflecting the underlying achievement growth. 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

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

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 $\sigma^2_{\Delta\eta_{i,g}} = 2\sigma^2_{\eta_i}$ 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; $\sigma^2_{\Delta S_g} = \sigma^2_{\Delta\tau_g} + \sigma^2_{\Delta\eta_\bullet}$ where $\sigma^2_{\Delta\tau_g}$ is the variance of the universe score growth in the population of students and $\sigma^2_{\Delta\eta_\bullet}$ is the mean value of $\sigma^2_{\Delta\eta_i}$. Here the generalizability coefficient $K^\Delta_g = \sigma^2_{\Delta\tau_g} \big/ \sigma^2_{\Delta S_g}$ is the proportion of the overall variance in gain scores that actually reflects variation in students' underlying growth in educational achievement. In general, $K^\Delta_g$ will be smaller than $K_g = \sigma^2_{\tau_g} \big/ \sigma^2_{S_g}$ so that test measurement error is especially problematic when analyzing achievement growth.[10]

**An Empirical Example: 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

[10] This point has been made in numerous publications. See, for example, Ballou (2002). Rogosa and Willett (1983) discuss 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.

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 3, 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 3 cut score for each grade.[11] 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.[12] 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 0.044; the average absolute differences from the mean standard deviation is less than five percent of the mean. Given these properties, we standardize the test scores by grade and year, with little, if any, loss in useful information.[13]

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 tests report reliability coefficients that range from 0.88 to 0.95 and average 0.92, indicating that eight percent of the variation in the scores for a test reflect measurement error associated with the test instrument. 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. This can result from differences in the measurement error variance, possibly due to differences in testing conditions, or the dispersion in the underlying achievements of students in New York City differing from that statewide. More importantly, the reliability measures are with respect to raw scores, not the scale scores typically employed in VA analyses. As a result of the nonlinear mapping between raw and scale scores, a given raw-score increase yields quite different increases in scale 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

[11] CTB/McGraw-Hill (2006).

[12] CTB/McGraw-Hill (2007).

[13] Rothstein (2007, p. 12) makes the point that when scores are measured on an interval scale, standardizing 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 standardizing scores are of little, if any, consequence.

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 scale scores.

The technical reports provide estimates of the standard errors of measurement (SEM) for the scale scores. These estimates have a conceptual foundation that differs from classical test theory because they are based on an IRT framework. Even so, the reported SEM 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.[14] The technical documents for New York report IRT standard errors of measurement for every scale-score value. Reflecting our standardizations of scale-scores discussed above, we standardize 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 NYC 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 the variance of the universe score and the universe score change, $\sigma_{\tau}^2 = \sigma_S^2 - \sigma_{\eta_\bullet}^2$ and $\sigma_{\Delta\tau}^2 = \sigma_{\Delta S}^2 - \sigma_{\Delta\eta_\bullet}^2 = \sigma_{\Delta S}^2 - 2\sigma_{\eta_\bullet}^2$. By construction, $\sigma_S^2 = 1$, and we estimate $\hat{\sigma}_{\Delta S}^2 = 0.398$ in the New York City data. With 0.102 being a lower-bound estimate of $\sigma_{\eta_*}^2$, 0.898 and 0.194 are upper-bound estimates of $\sigma_{\tau}^2$ and $\sigma_{\Delta\tau}^2$, respectively. Thus, effect sizes measured in relation to $\sigma_{\Delta\tau}$ are more than twice as large as effect sizes measured in relation to $\sigma_S^2$. (Our estimate of $\sigma_{\Delta\tau}$ is $0.439 = \sqrt{0.192}$.) By contrast, $\sigma_S$ is 1.0, which is 2.28 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, $\sigma_{\eta_\bullet}^2$. As noted in footnote 5, 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 (1) the selection of test items included in a test, (2) the timing (occurrence) of the test and (3) these factors *crossed with* students. Reliability or generalizability 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 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,

[14] Lee, Brennan and Kolen (2000).

forgetfulness, carelessness, subjectivity or impulsiveness in response and luck in random guessing"[15] 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 state achievement tests like those considered here.[16]

Rather than analyzing the consistency of student test scores over occurrences, 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 (1989) note, 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 $\sigma_{\eta_\bullet}^2$ 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.

**Analyzing the Overall Measurement-Error Variance**.

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

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

where $\Gamma$ is the auto-covariance matrix for the universe scores and $\mathrm{I}$ is a $6 \times 6$ identity matrix. For the population of all students, $\Omega_\bullet = E\,\Omega(i) = \Gamma + \sigma_{\eta_\bullet}^2 \mathrm{I}$ 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; $\Gamma$ and, therefore, the off diagonal elements of $\Omega(i)$ are assumed to be constant across students.[17]

[15] Feldt and Brennan (1989).

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

[17] To simplify notation we have assumed that $\sigma_{\eta_{i,g}}^2 = \sigma_{\eta_i}^2, \forall g$. However, this is not needed for much of our analysis. Taking expectations across all students, it is sufficient that $E\sigma_{\eta_{i,g}}^2 = E\sigma_{\eta_{i,g'}}^2 = \sigma_{\eta_\bullet}^2, \forall g, g'$.

We employ test-score data from New York City to estimate the empirical counterpart of $\Omega_{\bullet}$, $\tilde{\Omega}_{\bullet}=\sum_i S_i S_i' \big/ N$. 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 relatively minor 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 $g+1$, $g=3,4,...7$. 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.[18] 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 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.[19]

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), with the coefficient of

[18] For example, see Rubin (1987) and Schafer (1997).

[19] The Markov Chain Monte Carlo procedure was used to impute missing-score gaps (e.g., a missing fourth grade score for a student having scores for grades three and five). This yielded an imputed database with only *monotone* missing data (e.g., scores included for grades three through five and missing in all grades thereafter). The monotone missing data were then imputed using the parametric regression method.

dispersion being quite small (0.007). 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 $\Omega_{\bullet}$. In particular, let $\omega^s \equiv Cov(S_{i,g}, S_{i,g+s})$, $s = 1, 2, ..., 4$, starting with grade four.[20] Estimates of these measures are shown in Table 2, along with the estimate of $\omega^0 = V(S_{i,g}) = \gamma^0 + \sigma^2_{\eta_{\bullet}}$.

In the following section, we describe the approach used to estimate $\sigma^2_{\eta_{\bullet}}$, $\gamma^0$, and $\gamma^1$ which yields an estimate of the variance in the gain in universe scores; $\sigma^2_{\Delta\tau} = V(\tau_{i,g+1} - \tau_{i,g}) = 2(\gamma^0 - \gamma^1) = 2(\gamma^0 - \omega^1)$. Alternatively, $\sigma^2_{\Delta\tau} = \sigma^2_{\Delta S} - 2\sigma^2_{\eta_{\bullet}}$. 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. The approach, developed by Abowd and Card (1989), has been applied and extended in numerous papers.

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

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

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 $\beta < 1$, where the rate of decay, $1 - \beta$, is assumed to be constant across grades. Repeated substitution yields $\tau_{i,g} = \beta^g \tau_{i0} + \sum_{s=1}^{g} \beta^{g-s}\theta_{i,s}$ where $\tau_{i0}$ is the initial condition. In the special case where $\beta = 1$, $\theta_{i,g}$ is the student's gain in achievement while in grade g.[21] This special case is the basic structure maintained in many value-added analyses, including the layered model employed by Sanders.[22] 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., $\theta_{i,1}, \theta_{i,2}, ...$) together determine the dynamic pattern of the universe scores as reflected in the parameterization of $\Gamma = E(\tau_i \tau_i^{'})$ which, given stationarity, is completely characterized by $\gamma^0, \gamma^1, \cdots, \gamma^4$ 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 $\Gamma$, several general implications of

[20] We hypothesize that the patterns for third-grade scores differ because this is the first tested grade, resulting in relatively greater test measurement error due, at least in part, from confusion about test instructions, testing strategies, etc..

[21] 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; $\theta_{i,g} = S_{i,g+1} - S_{i,g} + (1-\beta)\tau_{i,g}$.

[22] Wright (2007).

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.[23] 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 $\Omega_\bullet$ shown in (9)

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

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

**Model 1:** Consider an individual-effects specification for the $\theta_{i,g}$; $\theta_{i,g} = \mu_i + \varepsilon_{i,g}$ where $\mu_i$ is a 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, $E\varepsilon_{i,g}\varepsilon_{i,g'} = 0\ \forall g \neq g'$. This structure implies that $\psi^s = E\tau_{i,g}\,\theta_{i,g+s} = E\tau_{i,g}\left(\mu_i + \varepsilon_{i,g+s}\right) = E\tau_{i,g}\,\mu_i \equiv \lambda$ for all s > 0 as well as the test-score auto-covariances shown in (10).

[23] The expression $\gamma^1 = E\tau_{i,g}\tau_{i,g+1} = E\tau_{i,g}\left(\beta\tau_{i,g} + \theta_{i.g+1}\right) = \beta\gamma^0 + E\tau_{i,g}\theta_{i,g+1}$ implies that $E\tau_{i,g}\theta_{i,g+1} = \gamma^1 - \beta\gamma^0$.

$$\begin{aligned}
\omega^0 &= \gamma^0 + \sigma^2_{\eta\bullet} \\
\omega^1 &= \beta\gamma^0 + \lambda \\
\omega^2 &= \beta^2\gamma^0 + (\beta+1)\lambda \\
\omega^3 &= \beta^3\gamma^0 + \left(\beta^2+\beta+1\right)\lambda \\
\omega^4 &= \beta^4\gamma^0 + \left(\beta^3+\beta^2+\beta+1\right)\lambda
\end{aligned} \qquad (10)$$

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 each student 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 (i.e., $\beta = 1$), the test-score covariances are of the form $\omega^s = \gamma^0 + s\lambda$.

**Model 2:** To explore whether estimates of $\sigma^2_{\eta\bullet}$ are robust to maintaining a more general model structure, we can specify a reduced-form parameterization of the $\psi^s = E\tau_{i,g}\theta_{i,g+s}$ in (9), rather than specify the structure of $\theta_{i,g}$ and infer the structures of $\psi^s$ and $\omega^0, \omega^1, \ldots$ . In particular, consider the case $\psi^s = \alpha^{s-1}\psi$ where it is anticipated that $\alpha \le 1$. The implied test-score covariance structure is shown in (11), which includes Model 1 as the special case where $\alpha = 1$. For $\alpha < 1$ and $\psi > 0$, the specification in (11) corresponds to the case where student gains follow an AR(1) process; $\theta_{i,g} = \alpha\theta_{i,g-1} + \varepsilon_{i,g}$ where $\varepsilon_{i,g}$ is *i.i.d.* as above and $E\theta_{i,g}\,\varepsilon_{i,g+s} = 0,\ s>0$. However, Model 2 also allows for the possibility that $\psi = E\tau_{i,g}\theta_{i,g+1} < 0$.

$$\begin{aligned}
\omega^0 &= \gamma^0 + \sigma^2_{\eta\bullet} \\
\omega^1 &= \beta\gamma^0 + \psi \\
\omega^2 &= \beta^2\gamma^0 + (\beta+\alpha)\psi \\
\omega^3 &= \beta^3\gamma^0 + \left(\beta^2+\beta\alpha+\alpha^2\right)\psi \\
\omega^4 &= \beta^4\gamma^0 + \left(\beta^3+\beta^2\alpha+\beta\alpha^2+\alpha^3\right)\psi
\end{aligned} \qquad (11)$$

Let $\chi$ 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, $\chi \equiv \left[\sigma^2_{\eta\bullet}\ \gamma^0\ \beta\ \lambda\right]$ in (10) for Model 1. Let $\hat{\omega} \equiv \left[\hat{\omega}^0\quad \hat{\omega}^1\quad \hat{\omega}^2\quad \hat{\omega}^3\quad \hat{\omega}^4\right]$ represent the empirical counterpart of the unique elements of the auto-covariance matrix $\Omega_\bullet$, i.e., $\tilde{\Omega}_\bullet$, shown in Table 2. The parameters in $\chi$ can be estimated using a minimum distance estimator where $\hat{\chi}$ is the value of $\chi$ that minimizes the distance between $\omega(\chi)$, and $\hat{\omega}$ as measured by $Q = (\hat{\omega} - \omega(\chi))(\hat{\omega} - \omega(\chi))' = \sum_j \left(\hat{\omega}^j - \omega^j(\chi)\right)^2$. This *equally weighted minimum distance estimator* is commonly used in empirical analyses where parameters characterizing covariance structures are estimated (e.g., the auto-covariance structure of earnings).[24]

It is the over-identification of parameters in Model 1 that leads us to estimate the parameters by minimizing $Q$. With Model 2 having five equations (i.e., $\hat{\omega}^j = \omega^j(\chi),\ j = 0,1,...,4$) in five unknown parameters, we are able to directly solve for estimates of those parameters, as discussed in the Appendix. Dropping the last equation in (10), one also can directly obtain estimates of the parameters in Model 1 in a similar manner. In this case, $\hat{\beta} = (\hat{\omega}^2 - \hat{\omega}^3)/(\hat{\omega}^1 - \hat{\omega}^2)$, $\hat{\lambda} = \hat{\omega}^2 - \hat{\beta}\hat{\omega}^1$, $\hat{\gamma}^0 = (\hat{\omega}^1 - \hat{\lambda})/\hat{\beta}$ and $\hat{\sigma}^2_{\eta_\bullet} = \hat{\omega}^0 - \hat{\gamma}^0$.[25]

Such direct solution illustrates the intuition behind our general approach for estimating the extent of measurement error. The equations characterizing the covariances $\omega^1, \omega^2, \cdots$ allow us to infer an estimate of $\gamma^0$ which, along with the first equation in (10) and (11), yields an estimate of $\sigma^2_{\eta_\bullet}$. This underscores the importance of two key assumptions. First, identification requires the universe test scores to reflect a cumulative process in which there is some degree of persistence (i.e., $\beta > 0$) that is constant across grades. When $\beta = 0$, $\gamma^0$ and $\sigma^2_{\eta_\bullet}$ only enter the first equation, implying that they are not separately identified. Second, there is no persistence (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 a hierarchical linear model (HLM) would be a computational challenge. Instead, we simply compute $\hat{\omega}$ and then need only minimize Q or use the simple formulas applicable when the parameters are exactly identified. 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

[24] 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 earning.
[25] See the Appendix for derivations of these estimators and the estimation formulas for the two special cases of Model 1.

distributions from which the various random components are drawn. Such explicit assumptions are integral to the hierarchical approach. The covariance structure could also be estimated useing panel-data methods that would employ the student-level data, rather than $\hat{\omega}$ which summarizes certain features of that data.[26]

**Results** Parameter estimates for the alternative models discussed above are shown in Table 3. 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 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 two column of Table 3 are for the specification in (11), which includes the other three models as special cases. As discussed in the Appendix, $\beta$, $\alpha$ and $\psi$ in (11) are not uniquely identified in that the system of equations in (11) can be manipulated to show that that $\beta$ and $\alpha$ enter in identical ways (i.e., $\beta$ and $\alpha$ can be exchanged – their interpretation can be switched) so that it is not possible to identify unique estimates of each; as shown in the last two columns of Table 3, we estimates $\beta$ and $\alpha$ to be 0.653 and 0.978, respectively, or these same values in reverse order. Even so, as explained in the appendix, we are able to uniquely identify estimates of $\gamma^0$ and $\sigma^2_{\eta_\bullet}$, with the estimates shown in the last two columns of Table 3. For all the models, standard errors are shown in parentheses.[27]

Note the meaningful difference in the estimates of $\beta$ and $\psi$ across the four sets of estimates. The qualitative differences can be seen to be linked to the stationarity in test-score variances across grades. Given the time-series pattern of test scores maintained in equation (8), it follows that $E\tau^2_{i,g} = E\left(\beta\tau_{i,g-1} + \theta_{i,g}\right)^2 = \beta^2 E\tau^2_{i,g-1} + E\theta^2_{i,g} + 2\beta E\tau_{i,g-1}\theta_{i,g}$. Stationarity of $E\tau^2_{i,g} = E\tau^2_{i,g-1} = \gamma^0$ implies that $\left(1-\beta^2\right)\gamma^0 = \sigma^2_\theta + 2\beta\psi$, which establishes a relationship between $\beta$ and $\psi$. For example, when $\beta = 1$, $\psi = -\sigma^2_\theta \big/ 2 \le 0$; this particular value of $\psi$ is needed in order to maintain the constant test-

[26] For example, see Baltagi (2005, chapter 8) for a general discussion of such dynamic panel data models.

[27] The standard errors reported in Table 4 are the square roots of the diagonal elements of the estimated covariance matrix of $\hat{\chi}$, $V(\hat{\chi}) = \left[D'D\right]^{-1}\left[D'V(\hat{\omega})D\right]\left[D'D\right]^{-1}$. Here $D$ is the first derivative of $\omega(\chi)$ with respect to $\chi$ evaluated at $\hat{\chi}$. Standard deviations for the parameter estimates in model 2 are large because of $D'D$ being close to singular; in this case the determinant $\left|D'D\right|$ equals 2.65E-7. In contrast, $\left|D'D\right|$ equals 8.55E-3, 10.3 and 20.0 for models 1, 1a and 1b, respectively.

score variance. Thus, $\hat{\psi} < 0$ in Model 1b is to be expected and is consistent with the estimate of $\psi$ being negative when the values of $\hat{\beta}$ in Model 2 is close to 1.0.

The variability in the estimates of $\beta$ and $\psi$ across the alternative specifications and the indeterminacy of estimates of $\beta$, $\psi$ and $\alpha$ in Model 2 imply that our relatively simple empirical approach does not allow us to identify the dynamic structure underlying the covariance of the universe scores. However, the estimates of $\sigma^2_{\eta_\bullet}$ shown in the first row of Table 3 are quite robust across the range of specifications, only varying by 0.012, thereby increasing our confidence that approximately 17 percent of the overall dispersion in the NYS tests is attributable to various forms of test measurement error. Furthermore, this robustness supports the proposition that the approach we employ generally can be used to isolate the overall extent of test measurement error.

In the following analysis, we will employ the estimates $\hat{\sigma}^2_{\eta_\bullet} = 0.168$ from Model 2, since this is the more conservative estimates of $\sigma^2_{\eta_\bullet}$. The corresponding estimates of $V\left(\tau_{i,g}\right)$ and $V\left(S_{i,g}\right)$ – that is $\hat{\gamma}^0 =$ 0.824 and $\hat{\omega}^0 = 0.992$ – imply the overall generalizability coefficient is estimated to be $\hat{K}_g = \hat{\sigma}^2_\tau \big/ \hat{\sigma}^2_S = \hat{\gamma}^0 \big/ \hat{\omega}^0 =$ 0.831. 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 Carolina's reading test reports a test-retest reliability equal to 0.86,[28] somewhat larger than our estimate. However, 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' universe-score gains measuring growth in skills and knowledge for the relevant student population. Utilizing Model 2 estimates, we calculate the variance of gain scores net of measurement error to be 0.062: $\hat{\sigma}^2_{\Delta\tau} = \hat{\sigma}^2_{\Delta S} - 2\hat{\sigma}^2_{\eta_*} =$ 0.398 - 2(0.168). Thus, we estimate the standard deviation of universe score gains to be 0.259, indicating that effect sizes based on the dispersion in the gains in actual student achievement are 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

[28] 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.

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 $\hat{\sigma}_{\Delta\tau} = 0.249$ rather than $\hat{\sigma}_{\Delta S} = 0.630$ as the measure of gain score dispersion. This large difference reflects that only one-sixth of the dispersion in gain scores is actually attributable to the dispersion of academic achievement gains.[29]

We have focused on the mean measurement error variance for the population of students, $\sigma^2_{\eta_\bullet}$, 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 (12).

$$V\left(S_{i,g+1} - \beta S_{i,g}\right) = \sigma^2_\theta + \sigma^2_{\eta_i} + \beta^2\sigma^2_{\eta_i} = \sigma^2_\theta + \left(1+\beta^2\right)\sigma^2_{\eta_i} \qquad (12)$$

This, along with the formula $\sigma^2_\theta = \left(1-\beta^2\right)\gamma^0 - 2\beta\psi$ [30] and our estimates of $\sigma^2_{\eta_\bullet}$, $\gamma^0$, $\beta$, and $\psi$, imply the estimator of the measurement error variance for each student, $\hat{\sigma}^2_{\eta_i}$, shown in (13) where $N_i^G$ is the number of grades for which the student has scores.

$$\hat{\sigma}^2_{\eta_i} = \frac{\left[\frac{1}{N_i^G-1}\sum_{g=1}^{N_i^G-1}\left(S_{i,g+1} - \hat{\beta}S_{i,g}\right)^2 - \hat{\sigma}^2_\theta\right]}{\left(1+\hat{\beta}^2\right)} = \frac{\left[\frac{1}{N_i^G-1}\sum_{g=1}^{N_i^G-1}\left(S_{i,g+1} - \hat{\beta}S_{i,g}\right)^2 - \left(\left(1-\hat{\beta}^2\right)\hat{\gamma}^0 - 2\hat{\beta}\hat{\psi}\right)\right]}{\left(1+\hat{\beta}^2\right)} \qquad (13)$$

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

The solid line in Figure 1 shows the estimated relationship between $\hat{\sigma}^2_{\eta_i}$ 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 $\hat{\sigma}^2_{\eta_i}$ for the students whose values of $\bar{S}_i$ fall in each interval. In this way, the solid line is a simple non-parametric characterization of how the overall measurement error

[29] $\hat{\sigma}^2_{\Delta\tau} = 0.062$ implies that the generalizability coefficient for student gain scores, $\hat{K}^\Delta = \left(\hat{\sigma}^2_{\Delta\tau} / \hat{\sigma}^2_{\Delta S}\right) = (0.062/0.398) = 0.156$, is much smaller than that for scores.

[30] This follows from the formula $\left(1-\beta^2\right)\gamma^0 = \sigma^2_\theta + 2\beta\psi$ derived above.

varies across the range of universe 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 venders.

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 universe scores follows a pattern almost identical to that implied by the measurement error variances associated with the test instrument, 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 instrument. 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 universe score gains and, in turn, effect sizes.

Beyond increasing our confidence in the statistical approach we used to estimate the extent of measurement error for the overall population of students, the relationship between the measurement error variance for individual students and their universe scores, as illustrated in Figure 1, can be utilized in several ways. First consider analyses in which student test scores are entered as right-hand-side variables in regression equations, as is often done in value-added modeling. Some researchers have expressed reservations regarding the use of this approach because of errors-in-variables resulting from test measurement error. However, any such problems can be avoided using information about the pattern of measurement error variances, like that shown in Figure 1, and the approach Sullivan (2001) lays out for estimating regression models with explanatory variables having heteroskedastic measurement error. The method we employ to estimate the overall test measurement error and how the measurement error variance differs across students can be used to compute empirical Bayes estimates of universal scores conditional on the observed test scores, as discussed below. Sullivan's results imply that including such empirical Bayes "shrunk" universal score estimates, rather than actual test scores, as right-hand-side variables will yield consistent estimates of regression coefficients, avoiding any bias resulting from measurement error.[31]

The estimated pattern of measurement error variances in Figure 1 also can be employed to estimate the distributions of universe scores and universe score gains. For example, the more dispersed line in Figure 3 (short dashes) shows the distribution of gains in standardized scale scores between grades four and five. Because of the measurement error embedded in these gain scores, this distribution

[31] Jacob and Lefgren (2005) employ Sullivan's approach to deal with measurement error in estimated teacher effects used as explanatory variables in their analysis. The same logic applies when student test scores are entered as right-hand-side variables.

overstates the dispersion in the universe score gains, $\Delta\tau_{i,5}$. The individual 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 universe score gains, computed using the formula $\Delta S_{i,5}^{EB} = G_i^{\Delta}\,\Delta S_{i,5} + (1-G_i^{\Delta})\,\overline{\Delta S}_5$ where $G_i^{\Delta} \equiv \sigma_{\Delta\tau}^2 \big/ (\sigma_{\Delta\tau}^2 + \sigma_{\Delta\eta_i}^2)$ and $\overline{\Delta S}_5$ is the mean value of $\Delta S_{i,5}$. Even though this empirical Bayes estimator is the best linear unbiased estimator of the underlying parameters for individual students ($\Delta\tau_{i,5}$)[32], the empirical distribution of the empirical Bayes estimates understates the actual dispersion in the distribution of the parameters estimated.[33] Thus, the empirical distribution of the $\Delta S_{i,5}^{EB}$ shown in Figure 3 understates the dispersion in the empirical distribution of universe score gains, $F_N(z) = \sum_i I\left(\Delta\tau_{i,g} \le z\right) \Big/ N$. 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, $E\left[F_N(z)\middle|S\right] = \sum_i \Phi\left(\frac{z-\Delta S_{i,5}^{EB}}{\sigma_{\Delta\eta_\bullet}\sqrt{G_i^{\Delta}}}\right) \Bigg/ N$. This motivates our use of the formula $\hat{F}_N(z)\Big|S = \sum_i \Phi\left(\frac{z-\Delta S_{i,5}^{EB}}{\hat{\sigma}_{\Delta\eta_\bullet}\sqrt{\hat{G}_i^{\Delta}}}\right) \Bigg/ N$ to estimate the empirical density of universe score gain shown by the solid line in Figure 3.[34]

In a similar way, the distributions of universe scores can be analyzed. The more dispersed line in Figure 4 (short dashes) shows the distribution of standardized scale scores in grade five. The line with long dashes is the distribution of empirical Bayes estimates of universe scores, computed using the formula $S_{i,5}^{EB} = G_i\; S_{i,5} + (1-G_i)\;\overline{S}_5$ where $G_i \equiv \sigma_\tau^2 \big/ (\sigma_\tau^2 + \sigma_{\eta_i}^2)$ and $\overline{S}_5$ is the mean value of $S_{i,5}$. As noted above, the empirical distribution of the empirical Bayes estimates understates the actual dispersion in the distribution of the parameters estimated. This motivates our using of the formula $\hat{F}_N(z)\Big|S = \sum_i \Phi\left(\frac{z-S_{i,5}^{EB}}{\hat{\sigma}_{\eta_\bullet}\sqrt{\hat{G}_i}}\right) \Bigg/ N$ to estimate the empirical density of universe scores shown by the solid

[32] $\Delta S_{i,5}^{EB}$ is the value of $\widehat{\Delta\tau_{i,g}}$ which minimizes the loss function $\sum_i \left(\Delta\tau_{i,g} - \widehat{\Delta\tau_{i,g}}\right)^2$.

[33] Louis (1984) and Ghosh (1992).

[34] An alternative would be to utilize the distribution of constrained empirical Bayes estimators, as discussed by Louis (1984), Ghosh (1992) and others.

line in Figure 4. Comparing Figures 3 and 4, it is clear that accounting for test measurement error is far more important in the analysis of gain scores.

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 needs to be reconsidered.

## An Analysis of Teacher Attribute Effect Sizes

In a recent paper, Boyd, Lankford, Loeb, Rockoff and Wyckoff (in press) 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 (14).

$$\mathrm{S}_{\mathrm{ikgty}} - \mathrm{S}_{\mathrm{ik'(g\text{-}1)t'(y\text{-}1)}} = \mathrm{B}_0 + \mathrm{B}_1 Z_{iy} + \mathrm{B}_2 C_{gty} + \mathrm{B}_3 X_{ty} + \pi_i + \pi_g + \pi_y + \varepsilon_{ikgty} \qquad (14)$$

Here the standardized achievement gain score of student *i* in school *k* in grade *g* with teacher *t* in year *y* is a linear function of time-varying characteristics of the student (*Z)*, characteristics of the other students in the same grade having the same teacher in that year (*C)*, and the teacher's qualifications (*X)*. The model also includes student, grade and year 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 (LAST) certification exam on the first attempt, the certification test score, 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.

As shown in Table 5, Boyd et al. (in press) find that teacher experience, teacher certification, SAT scores, competitiveness of the teachers' undergraduate institution, and whether the teacher was recommended for certification by a university-based teacher education program are all statistically significant predictors of achievement, but the size of the effects appear small. We reproduce the parameter estimates for selected measures of teacher attributes from Table 5 in the first column of Table 6. These estimated effects, measured relative to the standard deviation of observed student achievement scores (1.0), 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, most observers believe that the difference between a first- and second-year teacher is meaningful, and the effect of not being certified, and the effect of 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.

The second column of Table 6 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 universe score gains corresponding to Model 2 in Tables 4 and 5: $\hat{\sigma}_{\Delta\tau} = 0.249$. 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 a teacher having no prior experience, is estimated to be over a quarter of a standard deviation in the (universe) 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 16 percent of the standard deviation of the gain in achievement net of measurement error.

Finally, Boyd et al. (in press) 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 observed teacher attributes, holding teacher experience and all of the other variables in Table 5 constant. The teachers in the poorest quartile of schools are divided into quintiles based on their predicted value-added. As shown in the second column of Table 7, the difference in mean estimated teacher effects between teachers in the highest and lowest quintiles is 0.11 (0.18 when experience is not 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 almost half a standard deviation of the universe score gains. As shown in columns 3-8 of Table 7, this meaningful difference in teacher value added is systematically related to teacher attributes – attributes that many have concluded are unrelated to teacher effectiveness. However, we see that only one percent of the teachers in the top quintile of effectiveness are not certified, compared to 73 percent in the bottom quintile. The more effective teachers are less likely to initially have failed the general knowledge certification exam and more likely to have higher scores on this exam as well as on the SAT. Furthermore, almost half of the teachers in the most effective quintile graduated from a college ranked competitive or higher by Barron's, compared to only ten percent of the teachers in the least effective quintile. These differences in effectiveness and teacher qualifications reflect differences *within* the poorest quartile of New York City schools. Given the systematic sorting of teachers between high-poverty and other schools, the differences in teacher effects and attributes likely would be larger had we considered teachers in all NYC schools. The bottom line is that there are important differences in teacher effectiveness that are systematically related to observed teacher attributes.

## Summary

VAM estimation increasingly is being employed to inform policy decisions. The resulting estimates of the effects of teacher attributes using state and district student achievement tests are frequently small by traditional standards. In this paper we explore the role that measurement error plays in creating the perception that observed attributes of teachers matter little. First, we lay out a relatively simple approach for estimating test measurement error from all sources and calculating the standard deviations of universe scores and universe score gains. Second, we apply this approach to estimates of the effect of teacher attributes commonly observed in the literature and find that accounting for measurement error meaningfully increases the estimated importance of teacher attributes for explaining 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., $\beta > 0$) with the degree of persistence reflected in $\beta$ constant across grades. 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 universe scores as well as the dispersion in universe score gains.

We apply these methods to a recent paper that reports VAM estimates of various teacher attributes (Boyd et al., 2008). Many of these estimates appear small when compared to the standard deviation of student achievement – that is effect sizes of less than 0.05. However, the effects are four times larger when measurement error is taken into account, implying that the associated effect sizes are often about 0.16. Furthermore, when teacher attributes are considered jointly, based on the teacher attribute combinations commonly observed, the overall effect of teacher attributes is roughly half a standard deviation of universe score gains – even larger when teaching experience is also allowed to vary. These effects are important from a policy perspective, as in the case of the formulation and implementation of personnel policies.

We have using an analysis of effect sizes associated with teacher attributes to illustrated the importance of accounting for any error in measuring the outcome of interest (e.g., gains in student achievement). More generally, it is important to account for test measurement error in when estimating how any intervention affects student achievement.

Table 1 Auto-Covariance Matrix of Test Scores, $\tilde{\Sigma}_{\bullet}$
Cohorts of New York City Students Entering Grade Three, 1999-2005

| | 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 |

Table 3
Estimates of Underlying Parameter for Alternative Test-Score Auto-Covariance Structures

| | Model 1 | Model 1a | Model 1b | Model 2 | Model 2 |
|---|---|---|---|---|---|
| $\sigma^2_{\eta_\bullet}$ | 0.1699<br>(0.044) | 0.1775<br>(0.026) | 0.1795<br>(0.025) | 0.1680<br>(0.167) | 0.1680<br>(0.167) |
| $\gamma^0$ | 0.8225<br>(0.058) | 0.8149<br>(0.038) | 0.8129<br>(0.038) | 0.8244<br>(0.164) | 0.8244<br>(0.164) |
| $\beta$ | 0.8647<br>(0.432) | 0.9687<br>(0.008) | | 0.6533<br>(12.912) | 0.9778<br>(0.440) |
| $\lambda$ or $\psi$ | 0.0795<br>(0.330) | | -0.0239<br>(0.006) | 0.2521<br>(10.545) | -0.0154<br>(0.220) |
| $\alpha$ | | | | 0.9778<br>(0.440) | 0.6533<br>(12.912) |
| Q | 4.059E-08 | 7.344E-06 | 1.202E-05 | 0.0 | 0.0 |

Table 4
Variance Estimates Associated with the Four Models in Table 3

| | Model 1 | Model 1a | Model 1b | Model 2 |
|---|---|---|---|---|
| Variance in scores for a particular grade ($\hat{\omega}^0$) | 0.9924 | 0.9924 | 0.9924 | 0.9924 |
| Variance in universe scores for grade ($\hat{\gamma}^0$) | 0.8225 | 0.8149 | 0.8129 | 0.8244 |
| Variance in gain scores ($\hat{\sigma}^2_{\Delta S}$) | 0.3980 | 0.3980 | 0.3980 | 0.3980 |
| Variance of the gain in universe scores ($\hat{\sigma}^2_{\Delta\tau}$) | 0.0582 | 0.0430 | 0.0390 | 0.0620 |
| Standard deviation of universe score gains ($\hat{\sigma}_{\Delta\tau}$) | 0.2412 | 0.2074 | 0.1975 | 0.2490 |

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

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Constant | 0.17147 | SD ELA score t-1 | -0.02332 | 14 | 0.1263 | Not certified | -0.04235 |
| | [1.51] | | [1.91] | | [8.21]** | | [5.72]** |
| Student changed schools | -0.03712 | SD math score t-1 | -0.11722 | 15 | 0.1252 | Barrons undergrad college | |
| | [6.60]** | | [8.27]** | | [6.82]** | Most competitive | 0.01498 |
| **Class Variables** | | **Teacher Variables** | | 16 | 0.12464 | | [1.48] |
| Proportion Hispanic | -0.4576 | Experience | | | [6.36]** | Competitive | 0.01426 |
| | [12.89]** | 2 | 0.06549 | 17 | 0.08298 | | [2.24]* |
| Proportion Black | -0.57974 | | [10.61]** | | [3.10]** | Least Competitive | 0.00686 |
| | [16.16]** | 3 | 0.1105 | 18 | 0.14161 | | [1.25] |
| Proportion Asian | -0.07711 | | [16.56]** | | [4.02]** | Imputed Math SAT | 0.00043 |
| | [1.75] | 4 | 0.13408 | 19 | 0.13686 | | [9.05]** |
| Proportion other | -0.56887 | | [17.91]** | | [2.62]** | Imputed Verbal SAT | -0.00034 |
| | [3.95]** | 5 | 0.117 | 20 | 0.24658 | | [6.06]** |
| Class size | 0.002 | | [14.24]** | | [2.50]* | SAT missing | -0.01535 |
| | [3.36]** | 6 | 0.13365 | 21 or more | 0.38977 | | [2.94]** |
| Proportion Eng Lang Learn | -0.42941 | | [14.58]** | | [3.89]** | Initial path into teaching | |
| | [14.16]** | 7 | 0.12307 | Cert pass first | 0.00657 | College Recommended | 0.03108 |
| Proportion home lang Eng | -0.02902 | | [12.27]** | | [0.94] | | [4.95]** |
| | [1.16] | 8 | 0.11898 | Imputed LAST score | 0.00025 | NYC Teaching Fellows | 0.01173 |
| Proportion free lunch | -0.00181 | | [10.81]** | | [0.57] | | [1.10] |
| | [0.01] | 9 | 0.12433 | LAST missing | 0.00188 | Teach for America | 0.02364 |
| Proportion reduced lunch | 0.10521 | | [10.04]** | | [0.26] | | [1.20] |
| | [3.40]** | 10 | 0.13693 | Certified Math | 0.07086 | Individual evaluation | 0.00866 |
| Mean absences t-1 | -0.01367 | | [9.85]** | | [1.30] | | [1.00] |
| | [15.10]** | 11 | 0.12592 | Certified Science | -0.04852 | Other | -0.00138 |
| Mean suspensions t-1 | 0.14069 | | [9.41]** | | [0.95] | | [-0.09] |
| | [2.78]** | 12 | 0.10209 | Certified special ed | 0.01086 | Teacher LAST* | -0.00024 |
| Mean ELA score t-1 | 0.33811 | | [7.66]** | | [1.05] | class proportion free lunch | [0.49] |
| | [31.29]** | 13 | 0.11831 | Certified other | -0.00521 | | |
| Mean math score t-1 | -0.88479 | | [8.23]** | | [0.62] | | |
| | [58.78]** | | | | | Observations | 578,630 |

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 universe 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 |

Table 7
Average Qualifications of Teachers in Poorest Quartile of Schools
by Math Achievement Quintiles Predicted Solely Based on
Teacher Qualifications (excluding experience), 2000-20005

| VA Quintile | Mean VA | Not Certified | LAST Pass First | LAST Score | Math SAT | Verbal SAT | College Ranking Competitive or Higher |
|---|---|---|---|---|---|---|---|
| 1 | -0.068 | 0.731 | 0.46 | 227 | 355 | 440 | 0.101 |
| 2 | -0.032 | 0.141 | 0.656 | 239 | 414 | 467 | 0.121 |
| 3 | -0.01 | 0.076 | 0.779 | 245 | 423 | 462 | 0.224 |
| 4 | 0.01 | 0.031 | 0.851 | 252 | 450 | 470 | 0.352 |
| 5 | 0.045 | 0.013 | 0.908 | 254 | 512 | 474 | 0.494 |
| **Range** | 0.113 | -0.718 | 0.448 | 27 | 157 | 34 | 0.393 |

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

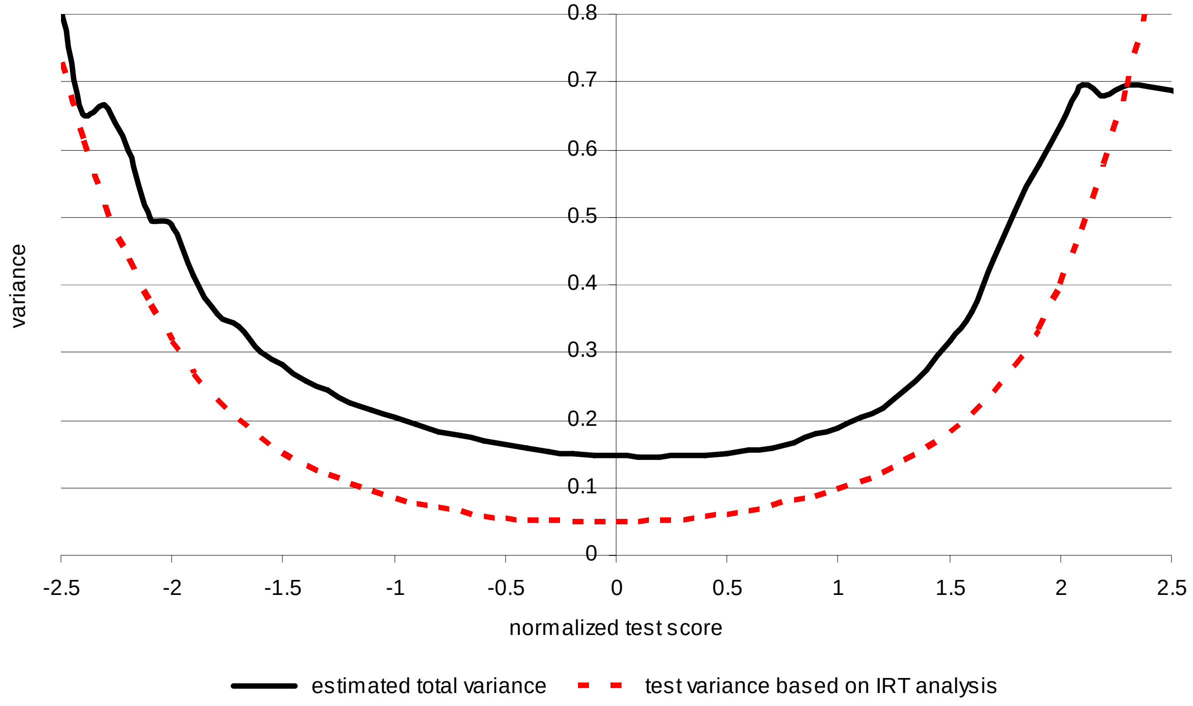


Figure 2

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

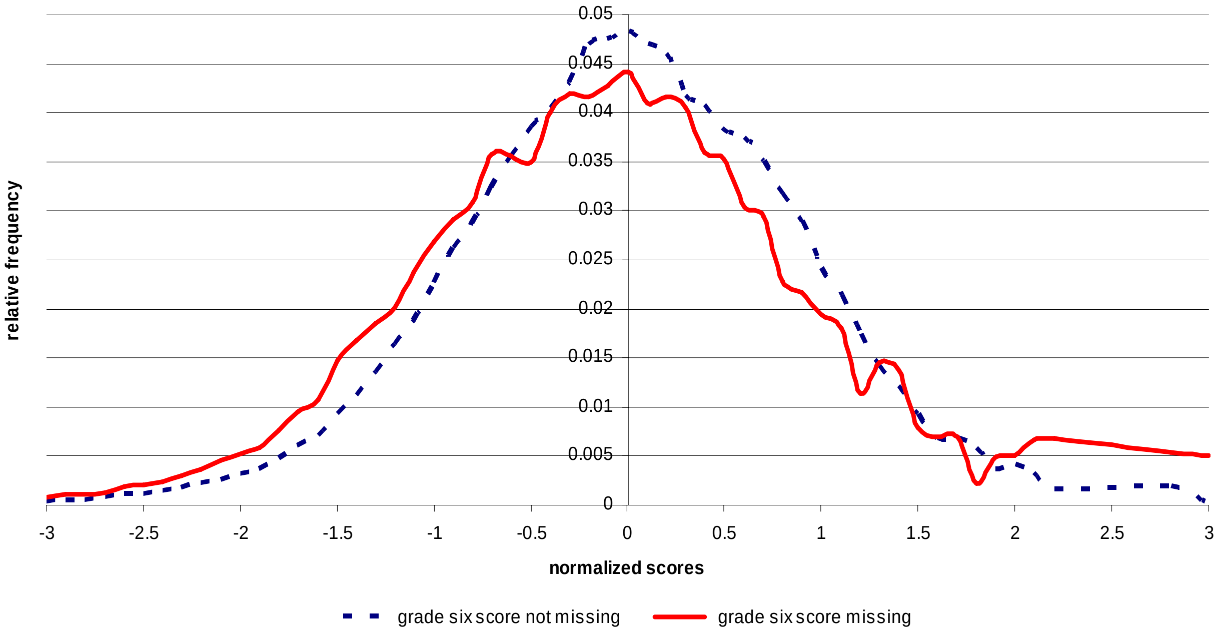


Figure 3
Distribution of Gain Scores, Distribution of the Empirical Bayes Estimates of Universe Score Gains and the Estimated Empirical Distribution of Universe Score Gains, Grade 5

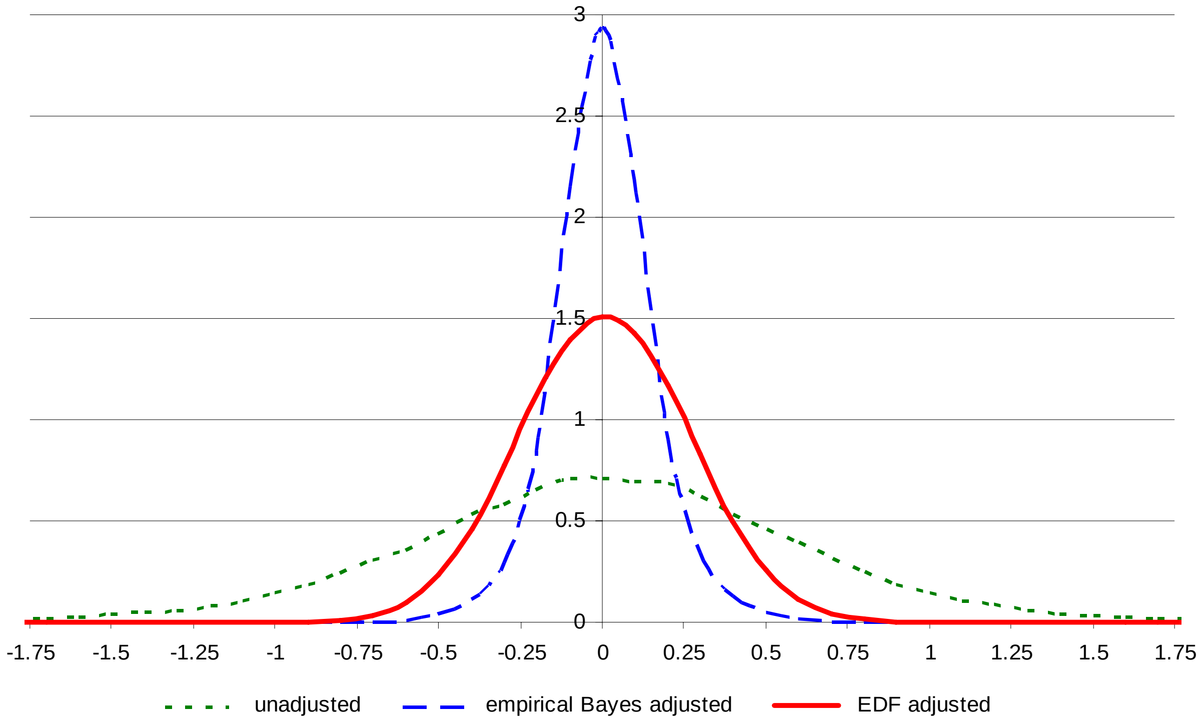
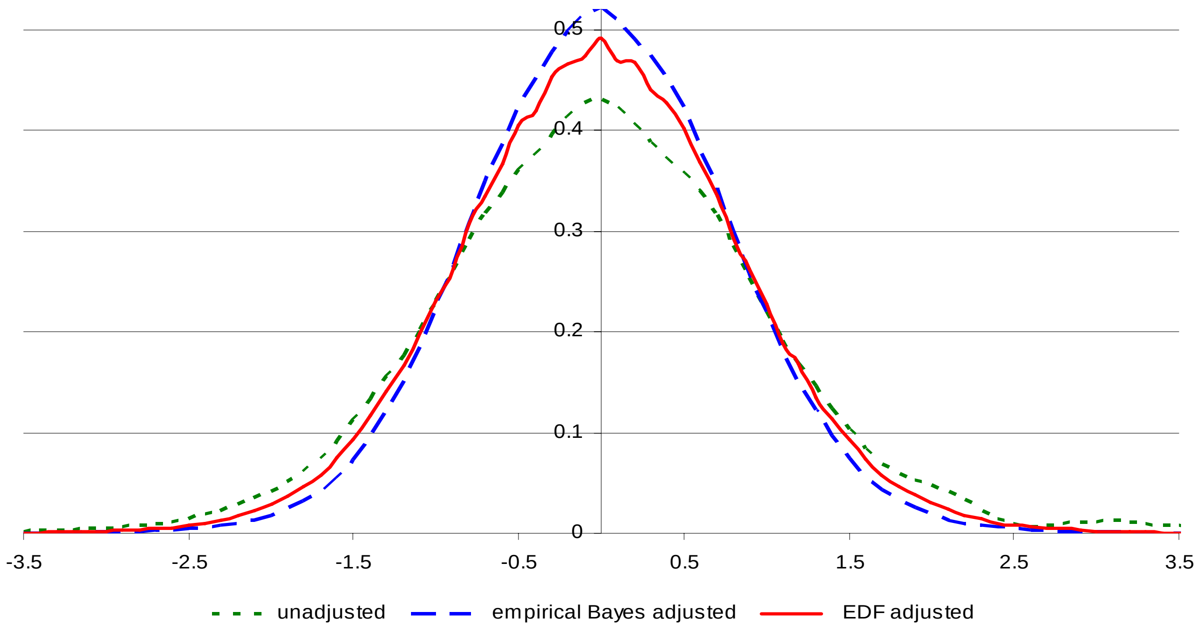


Figure 4
Distribution of Universe Scores, Distribution of the Empirical Bayes Estimates of Universe Scores and the Estimated Empirical Distribution of Universe Scores, Grade 5



## Appendix

Here we derive the formulas for the estimators of the parameters in Model 2. The expressions in (A1) follow from (11) and $\hat{\omega}^j = \omega^j(\chi)$.

$$\begin{aligned}
\hat{\omega}^0 &= \gamma^0 + \sigma^2_{\eta\bullet} \\
\hat{\omega}^1 &= \beta\gamma^0 + \psi \\
\hat{\omega}^2 &= \beta\hat{\omega}^1 + \alpha\ \psi \qquad \text{(A1)} \\
\hat{\omega}^3 &= \beta\hat{\omega}^2 + \alpha^2\psi \\
\hat{\omega}^4 &= \beta\hat{\omega}^3 + \alpha^3\psi
\end{aligned}$$

The last three equations in (A1) can be manipulated to yield $\left(\hat{\omega}^2 - \beta\hat{\omega}^1\right)\left(\hat{\omega}^4 - \beta\hat{\omega}^3\right) - \left(\hat{\omega}^3 - \beta\hat{\omega}^2\right)^2 = 0$. With this being a quadratic function of $\beta$, the expression yields two estimates of $\beta$. In turn, there are two corresponding values of $\hat{\alpha} = \left(\hat{\omega}^3 - \hat{\beta}\hat{\omega}^2\right) \big/ \left(\hat{\omega}^4 - \hat{\beta}\hat{\omega}^3\right)$. However, there is a simple relationship between the two sets of estimates. A different manipulation of the last three equations yields the equations $\left(\hat{\omega}^2 - \alpha\hat{\omega}^1\right)\left(\hat{\omega}^4 - \alpha\hat{\omega}^3\right) - \left(\hat{\omega}^3 - \alpha\hat{\omega}^2\right)^2 = 0$ and $\hat{\beta} = \left(\hat{\omega}^3 - \hat{\alpha}\hat{\omega}^2\right) \big/ \left(\hat{\omega}^4 - \hat{\alpha}\hat{\omega}^3\right)$. Note that the two equation-pairs have the same structure except that the placements of $\alpha$ and $\beta$ are reversed. Thus, the values of $\hat{\beta}$ and $\hat{\alpha}$ are merely reversed in the two cases; one of the roots of the equation $\left(\hat{\omega}^2 - \beta\hat{\omega}^1\right)\left(\hat{\omega}^4 - \beta\hat{\omega}^3\right) - \left(\hat{\omega}^3 - \beta\hat{\omega}^2\right)^2 = 0$ has a corresponding value of $\hat{\alpha} = \left(\hat{\omega}^3 - \hat{\beta}\hat{\omega}^2\right) \big/ \left(\hat{\omega}^4 - \hat{\beta}\hat{\omega}^3\right)$ that is the second root of the former equation. In turn, there are two estimates of $\hat{\psi} = \left(\hat{\omega}^2 - \hat{\beta}\,\hat{\omega}^1\right) \big/ \hat{\alpha}$. Even with this ambiguity regarding the estimation of $\beta$, $\alpha$ and $\psi$, there is a unique estimate of $\hat{\gamma}^0 = \left(\hat{\omega}^1 - \hat{\psi}\right) \big/ \hat{\beta} = \dfrac{\left(\hat{\alpha} + \hat{\beta}\right)\hat{\omega}^1 - \hat{\omega}^2}{\hat{\alpha}\hat{\beta}}$, as a result of the symmetry in how $\hat{\beta}$ and $\hat{\alpha}$ enter the formula. In turn, we can identify $\hat{\sigma}^2_{\eta\bullet} = \hat{\omega}^0 - \hat{\gamma}^0$.

Model 2 illustrates limitations associated with using our empirical approach to identify the parameters characterizing a relatively general dynamic structure underlying the covariance of universe scores. Even so, we are able to estimate $\sigma^2_{\eta\bullet}$ thereby isolating the overall extent of test measurement error.

Estimation of the parameters in Model 2 requires test scores for students spanning five grades. However, analysts often only have access to test data for a shorter grade span. Thus, it is pertinent to consider whether estimates of $\sigma^2_{\eta\bullet}$ based on such shorter grade spans are consistent with those reported in

Table 3, where the parameter estimates are all based on the five moments $\hat{\omega} \equiv \left[ \hat{\omega}^0 \quad \hat{\omega}^1 \quad \hat{\omega}^2 \quad \hat{\omega}^3 \quad \hat{\omega}^4 \right]$. Thus, the estimates of $\sigma^2_{\eta_\bullet}$ in Table 3 only differ because of differences in the model specifications. Estimates would also differ to some degree if we vary the number of moment conditions employed. Below we show the estimation formulas for the parameters in each of the models considered employing the minimum number of moments needed for identification. We then report these exactly identified estimates of the parameters for the four models.

For completeness, the top panel of Table A1 summarizes the structures characterizing the four model specifications as well as the minimum number of moments needed for estimating the parameters of each model. Note that the number of moments is the same as the minimum number of grades for which test scores are needed. Estimation formulas for the parameters of each model are shown in the bottom panel of Table A1. For example, the formulas for Model 2 in the last column summarize results discussed above. For completeness, equation A2 shows the corresponding formula for $\hat{\sigma}^2_{\eta_i}$.

$$\hat{\sigma}^2_{\eta_i} = \frac{\left[ \frac{1}{N_G - 1} \sum_{g=1}^{N_G - 1} \left( S_{i,g+1} - \hat{\beta} S_{i,g} \right)^2 - \hat{\sigma}^2_{\theta} \right]}{\left( 1 + \hat{\beta}^2 \right)} = \frac{\left[ \frac{1}{N_G - 1} \sum_{g=1}^{N_G - 1} \left( S_{i,g+1} - \hat{\beta} S_{i,g} \right)^2 - \left( \left( 1 - \hat{\beta}^2 \right) \hat{\gamma}^0 - 2 \hat{\beta} \hat{\psi} \right) \right]}{\left( 1 + \hat{\beta}^2 \right)} \tag{A2}$$

With test scores for students spanning four grades, the parameter estimates for Model 1 can be obtained using the formulas shown in the first column of Table A1. The corresponding formula for $\hat{\sigma}^2_{\eta_i}$ is the same as that shown above. Test scores for students need only span three grades to estimate the parameters of either Model 1a or Model 1b. The relatively simple estimation formulas for these models are shown in columns (2) and (3), respectively. For Model 1a, the formula for $\hat{\sigma}^2_{\eta_i}$ is as shown in (A3) with the corresponding formula for Model 1b shown in (A4).

$$\hat{\sigma}^2_{\eta_i} = \frac{\left[ \frac{1}{N_G - 1} \sum_{g=1}^{N_G - 1} \left( S_{i,g+1} - \hat{\beta} S_{i,g} \right)^2 - \hat{\sigma}^2_{\theta} \right]}{\left( 1 + \hat{\beta}^2 \right)} = \frac{\left[ \frac{1}{N_G - 1} \sum_{g=1}^{N_G - 1} \left( S_{i,g+1} - \hat{\beta} S_{i,g} \right)^2 - \left( 1 - \hat{\beta}^2 \right) \hat{\gamma}^0 \right]}{\left( 1 + \hat{\beta}^2 \right)} \tag{A3}$$

$$\hat{\sigma}^2_{\eta_i} = \frac{\left[ \frac{1}{N_G - 1} \sum_{g=1}^{N_G - 1} \left( S_{i,g+1} - S_{i,g} \right)^2 - \hat{\sigma}^2_{\theta} \right]}{2} = \frac{\left[ \frac{1}{N_G - 1} \sum_{g=1}^{N_G - 1} \left( S_{i,g+1} - S_{i,g} \right)^2 + 2 \hat{\psi} \right]}{2} \tag{A4}$$

Based on the empirical moments in Table 2, the formulas in Table A1 for the four models imply the parameter estimates reported in Table A2. Comparing these estimates to those in Table 3, the estimates are seen to be robust to the number of moments employed (tested grades needed) in estimation.

Table A1
Summaries of Alternative Models and Formulas for Estimation of Model Parameters
Using the Minimum Number of Empirical Moments Needed for Identification

| | **Model 1** | **Model 1a** | **Model 1b** | **Model 2** |
|---|---|---|---|---|
| | **(1)** | **(2)** | **(3)** | **(4)** |
| **Model Structure** | $\tau_{i,g} = \beta\tau_{i,g-1} + \theta_{i,g}$ | $\tau_{i,g} = \beta\tau_{i,g-1} + \theta_{i,g}$ | $\tau_{i,g} = \tau_{i,g-1} + \theta_{i,g}$ | $\tau_{i,g} = \beta\tau_{i,g-1} + \theta_{i,g}$ |
| | $\theta_{i,g} = \mu_i + \varepsilon_{i,g}$ | $\theta_{i,g} = \varepsilon_{i,g}$ | $\theta_{i,g} = \mu_i + \varepsilon_{i,g}$ | (unspecified) |
| | $\psi^s = E\tau_{i,g}\,\theta_{i,g+s}$<br>$= E\tau_{i,g}\,\mu_i \equiv \lambda$ | $\psi^s = E\tau_{i,g}\,\theta_{i,g+s}$<br>$= 0$ | $\psi^s = E\tau_{i,g}\,\theta_{i,g+s}$<br>$= E\tau_{i,g}\,\mu_i \equiv \lambda$ | $\psi^s = E\tau_{i,g}\,\theta_{i,g+s}$<br>$= \alpha^{s-1}\psi$ |
| | | | | |
| Empirical moments needed for estimation | $\hat{\omega}^0$, $\hat{\omega}^1$, $\hat{\omega}^2$, and $\hat{\omega}^3$<br>(four grades) | $\hat{\omega}^0$, $\hat{\omega}^1$, and $\hat{\omega}^2$<br>(three grades) | $\hat{\omega}^0$, $\hat{\omega}^1$, and $\hat{\omega}^2$<br>(three grades) | $\hat{\omega}^0$, $\hat{\omega}^1$, $\hat{\omega}^2$, $\hat{\omega}^3$, and $\hat{\omega}^4$<br>(five grades) |
| | | | | |
| **Formulas for Estimation** | | | | |
| $\hat{\alpha}$ | | | | $(\hat{\omega}^2 - \hat{\alpha}\hat{\omega}^1)(\hat{\omega}^4 - \hat{\alpha}\hat{\omega}^3) = (\hat{\omega}^3 - \hat{\alpha}\hat{\omega}^2)^2$<br>(Here $\hat{\alpha}$ is implicitly defined.) |
| $\hat{\beta}$ | $\hat{\beta} = (\hat{\omega}^2 - \hat{\omega}^3)\big/(\hat{\omega}^1 - \hat{\omega}^2)$ | $\hat{\beta} = \hat{\omega}^2 \big/ \hat{\omega}^1$ | | $\hat{\beta} = \left(\hat{\omega}^3 - \hat{\alpha}\hat{\omega}^2\right)\Big/\left(\hat{\omega}^4 - \hat{\alpha}\hat{\omega}^3\right)$ |
| $\hat{\psi}$ or $\hat{\lambda}$ | $\hat{\lambda} = \hat{\omega}^2 - \hat{\beta}\hat{\omega}^1$ | | $\hat{\lambda} = \hat{\omega}^2 - \hat{\omega}^1$ | $\hat{\psi} = \left(\hat{\omega}^2 - \hat{\beta}\,\hat{\omega}^1\right)\Big/\hat{\alpha}$ |
| $\hat{\gamma}^0$ | $\hat{\gamma}^0 = \left((1+\hat{\beta})\hat{\omega}^1 - \hat{\omega}^2\right)\Big/\hat{\beta}$ | $\hat{\gamma}^0 = \left(\hat{\omega}^1\right)^2 \Big/ \hat{\omega}^2$ | $\hat{\gamma}^0 = \hat{\omega}^1 - \hat{\lambda} = 2\hat{\omega}^1 - \hat{\omega}^2$ | $\hat{\gamma}^0 = \left(\hat{\omega}^1 - \hat{\psi}\right)\Big/\hat{\beta} = \left(\left(\hat{\alpha} + \hat{\beta}\right)\hat{\omega}^1 - \hat{\omega}^2\right)\Big/\hat{\alpha}\hat{\beta}$ |
| $\hat{\sigma}^2_{\eta_\bullet}$ | $\hat{\sigma}^2_{\eta_\bullet} = \hat{\omega}^0 - \hat{\gamma}^0$ | $\hat{\sigma}^2_{\eta_\bullet} = \hat{\omega}^0 - \hat{\gamma}^0$ | $\hat{\sigma}^2_{\eta_\bullet} = \hat{\omega}^0 - \hat{\gamma}^0$ | $\hat{\sigma}^2_{\eta_\bullet} = \hat{\omega}^0 - \hat{\gamma}^0$ |

Table A2
Estimates of Underlying Parameter for Alternative
Test-Score Auto-Covariance Structures, Minimum Number
of Moments Needed for Identification

| | Model 1 | Model 1a | Model 1b | Model 2 | Model 2 |
|---|---|---|---|---|---|
| $\sigma^2_{\eta_\bullet}$ | 0.1693 | 0.1731 | 0.1741 | 0.1680 | 0.1680 |
| $\gamma^0$ | 0.8231 | 0.8193 | 0.8183 | 0.8244 | 0.8244 |
| $\beta$ | 0.8514 | 0.9651 | | 0.6533 | 0.9778 |
| $\lambda$ or $\psi$ | 0.0899 | | -0.0276 | 0.2521 | -0.0154 |
| $\alpha$ | | | | 0.9778 | 0.6533 |
| Q | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

## References

Aaronson, Daniel,  Lisa Barrow, and William Sander. 2003. "Teachers and Student Achievement in the Chicago Public High Schools."  Federal Reserve Bank Working Paper No 2002-28. Chicago: Federal Reserve Bank of Chicago.

Abowd, John. M., and David Card. 1989. "On the Covariance Structure of Earnings and Hour Changes." *Econometrica* 57 (2): 411-445.

Ballou, Dale. 2002. "Sizing Up Test Scores." *Education Next* 2(2): 10-15.

Baltagi, Badi H.2005. *Econometric Analysis of Panel Data.* West Sussex, England: John Wiley and Sons, Ltd.

Boyd, Donald J., Pamela Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff. 2006. "How Changes in Entry Requirements Alter the Teacher Workforce and Affect Student Achievement." *Education Finance and Policy* 1(2):176-216.

Boyd, Donald J., Pamela L. Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff. 2008. "Who Leaves? Teacher Attrition and Student Achievement." NBER Working Paper No. 14022. Cambridge, MA: National Bureau of Economic Research, Inc.

Boyd, Donald J., Hamilton Lankford, Susanna Loeb, Jonah E. Rockoff, and James Wyckoff. 2008. "The Narrowing Gap in New York City Teacher Qualifications and its Implications for Student Achievement in High-Poverty Schools." *Journal of Policy Analysis and Management* 27(4): 793-818. Fall 2008.

Brennan, Robert. L. 2001. *Generalizability Theory*. New York: Springer-Verlag.

Cameron, A. Colin, and Pravin.K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

Carlin, Bradley P., and Thomas A. Louis.1996. *Bayes and Empirical Bayes Methods for Data Analysis*. Boca Raton: Chapman and Hall/CRC.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2006. "Teacher-Student Matching and the Assessment of Teacher Effectiveness." J*ournal of Human Resources* 41(4): 778–820.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2007. "How and Why Do Teacher Credentials Matter for Student Achievement?" CALDER Working Paper No. 2. Washington, D.C.: National Center for Analysis of Longitudinal Data in Education Research. March 2007.

Cronbach, Lee J., Robert L. Linn, Robert L. Brennan**,** and Edward H. Haertel. 1997. "Generalizability Analysis for Performance Assessments of Student Achievement or School Effectiveness." *Educational and Psychological Measurement* 57: 373-399.

CTB/McGraw-Hill. 2006. "New York State Testing Program 2006: Mathematics, Grades 3-8: Technical Report." Monterey, CA: CTB/McGraw-Hill.

CTB/McGraw-Hill. 2007. "New York State Testing Program 2007: Mathematics, Grades 3-8: Technical Report." Monterey, CA: CTB/McGraw-Hill.

Feldt, Leonard S., and Robert L. Brennan. 1989. “Reliability.” In *Educational Measurement*, edited by Robert L. Linn (3rd ed., 105-146). Washington, D.C.: American Council on Education; Macmillan.

Goldhaber, Dan. 2007. “Everyone’s Doing It, But What Does Teacher Testing Tell Us about Teacher Effectiveness?” *Journal of Human Resources* 42(4) 765-794.

Goldhaber, Dan., and Emily Anthony. 2007. “Can Teacher Quality Be Effectively Assessed? National Board Certification as a Signal of Effective Teaching.” *Review of Economics and Statistics* 89(1): 134-150.

Gordon, Robert, Thomas J. Kane, and Douglas O. Staiger. 2006. *Identifying Effective Teachers Using Performance on the Job*. Washington, D.C.: The Brookings Institution. *The Hamilton Project* Discussion Paper 2006-01.

Gosh, Malay. 1992. “Constrained Bayes Estimation with Applications.” *Journal of the American Statistical Association* 87(418): 533-540.

Haertel, Edward H. 2006. “Reliability.” In *Educational Measurement* edited by Robert L. Brennan (4th ed., 65-110).Westport, CT: American Council on Education/Praeger.

Harris, Douglas N., and Tim R. Sass. 2007. “The Effects of NBPTS Teachers on Student Achievement.” CALDER Working Paper No. 4. Washington, D.C.: National Center for Analysis of Longitudinal Data in Education Research. March 2007.

Hill, Carolyn J., Howard S. Bloom, Alison R. Black, and Mark W. Lipsey. 2007. “Empirical Benchmarks for Interpreting Effect Sizes in Research.” MDRC Working Paper on Research Methodology. New York: MDRC.

Jacob, Brian A., and Lars Lefgren. 2005. “Principals as Agents: Subjective Performance Measurement in Education.” NBER Working Paper No. 11463.Cambridge, MA: National Bureau of Economic Research, Inc.

Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger. 2008. “What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City.” *Economics of Education Review* 27(6): 615-631.December 2008.

Lee, Won-Chan, Robert L. Brennan, and Michael J. Kolen. 2000. “Estimators of Conditional Scale-Score Standard Errors of Measurement: A Simulation Study.” *Journal of Educational Measurement* 37(1): 1-20.

Louis, Thomas A.1984. “Estimating a Population of Parameter Values Using Bayes and Empirical Bayes Methods,” *Journal of the American Statistical Association* 79(386): 393-398.

McCaffrey, Daniel F., J. R. Lockwood, Daniel M. Koretz, Thomas A. Louis, and Laura S. Hamilton. 2004. “Models for Value-Added Modeling of Teacher Effects.” *Journal of Educational and Behavioral Statistics* 29(1): 67-101.

Nye, Barbara, Spyros Konstantopoulos, and Larry V. Hedges. 2004. “How Large are Teacher Effects?” *Educational Evaluation and Policy Analysis* 26(3): 237-257.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. “Teachers, Schools and Academic Achievement.” *Econometrica* 73(2): 417-58.

Rockoff, Jonah E. 2004. “The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data.” *American Economic Review* 94(2): 247-252.

Rogosa, David R., and John B. Willett. 1983. “Demonstrating the Reliability of Difference Scores in the Measurement of Change.” *Journal of Educational Measurement* 20(4): 335-343.

Rothstein, Jesse. 2007. “Do Value-Added Models Add Value? Tracking, Fixed Effects, and Causal Inference.” CEPS Working Paper No. 159. Princeton, NJ: Center for Economic Policy Studies.

Rubin, Donald B. 1987. *Multiple Imputation  for Nonresponse in Surveys*. New York: John Wiley and Sons, Inc.

Sanders, William L., and June C. Rivers. 1996. “Cumulative and Residual Effects of Teachers on Future Student Academic Achievement.” Value-Added Research and Assessment Center Research Progress Report. Knozville, TN: University of Tennessee Value-Added Research and Assessment Center.

Sanford, Eleanor E. 1996. “North Carolina End-of-Grade Tests: Reading Comprehension, Mathematics.” Division of Accountability/Testing, Office of Instruction and Accountability Services Technical Report #1. Raleigh, NC: North Carolina Department of Public Instruction.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall/CRC.

Shen, Wei, and Thomas A. Louis. 1998. “Triple-goal Estimates in Two-Stage Hierarchical Models.”*Journal of the Royal Statistical Society. Series B (Statistical Methodology)*. 60(2): 455-471.

Sullivan, Daniel G. 2001. “A Note on the Estimation of Linear Regression Models with Heteroskedastic Measurement Errors.” Federal Reserve Bank of Chicago Working Paper No. 2001-23. Chicago, IL: Federal Reserve Bank of Chicago.

Thorndike, Robert L. 1951. “Reliability.” In *Educational Measurement* edited by Everet F. Lindquist. Washington, DC: American Council on Education.

Wright, S. Paul, and William L. Sanders. 2007. “Decomposition of Estimates in a Layered Value-Added Assessment Model.” Value-Added Assessment and Research, SAS Institute.

CALDER