NBER WORKING PAPER SERIES

MEASURING TEST MEASUREMENT ERROR: A GENERAL APPROACH

Donald Boyd
Hamilton Lankford
Susanna Loeb
James Wyckoff

Working Paper 18010
http://www.nber.org/papers/w18010

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2012

We gratefully acknowledge support from the National Science Foundation and the Center for Analysis of Longitudinal Data in Education Research (CALDER). Thanks also to Dale Ballou, Daniel McCaffrey and Jeffery Zabel for their helpful comments. The authors are solely responsible for the content of this paper. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Measuring Test Measurement Error: A General Approach
Donald Boyd, Hamilton Lankford, Susanna Loeb, and James Wyckoff
NBER Working Paper No. 18010
April 2012
JEL No. I21

ABSTRACT

Test-based accountability including value-added assessments and experimental and quasi-experimental research in education rely on achievement tests to measure student skills and knowledge. Yet we know little regarding important properties of these tests, an important example being the extent of test measurement error and its implications for educational policy and practice. While test vendors provide estimates of split-test reliability, these measures do not account for potentially important day-to-day differences in student performance.

We show there is a credible, low-cost approach for estimating the total test measurement error that can be applied when one or more cohorts of students take three or more tests in the subject of interest (e.g., state assessments in three consecutive grades). Our method generalizes the test-retest framework allowing for either growth or decay in knowledge and skills between tests as well as variation in the degree of measurement error across tests. The approach maintains relatively unrestrictive, testable assumptions regarding the structure of student achievement growth. Estimation only requires descriptive statistics (e.g., correlations) for the tests. When student-level test-score data are available, the extent and pattern of measurement error heteroskedasticity also can be estimated. Utilizing math and ELA test data from New York City, we estimate the overall extent of test measurement error is more than twice as large as that reported by the test vendor and demonstrate how using estimates of the total measurement error and the degree of heteroskedasticity along with observed scores can yield meaningful improvements in the precision of student achievement and achievement-gain estimates.

Donald Boyd
The Center for Policy Research
University of Albany
135 Western Ave.
Albany, NY 12222
donboyd5@gmail.com

Susanna Loeb
524 CERAS, 520 Galvez Mall
Stanford University
Stanford, CA 94305
and NBER
sloeb@stanford.edu

Hamilton Lankford
School of Education, ED 317
University at Albany
State University of New York
Albany, NY 12222
hamp@albany.edu

James Wyckoff
Curry School of Education
University of Virginia
P.O. Box 400277
Charlottesville, VA 22904-4277
wyckoff@virginia.edu

Recent educational policies such as test-based accountability, teacher evaluation and experimental and quasi-experimental research in education rely on achievement tests as an important metric to assess student skills and knowledge. Yet we know little regarding the properties of these tests that bear directly on their use and interpretation. For example, evidence is often scarce regarding the extent to which standardized tests are aligned with educational standards or the outcomes of interest to policymakers or analysts. Similarly, we know little about the extent of test measurement error and the implications of such error for educational policy and practice. While test vendors provide estimates of reliability, these estimates capture only one of a number of different sources of error.

This paper focuses on test measurement error and demonstrates a credible approach for estimating the overall extent of error. For the tests we analyze, the measurement error is at least twice as large as that indicated in the technical reports provided by test vendors. Such error in measuring student performance results in measurement error in the estimation of teacher effectiveness, school effectiveness and other measures based on student test scores. The relevance of test measurement error in assessing the usefulness of measures such as teacher value-added or schools' adequate yearly progress often is noted but not addressed, due to the lack of easily implemented methods for quantifying the overall extent of measurement error. This paper demonstrates a technique for estimating the total measurement error and provides evidence of the importance of doing so

Thorndike (1951) articulates a variety of factors which can result in a test score being a noisy measure of student achievement. 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., randomness in the selection of test items and the raw-score to scale-score conversion). This information is useful, but provides no information regarding the measurement error from other sources (e.g., students having particularly good or bad days).

Reliability coefficients based on the test-retest approach using parallel test forms is recognized in the psychometric literature as the gold standard for quantifying 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, forgetfulness, carelessness, subjectivity or impulsiveness in response and luck in random guessing"[1] but sufficiently close in time that the knowledge, skills and abilities of individuals taking the tests are unchanged. However, there are relatively few examples of this approach to measurement error estimation in practice, especially in the analysis of student achievement tests used in high-stakes settings.

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 measured with respect to the consistency (i.e., correlation) of students' scores across these parts only accounts for the measurement error resulting from the random selection of a set of test items from the relevant population of items. As Feldt and Brennan (1989) note, this approach "frequently present[s] 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 is that measures based on a single test occurrence ignore potentially important day-to-day differences in student performance.

[1] Feldt and Brennan (1989).

In this paper we show that there is a credible approach for measuring the overall extent of test measurement error that can be applied in a wide variety of settings. Estimation is straightforward and only requires estimates of the correlation or covariance of test scores in the subject of interest at several points in time (e.g., the correlations between third-, fourth- and fifth-grade math scores for one cohort of students).[2] Note that student-level test-score data is not needed, provided that estimates of test-score correlations or covariances are available. Our approach generalizes the test-retest framework to allow for either growth or decay in the knowledge, skills and abilities of students between the test administrations as well as variation across tests in the extent of measurement error. Utilizing the estimated test-score covariance or correlation matrix and a few assumptions regarding the structure of student achievement growth, it is possible to estimate the overall extent of test measurement error and decompose the variance of test scores into the part attributable to real differences in academic achievement and the part attributable to measurement error.

In the following section we briefly introduce generalizability theory, a framework for characterizing multiple sources of test measurement error, and show how the total measurement error is reflected in the covariance structure of observed test scores. This is followed by an explanation of our statistical approach. In turn, we report estimates of the overall extent of measurement error associated with New York State assessments in math and English language arts (ELA) and how the extent of test measurement error varies across ability levels. We conclude with a summary and a brief discussion of ways in which information regarding the

[2] As discussed below, it is necessary that the underlying knowledge one is attempting to measure is measurable using a vertical scale. However, each test instrument employed need only be measured on an interval scale. The interval scales can differ as long as they are linear transformations of the underlying vertical scale (even if this linear transformation is unknown).

extent of test measurement error can be informative in analyses related to educational practice and policy.

1.0 Defining Test Measurement Error

From the perspective of classical test theory, an individual's observed test score is the sum of two components: the *true score* representing the expected value of test scores over some set of test replications, and 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 with the student doing that task at some point in time. The universe of possible occurrences is such that the student's knowledge/skills/ability is the same for all feasible times. Here students are the object of measurement and are assumed to be drawn from some population. As is typical, 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*th student's score on task *t* carried out at occurrence o, which can be decomposed using the random-effects specification in Equation 1.

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

[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 Conbach, Linn, Brennan and Haertel (1997) as well as Feldt and Brennan (1988).

The *universe score* of a student, $\tau_i \equiv \tau + \upsilon_i$, 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, τ_i measures the student's underlying academic achievement, e.g., ability, knowledge and skills. The υ's represent a set of uncorrelated random effects which, along with ε_{ito} and the student's universe score, sum to S_{ito}. Here υ_t and υ_o, respectively, reflect the random effect, common to all test-takers, associated with scores for a particular task and a particular occurrence differing from the population mean, τ. υ_{it} reflects the fact that a student might do especially well or poorly on a particular task. υ_{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). υ_{to} reflects the possibility that the performance of all students on a particular task might vary across occurrences. ε_{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]

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 Equation 2. In turn, the measurement error variance decomposition for a particular student and a single task is shown in Equation 3.

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

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

[5] As noted above, 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 classrooms and there being common random components of measurement error at those levels.

Now consider a test defined in terms of its timing (occurrence) and the N_T tasks making up the examination. The student's actual score, S_{iT}, will equal $\tau_i + \eta_{iT}$ as shown in Equation 4, where η_{iT} is a composite measure reflecting the errors in test measurement from all sources.[6]

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

The variance of η_{iT} for student *i* equals

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

2.0 Test-Score Covariance Structure

We generalize the notation in Equation 4 to allow for multiple tests, for exposition here assumed to be in multiple grades. In Equation 5 S_{ij} is the ith student's score on a test for a

$$S_{ij} = \tau_{ij} + \eta_{ij} \quad (5)$$

particular subject taken in the jth tested grade.[7] τ_{ij} is the i^{th} student's true academic achievement in that subject and grade. We drop subscript "T' to simplify notation, but maintain that a different test in a single occurrence is given in each grade and period. η_{ij} is the corresponding test measurement error from all sources, where $E\eta_{ij} = 0$. Allowing for the possibility of heteroskedasticity across grades and students, $E\eta_{ij}^2 = \sigma^2_{\eta_{ij}}$. Let $\sigma^2_{\eta_{\bullet j}}$ equal $\sigma^2_{\eta_{ij}}$ for all pupils in the homoskedastic case or, more generally, the mean value of $\sigma^2_{\eta_{ij}}$ for the universe of students in

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

[7] In general, time intervals between tests need not be annual nor constant. For example, from a randomized control trial one might know test-score correlations for tests administered at the start and end of the experiment as well as a test given at some point during the experiment.

grade j. The υ in Equation 1 being uncorrelated implies that $E\eta_{ij}\eta_{ik} = 0,\ \forall\, j \neq k$ and $E\eta_{ij}\tau_{ik} = 0,\ \forall j,k$.

Using vector notation, $S_i = \tau_i + \eta_i$ where $S_i' = \left[S_{i1}\ \ S_{i2}\ \ \cdots\ \ S_{iJ}\right]$, $\tau_i' = \left[\tau_{i1}\ \ \tau_{i2}\ \ \cdots\ \ \tau_{iJ}\right]$, and $\eta_i' = \left[\eta_{i1}\ \ \eta_{i2}\ \ \cdots\ \ \eta_{iJ}\right]$ for the first (*j*=1) through the J^{th} tested grades[8]. Equation 6 defines $\Omega(i)$ to be the auto-covariance matrix for the i$^{\text{th}}$ student's observed test scores. Γ is the auto-covariance matrix for the universe scores in the population of students. Z_i is the diagonal matrix with the measurement-error variances across grades for the *i*th student (e.g., $\sigma^2_{\eta_{ij}}$) on the diagonal.

$$\Omega(i) = E\left[\left(S_i - ES_i\right)\left(S_i - ES_i\right)'\right] = E\left[\left(\tau_i - E\tau_i\right)\left(\tau_i - E\tau_{ii}\right)'\right] + \quad E(\eta_i\eta_i^{'})$$

$$= \begin{bmatrix} \omega_{i11} & \omega_{i12} & \cdots & \omega_{i1J} \\ \omega_{i21} & \omega_{i22} & \cdots & \omega_{i2J} \\ \vdots & \vdots & \ddots & \\ \omega_{iJ1} & \omega_{iJ2} & & \omega_{iJJ} \end{bmatrix} = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1J} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2J} \\ \vdots & \vdots & \ddots & \\ \gamma_{J1} & \gamma_{J2} & & \gamma_{JJ} \end{bmatrix} + \begin{bmatrix} \sigma^2_{\eta_{i,1}} & 0 & \cdots & 0 \\ 0 & \sigma^2_{\eta_{i,2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \sigma^2_{\eta_{i,J}} \end{bmatrix} = \Gamma + \mathrm{Z}_i$$

(6) For the population of all students, $\sigma^2_{\eta_{\bullet j}} = E\sigma^2_{\eta_{ij}}$ and $\Omega_{\bullet} = \mathrm{E}\Omega(i) = \Gamma + \mathrm{Z}_{\bullet}$ where $\mathrm{Z}_{\bullet}$ is the diagonal matrix with $\sigma^2_{\eta_{\bullet 1}}, \sigma^2_{\eta_{\bullet 2}}, ..., \sigma^2_{\eta_{\bullet J}}$ on the diagonal. Note that $\Omega(i)$ differs from $\Omega(i')$ only because of possible heteroskedastic measurement error across test-takers.

For a variety of reasons, researchers and policymakers are interested in the decomposition of the overall variance of observed scores for students in a particular grade, ω_{jj}, into the variance in universe scores across the student population, γ_{jj}, and the measurement-error variance; $\omega_{jj} = \gamma_{jj} + \sigma^2_{\eta_{\bullet j}}$. The corresponding *generalizability coefficient,* $G_j = \gamma_{jj} / \omega_{jj}$, measures the portion of the total variation in observed scores that is explained by the variance of universe

[8] For example, the third grade might be the first tested grade. To simplify exposition, we often will not distinguish between the *i*$^{\text{th}}$ grade and the *i*$^{\text{th}}$ tested grade, even though we will mean the latter. Again, the assessments need not be annual; the situation might be one in which several tests are given during a particular year.

scores. The reliability coefficient is the comparable measure in classical test theory. This measure implies the characterization of $\Omega_{\bullet}$ shown in Equation 7.

$$\Omega_{\bullet} = \begin{bmatrix} \omega_{11} & \omega_{12} & \omega_{13} & \cdots \\ & \omega_{22} & \omega_{23} & \cdots \\ & & \omega_{33} & \cdots \\ & & & \ddots \end{bmatrix} = \begin{bmatrix} \gamma_{11}/G_1 & \gamma_{12} & \gamma_{13} & \cdots \\ & \gamma_{22}/G_2 & \gamma_{23} & \cdots \\ & & \gamma_{33}/G_3 & \cdots \\ & & & \ddots \end{bmatrix} \quad (7)$$

$\Omega_{\bullet}$ can be estimated using its empirical counterpart $\tilde{\Omega}_{\bullet} = \sum_i S_i S_i^{'} / N_S$ where N_S is the number of students for whom test scores are observed.[9]

Let ρ_{jk} represent the correlation between the universe scores in grades *j* and *k*; $\rho_{jk} \equiv \gamma_{jk} / \sqrt{\gamma_{jj}\gamma_{kk}}$. This notation along with Equation 7 yields the test-score correlation matrix P shown in Equation 8. Note that the presence of test measurement error (e.g., $G_j < 1$) implies

$$\mathrm{P} = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots \\ & 1 & r_{23} & \cdots \\ & & 1 & \cdots \\ & & & \ddots \end{bmatrix} = \begin{bmatrix} 1 & \sqrt{G_1 G_2}\rho_{12} & \sqrt{G_1 G_3}\rho_{13} & \sqrt{G_1 G_4}\rho_{14} & \cdots \\ & 1 & \sqrt{G_2 G_3}\rho_{23} & \sqrt{G_2 G_4}\rho_{24} & \cdots \\ & & 1 & \sqrt{G_3 G_4}\rho_{34} & \cdots \\ & & & 1 & \cdots \\ & & & & \ddots \end{bmatrix} \quad (8)$$

that each correlation of test scores is smaller than the correlation of the corresponding universe scores (e.g., $r_{jk} < \rho_{jk}$). In contrast, $\omega_{jk} = \gamma_{jk}$, $\forall j \neq k$, so that estimates of the off-diagonal elements of the covariance matrix $\Omega_{\bullet}$ (i.e., $\hat{\omega}_{ij}$) directly imply estimates of the off-diagonal elements of Γ in Equation 7, i.e., $\hat{\gamma}_{ij}$. However, we are primarily interested in separate estimates

[9] This corresponds to the case where one or more student cohorts are tracked through all *J* grades, a key assumption being that the values of the ω_{jk} are constant across the cohorts. A subset of the ω_{jk} can be estimated when the scores for individual students only span a subset of the grades included; a particular ω_{jk} can be estimated provided one has test score data for students in both grades j and k. For example, $\omega_{j,j+1}$, $j = 1,2,3$ can be estimated using test-score data for first-, second- and third-grade students in year-one and the same students in the next grade a year later.

of the universe-score and measurement-error variances, both of which enter the diagonal elements of $\Omega_{\bullet}$. Test-score data can be used to estimate the ω_{jk}, but these estimates by themselves are not sufficient to infer estimates of the γ_{jj} and G_j.

Assuming nothing more than one of the structures typically maintained by researchers estimating models of student achievement growth, the parameters in Equation 7 characterizing the covariance of universe scores (i.e., the γ_{jk}) can be expressed as functions of a smaller set of elemental parameters (e.g., γ_{22} frequently is a function of γ_{11} and other elemental parameters). Taking advantage of such structure, estimates of the ω_{jk} can be used to infer estimates of these elemental parameters, including γ_{11} and G_j. In general, the structure maintained needs to imply that the values of at least *J* of the parameters on the right-hand side of Equation 7 are implied by the values of the remaining parameters.[10] In a similar way, the representation of P in Equation 8 can be used to estimate the G_j; the structures of underlying growth models imply restrictions on the structure of the ρ_{jk} so that test-score correlations can be used to infer estimates of the underlying parameters and the G_j. The central point of our paper is that these methods allow the overall extent of test measurement error to be easily estimated.

Our estimation strategy is closely linked to frameworks laid out by Joreskog (1971, 1978) and Abowd and Card (1989). Abowd and Card develop a framework for studying the covariance structure of individual- and household-level earnings, hours worked and other time-series variables. Their approach falls within the general framework for the analysis of covariance structures developed by Joreskog (1978). Joreskog (1971) employs the kernel of this approach to

[10] Suppose m parameters on the right-hand side of (7) are known functions of the elemental parameters. If $m \geq J$, the number of moments, $J(J+1)/2$, will equal or exceed the number of elemental parameters, $\left(J(J+3)/2\right)-m$, one needs to estimate.

analyze the covariance of congeneric tests. In classical test theory, the set of K tests $S_{ik} = \tau_{ik} + \eta_{ik}$, $k = 1, 2, \ldots, K$, is said to be *congenetric* if the true scores, τ_{ik}, are such that there is a common τ_i where $\tau_{ik} = \lambda_0^k + \lambda_1^k \tau_i, \forall k, i$; here the true scores across tests are perfectly correlated. We consider the case where the τ_{ik} need only be correlated to some degree, following some systematic pattern. This is a generalization of both the test-retest approach and the somewhat more general framework of Joreskog (1971) for estimating the extent of test measurement error and falls within his general approach for the analysis of covariances (Joreskog, 1978).

3.0 Estimation Strategy

3.1 General Approach

We assume that academic achievement, measured by universe scores, is cumulative:

$$\tau_{ij} = \beta_{j-1} \tau_{i,j-1} + \theta_{ij} \qquad (9)$$

This first-order autoregressive structure models student attainment in grade *j* as depending upon the level of knowledge and skills in the prior grade[11] possibly subject to decay (if $\beta_{j-1} < 1$) where the rate of decay can differ across grades. A key assumption is that decay is not complete, as would be the case if $\beta_j = 0$. $\beta_j = \beta$ for all *j* is a special case. The further simplification $\beta_j = 1$ is maintained in many value-added analyses. θ_{ij} is the gain in student achievement in grade j, gross of any decay. [12]

For empirical growth models to actually measure growth in the underlying achievement of students, the test(s) used to measure achievement must reflect a single *interval scale*, meaning

[11] Todd and Wolpin (2003) discuss the conditions under which this will be the case.

[12] In the special case where $\beta_j = 1$, θ_{ij} is the student's gain in achievement while in grade *j*. However, we will refer to θ_{ig} as the student's achievement gain even when $\beta_j \neq 1$.

that "equal-sized gains at all points on the scale represent the same increment of learning".[13] For example, the tests used to estimate $\tau_{i1}, \tau_{i2}, \cdots$ in Equation 9 must all reflect a common vertical scale. As discussed by Ballou (2009): (1) the underlying assumptions regarding test items and test takers needed to assure interval scaling are quite restrictive, (2) those employing test scores in empirical work typically cannot test those assumptions and (3) descriptive statistics for tests of venders claiming their tests are vertically scaled often have properties that bring into question whether this is actually the case. A set of exams not being vertically scaled could be the result of the knowledge and skills being tested not being measurable on a single interval scale. However, the lack of vertical scaling instead could be the result problems in test construction.

The prevalence of questions regarding whether test scales are the same across grades and years explains why analysts often standardize test scores by grade and year to have zero means and unit standard deviations. Empirical analysis employing standardized scores can only provide information regarding the movements of students within the achievement distribution from grade to grade. Fortunately, our approach need only employ test-score correlations. Thus, the individual tests each need to reflect an interval scale, but the scales can differ from grade to grade.[14] Whether or not the tests are vertically scaled, the extent of test measurement error for the individual tests, as measured by the G_j, can be inferred. After first considering situations in which tests are vertically scaled, we discuss the more general and simpler approach that can be employed even when the tests are not necessarily vertically scaled.

Equation 9 can be used to infer Equation 10, which shows that each τ_{ij} reflects the

$$\tau_{ij} = \theta_{ij} + \beta_{j-1}\theta_{i,j-1} + \beta_{j-1}\beta_{j-2}\theta_{i,j-2} + \cdots + (\beta_{j-1}\beta_{j-2}\cdots\beta_{j-(s-1)}\theta_{i,j-(s-1)}) + (\beta_{j-1}\beta_{j-2}\cdots\beta_{j-s}\tau_{i,j-s}) \quad (10)$$

[13] Ballou (2009).
[14] One can think of the underlying model corresponding to the case where actually achievement across grades falls on the same interval scale, but the tests instruments employed need not have that property.

accumulation of decayed values of prior θ_{ij}. As is true in other time-series models, one can assume that the sum in Equation 10 extends back to an infinite past (i.e., $s \to \infty$). A more attractive alternative in our application is to assume that the pertinent time-series for each student begins at a specified point in time (e.g., when she first enters school or the grade in which she is first tested) and employ initial conditions to measure the knowledge and skills of each student at that point in time (e.g., $\tau_{i,j-s}$ for student i where $j-s$ is the starting point). These initial conditions together with Equation 10 and the statistical structure of the θ_{ij} determine the dynamic pattern of universe scores reflected in the parameterization of $\Gamma = E(\tau_i \, \tau_i^{'})$ and $\Omega_{\bullet}$.

Two approaches can be used to characterize the statistical structure of the θ_{ij}. One approach is to fully specify the relationship of achievement gains across grades. For example, in one specification discussed in Appendix A, we assume that $\theta_{ij} = \mu_i + \varepsilon_{ij}$ where μ_i is a student-level random effect and ε_{ij} is white noise. An alternative approach is to assume nothing more than that the joint distribution of $\theta_{i,j+1}$ and τ_{ij} is such that the conditional mean $E(\theta_{i,j+1} \mid \gamma_{ij})$ is a linear function of γ_{ij}. Because of its simplicity and generality, we focus on the reduced-form framework. Several structural models are discussed in Appendix A.

3.2 A Reduced-Form Model

Note that $\theta_{i,j+1} = E\left(\theta_{i,j+1} \mid \tau_{ij}\right) + u_{i,j+1}$ where $u_{i,j+1} \equiv \theta_{i,j+1} - E\left(\theta_{i,j+1} \mid \tau_{ij}\right)$ and $E\, u_{i,j+1}\tau_{ij} = 0$. The assumption that such conditional mean functions are linear in parameters is at the core of regression analysis. We go a step further and assume that $E\left(\theta_{i,j+1} \mid \tau_{ij}\right)$ is a linear function of τ_{ij}, or more generally that such a linear relationship is a reasonably good approximation;

$E\left(\theta_{i,j+1} \mid \tau_{ij}\right) = a_j + b_j\,\tau_{ij}$ where a_j and b_j are parameters.[15] For example, τ_{ij} and $\theta_{i,j+1}$ having a bivariate normal distribution is sufficient, but not necessary, to assure linearity in τ_{ij}. In particular, if the random variables $\tau_{i0}, \theta_{i1}, \theta_{i2}, \cdots, \theta_{ij}, \theta_{i,j+1}, \cdots$ are multivariate normal, $\tau_{ij}, \theta_{i,j+1}, \theta_{i,j+2}, \cdots$ will also be multivariate normal, since τ_{ij} is a linear function of $\tau_{i0}, \theta_{i1}, \theta_{i2}, \cdots, \theta_{ij}$, as shown in Equation 10. For this distribution, $E\left(\theta_{i,j+1} \mid \tau_{ij}\right) = E\theta_{i,j+1} + \left(Cov\left(\tau_{ij}, \theta_{i,j+1}\right) / \gamma_{jj}\right)\left(\tau_{ij} - E\tau_{ij}\right)$, which is linear in τ_{ij}.

The assumption of linearity implies that $\theta_{i,j+1} = a_j + b_j\,\tau_{ij} + u_{i,j+1}$. This along with $\tau_{i,j+1} = \beta_j \tau_{ij} + \theta_{i,j+1}$ implies that $\tau_{i,j+1} = a_j + c_j \tau_{ij} + u_{i,j+1}$ where $c_j \equiv \beta_j + b_j$; a student's universe score in grade *j+1* is a linear function of the universe score in the prior grade. This implies that $\gamma_{j+1,j+1} = c_j^2 \gamma_{jj} + \sigma_{u_{j+1}}$, as well as that $\gamma_{j,j+1} = c_j \gamma_{jj}$, $\gamma_{j,j+2} = c_{j+1} c_j \gamma_{jj}$ and, more generally, $\gamma_{j,j+s} = c_{j+(s-1)} c_{j+(s-2)} \cdots c_j \gamma_{jj}$. These equations along with Equation 7 imply the moments shown in Equation 11 where $\gamma_{j+1,j+1} = c_j^2 \gamma_{jj} + \sigma_{u_{j+1}}$. This structure follows from only assuming that $E\left(\tau_{i,j+1} \middle| \tau_{ij}\right)$ is a linear function of τ_{ij} (e.g., $\tau_{i,j+1} = \beta_j \tau_{ij} + \theta_{i,j+1}$ and $E\left(\theta_{i,j+1} \middle| \tau_{ij}\right)$ is a linear function of τ_{ij}).

$$\Omega_{\bullet} = \begin{bmatrix} \omega_{11} & \omega_{12} & \omega_{13} & \omega_{14} & \cdots \\ & \omega_{22} & \omega_{23} & \omega_{24} & \cdots \\ & & \omega_{33} & \omega_{34} & \cdots \\ & & & \omega_{44} & \cdots \\ & & & & \ddots \end{bmatrix} = \begin{bmatrix} \gamma_{11}/G_1 & c_1\gamma_{11} & c_2c_1\gamma_{11} & c_3c_2c_1\gamma_{11} & \cdots \\ & \gamma_{22}/G_2 & c_2\gamma_{22} & c_3c_2\gamma_{22} & \cdots \\ & & \gamma_{33}/G_3 & c_3\gamma_{33} & \cdots \\ & & & \gamma_{44}/G_4 & \cdots \\ & & & & \ddots \end{bmatrix} \quad (11)$$

[15] This linear specification is a first-order Taylor series approximation of $E\left(\theta_{i,j+1} \mid \tau_{ij}\right)$.

When test-score data for students span J grades, the parameters of the reduced-form covariance structure will include the $2J$ parameters $\gamma_{11}, c_1, c_2, \cdots, c_{J-1}, G_1, G_2, \cdots, G_J$ and either $\sigma^2_{u_2}, \sigma^2_{u_3}, \cdots, \sigma^2_{u_J}$ or a smaller number of parameters that imply the values of $\sigma^2_{u_2}, \cdots, \sigma^2_{u_J}$ (e.g., $\sigma^2_{u_j} = \sigma^2_u, \forall j$). As an example of how such parameters can be estimated, suppose that $G_j = G$ and that test-score data for J=3 grades yields estimates of ω_{11}, ω_{12}, ω_{13}, ω_{22}, ω_{23}, and ω_{33}. The corresponding moment conditions are shown in Equation 12. Substitution of the $\hat{\omega}_{jk}$ for ω_{jk} and manipulation of the six moments yields the estimators for the elemental parameters shown in (13). As is the case here, a few back-of-the-envelope calculations often can yield estimates of the overall extent of test measurement error.

$$\begin{bmatrix} \omega_{11} & \omega_{12} & \omega_{13} \\ & \omega_{22} & \omega_{23} \\ & & \omega_{33} \end{bmatrix} = \begin{bmatrix} \gamma_{11}/G & c_1\gamma_{11} & c_2c_1\gamma_{11} \\ & (c_1^2\gamma_{11}+\sigma_u^2)/G & c_2(c_1^2\gamma_{11}+\sigma_u^2) \\ & & [c_2^2c_1^2\gamma_{11}+(c_2^2+1)\sigma_u^2]/G \end{bmatrix} \quad (12)$$

$$\begin{aligned} \hat{c}_2 &= \frac{\hat{\omega}_{13}}{\hat{\omega}_{12}} \qquad & \hat{G} &= \frac{\hat{\omega}_{12}\hat{\omega}_{23}}{\hat{\omega}_{13}\hat{\omega}_{22}} \qquad & \hat{\gamma}_{11} &= \hat{\omega}_{11}\hat{G} \\ \hat{c}_1 &= \frac{\hat{\omega}_{12}}{\hat{\gamma}_{11}} \qquad & \hat{\sigma}^2_{u_1} &= \hat{\omega}_{22}\hat{G} - c_1^2\hat{\gamma}_{11} \qquad & \hat{\sigma}^2_{u_2} &= (\hat{\omega}_{33} - c_2^2\hat{\omega}_{22})\hat{G} \end{aligned} \quad (13)$$

This example illustrates that an estimate of the covariance of observed test scores together with assumptions regarding the structure of student achievement growth are sufficient to estimate the variance(s) of test measurement error from all sources, as well as the variances in universe scores measuring the dispersion in student achievement in each tested grade. In general, this is possible if student achievement is to some extent cumulative (e.g., $\beta_1 > 0$) and one has an estimate of $\Omega_{\bullet}$ -- the covariance matrix for a sequence of exams measuring student achievement over time (e.g., math test scores of students in three consecutive grades). Achievement being cumulative implies that the universe score variances enter expressions characterizing the off-

diagonal elements of $\Omega_{\bullet}$. For instance, γ_{11} enters the expressions for $\omega_{12} = \gamma_{12}$ and $\omega_{13} = \gamma_{13}$ shown in Equation 12. Thus, estimates of the $\omega_{jk}, j \neq k,$ can be used to infer estimates of the universe score variances. In turn, the extent of test measurement error can be inferred utilizing the diagonal elements of $\Omega_{\bullet}$ (e.g., $\omega_{11} = \gamma_{11}/G_1$).

A similar, but simpler, approach can be employed whether or not the tests utilized are vertically scaled, provided that each is interval scaled. The reduced-form model $\tau_{i,j+1} = a_j + c_j \tau_{ij} + u_{i,j+1}$ and the formulae in Equation 11 imply the following empirical relationships: $\tau^*_{i,j+1} = \rho_{j,j+1}\tau^*_{ij} + u^*_{i,j+1}$ where $u^*_{i,j+1} = u_{i,j+1} \Big/ \sqrt{\gamma_{j+1,j+1}}$, $E\tau^*_{ij} u^*_{i,j+1} = 0$ and τ^*_{ij} and $\tau^*_{i,j+1}$ are standardized universe scores having correlation $\rho_{j,j+1} = c_j \sqrt{\gamma_{jj} / \gamma_{j+1,j+1}}$ (e.g., $\rho_{12} = c_1 \sqrt{\gamma_{11} / \gamma_{22}}$). In addition, $\rho_{j,j+2} = \rho_{j,j+1}\ \rho_{j+1,j+2}$ (e.g., $\rho_{13} = c_2 c_1 \sqrt{\gamma_{11} / \gamma_{33}} = c_1 \sqrt{\gamma_{11} / \gamma_{22}}\, c_2 \sqrt{\gamma_{22} / \gamma_{33}} = \rho_{12}\rho_{23}$), $\rho_{j,j+3} = \rho_{j,j+1}\ \rho_{j+1,j+2}\ \rho_{j+2,j+3}$, etc.. This structure along with Equation 8 implies the moment conditions in Equation 14, where r_{jk} is the

$$\begin{bmatrix} r_{12} & r_{13} & r_{14} & r_{15} & \cdots \\ & r_{23} & r_{24} & r_{25} & \cdots \\ & & r_{34} & r_{35} & \cdots \\ & & & r_{45} & \cdots \\ & & & & \end{bmatrix} = \begin{bmatrix} \sqrt{G_1 G_2}\ \rho_{12} & \sqrt{G_1 G_3}\ \rho_{12}\rho_{23} & \sqrt{G_1 G_4}\ \rho_{12}\rho_{23}\rho_{34} & \sqrt{G_1 G_5}\ \rho_{12}\rho_{23}\rho_{34}\rho_{45} & \cdots \\ & \sqrt{G_2 G_3}\ \rho_{23} & \sqrt{G_2 G_4}\ \rho_{23}\rho_{34} & \sqrt{G_2 G_5}\ \rho_{23}\rho_{34}\rho_{45} & \cdots \\ & & \sqrt{G_3 G_4}\ \rho_{34} & \sqrt{G_3 G_5}\ \rho_{34}\rho_{45} & \cdots \\ & & & \sqrt{G_4 G_5}\ \rho_{45} & \cdots \\ & & & & \ddots \end{bmatrix} \quad (14)$$

test-score correlation for grades *j* and *k* . Because $\sqrt{G_1}$ and ρ_{12} only appear as a multiplicative pair, the two parameters cannot be identified separately, but $\rho^*_{12} \equiv \sqrt{G_1}\ \rho_{12}$ can. The same is true for $\rho^*_{J-1,J} \equiv \sqrt{G_J}\ \rho_{J-1,J}$ where J is the last grade for which one has test scores. After substituting the expressions for ρ^*_{12} and $\rho^*_{J-1,J}$, the $N_m = J(J-1)/2$ moments in Equation 14 are functions of the $N_\pi = 2\mathrm{J} - 3$ parameters in $\pi = \left[G_2\ G_3 \cdots G_{J-1}\ \rho^*_{12}\ \rho_{23} \cdots \rho_{J-2,J-1}\ \rho^*_{J-1,J} \right]$, which can be

identified provided that $J \geq 4$. With one or more additional parameter restriction (e.g., $G_1 = G_2 = G_3$ or $\rho_{23} = \rho_{34}$), $J = 3$ is sufficient for identification.

Whether the moment conditions in Equations 11 or 14 are employed in estimation, the parameters can be estimated using a minimum-distance estimator. For example, suppose the elements of the column vector $r(\pi)$ are the moment conditions on the right-hand-side of Equation 14, after having substituted the expressions for ρ^*_{12} and $\rho^*_{J-1,J}$. With $\hat{r}$ representing the corresponding vector of N_m test-score correlations for a sample of students, the minimum-distance estimator is $\text{argmin}_{\pi}\ [\hat{r} - r(\pi)]'\text{B}[\hat{r} - r(\pi)]$ where B is any positive semi-definite matrix.[16] We employ the identity matrix so that $\hat{\pi}_{MD} = \text{argmin}_{\pi}\ [\hat{r} - r(\pi)]'\ [\hat{r} - r(\pi)]$.[17,18] The estimated generalizability coefficients, in turn, can be used to infer estimates of the pre-normalized universe-score variance, $\hat{\gamma}_{jj} = \hat{G}_j\ \hat{\omega}_{jj}$, as well as the measurement-error variances $\sigma^2_{\eta_{\cdot j}} = \gamma_{jj}(1 - G_j)\ /G_j = (1 - G_j)\,\omega_{jj}$ and $\sigma^2_{\eta^*_{\cdot j}} = 1 - G_j$.[19]

3.3 Additional Points

Before turning to our empirical analysis, consider six important points. First, noted in the introduction, the test-retest approach is a commonly-discussed, but infrequently-employed,

---

[16] If $B \xrightarrow{P} B_0$ and $\text{Rank}[B_0\ \partial r(\pi)/\partial \pi'] \geq N_\pi$, π is locally identified. In the case of a strict equality, the parameters are exactly identified with $\hat{r} = r(\hat{\pi})$ implicitly defining the estimator, which is the same for all B. See Cameron and Trivedi (2005).

[17] $\hat{\pi}_{MD}$, the equally-weighted minimum-distant estimator is consistent, but less efficient than the estimator corresponding to the optimally chosen B. However, $\hat{\pi}_{MD}$ does not have the finite-sample bias problem that arises from the inclusion of second moments. See Altonji and Segal (1996).

[18] $\hat{r}$ having the limit distribution $\sqrt{N_S}\left(\hat{r} - r_0\right) \xrightarrow{d} N[0, V(\hat{\pi})]$ implies that the variance of the minimum-distance estimator is $V\left(\hat{\pi}_{MD}\right) = [Q'Q]^{-1}Q'\,V(\hat{\pi})\,Q\,[Q'Q]^{-1}$ where Q is the matrix of derivatives $Q = \partial r(\pi)/\partial \pi$.

[19] $S_{ij} = \tau_{ij} + \eta_{ij}$ implies that $S^*_{ij} = \sqrt{\gamma_{jj}/\omega_{jj}}\ \tau^*_{ij} + \eta_{ij}/\omega_{jj} = \sqrt{G_j}\ \tau^*_{ij} + \eta^*_{ij}$ where the normalized test- and universe-scores having unit variances. It follows that $1 = G_j + \sigma^2_{\eta^*_{\cdot j}}$ and, in turn, $\sigma^2_{\eta^*_{\cdot j}} = 1 - G_j$.

method for estimating the overall extent of test measurement error. To see that this approach is a special case of our framework, consider the subset of elements in Equation 14 shown in (15), initially focusing on the first equation. The test-retest approach requires that (i) the time between

$$r_{12} = \sqrt{G_1 G_2}\,\rho_{12} \qquad r_{13} = \sqrt{G_1 G_3}\,\rho_{12}\rho_{23} \qquad r_{23} = \sqrt{G_2 G_3}\,\rho_{23} \quad (15)$$

the test and retest is sufficiently short that the skills and knowledge of those tested are unchanged so that $\rho_{12} = 1$ and (ii) the tests are administered under identical conditions so that the overall extent of measurement error is the same for the two tests (e.g., $G_1 = G_2$). Under these conditions, the first equation in (15) reduces to $r_{12} = G$; the estimated correlation of scores from the two tests is an estimate of the generalizability (reliability) coefficient for the tests. Joreskog (1971) maintains the assumption that the universe scores are perfectly correlated but allows the extent of measurement error to differ across the tests. (In this case the expressions in (15) imply that $G_1 = r_{12}\, r_{13}/r_{23}$, $G_2 = r_{12}\, r_{23}/r_{13}$, and $G_3 = r_{13}\, r_{23}/r_{12}$.) Our approach goes meaningfully further in allowing the universe-score correlations to be less than one and different between test pairs.

Second, to estimate the overall extent of measurement error for a population of students one only needs descriptive statistics of scores on each test and test-score correlations, an attractive feature of our approach. However, additional inferences are possible when student-level test-score data are available.

Third, our approach is applicable whether the measurement-error variance is constant across students in each grade (i.e., $\sigma^2_{\eta_{ij}} = \sigma^2_{\eta_{\bullet j}}, \forall i$) or there is heteroskedasticity, where $\sigma^2_{\eta_{\bullet j}}$ is the mean variance for the population of students (i.e., $\sigma^2_{\eta_{\bullet j}} = E\sigma^2_{\eta_{ij}}$). In the latter case, it is possible to explore the extent and pattern of heteroskedasticity, provided one has student-level test scores. Consider the case where $\pi = \left[G_2\ G_3\ G_4 \cdots \rho^*_{12}\ \rho_{23}\ \rho_{34} \cdots \right]$ has been estimated. The

relationship $\tau^*_{i,j+1} = \rho_{j,j+1}\tau^*_{ij} + u^*_{i,j+1}$ for the standardized universe scores τ^*_{ij} and $\tau^*_{i,j+1}$ implies that $\sigma^2_{u^*_{i,j+1}} = 1 - \rho^2_{j,j+1}$.[20] Note that $S^*_{ij} = \sqrt{G_j}\tau^*_{ij} + \eta^*_{ij}$ and, in turn, $\tau^*_{ij} = (S^*_{ij} - \eta^*_{ij})/\sqrt{G_j}$. These expressions imply that $S^*_{i,j+1} - \rho_{j,j+1}\sqrt{G_{j+1}/G_j}S^*_{ij} = \sqrt{G_{j+1}}\,u^*_{i,j+1} + \eta^*_{i,j+1} - \rho_{j,j+1}\sqrt{G_{j+1}/G_j}\eta^*_{ij}$. It follows that the variances of the expressions before and after the equality are also equal, which implies that $\sigma^2_{\eta^*_{i,j+1}} + \rho^2_{j,j+1}\left(G_{j+1}/G_j\right)\sigma^2_{\eta^*_{ij}} = V\left(S^*_{i,j+1} - \rho_{j,j+1}\sqrt{G_{j+1}/G_j}S^*_{ij}\right) - G_{j+1}\left(1-\rho^2_{j,j+1}\right)$.

To provide some needed structure, suppose that $\sigma^2_{\eta^*_{ij}} = \alpha_i\,\sigma^2_{\eta^*_{\bullet j}}$; the ratio of a student's measurement-error variance to the population mean variance is constant across grades. If so, Equation 16 follows, which suggests that α can be estimated for a group of students, *C*, having the same (unknown) value α_C, as shown in Equation 17.[21] Rather than grouping students based upon one or more observed attributes, student-level values of α_i can be estimated using a regression approach described below where we estimate the extent to which α_i varies with the level of student achievement.

$$\alpha_i = \frac{V\left(S^*_{i,j+1} - \rho_{j,j+1}\sqrt{G_{j+1}/G_j}S^*_{ij}\right) - G_{j+1}\left(1-\rho^2_{j,j+1}\right)}{\sigma^2_{\eta^*_{\bullet,j+1}} + \rho^2_{j,j+1}\left(G_{j+1}/G_j\right)\sigma^2_{\eta^*_{\bullet j}}} \quad (16)$$

$$\hat{\alpha}_C = \tfrac{1}{N_C}\sum_{i\in C}\frac{\left(S^*_{i,j+1} - \hat{\rho}_{j,j+1}\sqrt{\hat{G}_{j+1}/\hat{G}_j}S^*_{ij}\right)^2 - \hat{G}_{j+1}\left(1-\hat{\rho}^2_{j,j+1}\right)}{\hat{\sigma}^2_{\eta^*_{\bullet,j+1}} + \hat{\rho}^2_{j,j+1}\left(\hat{G}_{j+1}/\hat{G}_j\right)\hat{\sigma}^2_{\eta^*_{\bullet j}}} \quad (17)$$

The reduced-form framework provides a useful tool for estimating the extent of test measurement error from all sources. Estimation is straightforward and the key assumptions

[20] This follows because $E\,u_{i,j+1}\tau_{ij} = 0$ implies that $E\,u^*_{i,j+1}\tau^*_{ij} = 0$.

[21] The formula in (16) only maintains that α_i for each student is constant across grades j and j+1. The formula easily can be generalized to reflect α_i being constant across more than two adjacent grades.

underlying the empirical model (i.e., $\tau_{i,j+1} = \beta_j \tau_{ij} + \theta_{i,j+1}$ with $\beta_j \neq 0$ and $E\left(\theta_{i,j+1} \middle| \tau_{ij}\right)$ is a linear function of τ_{ij}) appear to be quite reasonable. A fourth point is that the assumptions need not be accepted as an article of faith; together they imply that $\tau_{i,j+1}$ is a linear function of τ_{ij} (i.e., $\tau_{i,j+1} = a_j + c_j \tau_{ij} + u_{i,j+1}$ where $E\,\tau_{ij}\, u_{i,j+1} = 0$), which can be tested utilizing test-score data, as demonstrated below.

The fifth point is that even though the following empirical analysis utilizes the reduced-form model, our general approach is not dependent on this particular specification. One can carry out empirical analyses employing fully-specified statistical structures for the θ_{ij}. A variety of specifications can be employed, provided the specifications imply covariance structures where the number of moment conditions are sufficient to estimate the number of parameters, including the generalizability coefficients. The estimation strategy for the structural approach is the same as above, the only difference being that the moment conditions employed will include the full set of structural parameters, not a subset of the structural parameters and a set of reduced-form parameters. We discuss the structural approach and alternative model specifications in more detail in Appendix A.

Finally, the parameters entering the covariance structure can be estimated without specifying the distributions of τ_{ij} and η_{ij}. However, additional inferences are possible when one assumes particular functional forms and has student-level test scores. When needed, we assume that τ_{ij} and η_{ij} are normally distributed. When η_{ij} is either homoskedastic or heteroskedastic with $\sigma^2_{\eta_{ij}}$ not varying with the level of ability, τ_{ij} and S_{ij} will be bivariate normal, implying that the conditional distribution of τ_{ij} given S_{ij} will be normal with moments $E\left(\tau_{ij} \middle| S_{ij}\right) =$

$(1-G_{ij})\mu_j + G_{ij}S_{ij}$ and $V\left(\tau_{ij}\left|S_{ij}\right.\right) = (1-G_{ij})\gamma_{jj}$ where $\mu_j \equiv E\tau_{ij} = ES_{ij}$ and $G_{ij} = \gamma_{jj}\Big/\left(\gamma_{jj} + \sigma^2_{\eta_{ij}}\right)$. Here $E\left(\tau_{ij}\left|S_{ij}\right.\right)$ is the Bayesian posterior mean of τ_{ij} given S_{ij} – the best linear unbiased predictor (BLUP) of the student's actual ability. $V\left(\tau_{ij}\left|S_{ij}\right.\right)$ and easily computed Bayesian confidence (credible) intervals can be employed to measure the precision of the BLUP estimator for each student.

When the extent of test measurement error systematically varies across ability levels (i.e., $\sigma_{\eta_{ij}} = \sigma_{\eta_j}(\tau_{ij})$) – as is the case in our application – the normal density of η_{ij} is $g^j\left(\eta_{ij}\left|\tau_{ij}\right.\right) = \phi\left(\eta_{ij}\Big/\sigma^2_{\eta_j}(\tau_{ij})\right)\Big/\sigma_{\eta_j}(\tau_{ij})$ where $\phi(\ )$ is the standard-normal density. The joint density of τ_{ij} and η_{ij} is $h^j\left(\eta_{ij},\tau_{ij}\right) = g^j\left(\eta_{ij}\left|\tau_{ij}\right.\right) f^j(\tau_{ij}) = \dfrac{1}{\sigma_{\eta_j}(\tau_{ij})\sqrt{\gamma_{jj}}}\phi\left(\eta_{ij}\Big/\sigma^2_{\eta_j}(\tau_{ij})\right)\phi\left((\tau_{ij}-\mu_j)\Big/\sqrt{\gamma_{jj}}\right)$ which is not bivariate normal, due to $\sigma_{\eta_{ij}}$ being a function of τ_{ij}. (In this case S_{ij} is a mixture of normal random variables.) The conditional density of τ_{ij} given S_{ij} is $h^j\left(S_{ij}-\tau_{ij},\tau_{ij}\right)\Big/k^j\left(S_{ij}\right)$. Here $k^j\left(S_{ij}\right) = \int_{-\infty}^{\infty} h^j\left(S_{ij}-\tau_{ij},\tau_{ij}\right)d\tau_{ij} = \int_{-\infty}^{\infty} g^j\left(S_{ij}-\tau_{ij}\left|\tau_{ij}\right.\right) f^j(\tau_{ij})\,d\tau_{ij}$ is the density of S_{ij}. Given any particular function $\sigma^2_{\eta_{ij}} = \sigma^2_{\eta_j}(\tau_{ij})$, this integral can be calculated using Monte Carlo integration with importance sampling; $k^j\left(S_{ij}\right) = \sum_{m=1}^{M} g^j\left(S_{ij}-\tau^*_{mj}\left|\tau^*_{mj}\right.\right)\Big/M$ where τ^*_{mj}, $m = 1,2,\cdots,M$, is a sufficiently large set of random draws from the distribution $f^j(\tau_{ij})$. Similarly, the posterior mean ability level given any particular score is $E\left(\tau_{ij}\left|S_{ij}\right.\right) = \dfrac{1}{k^j\left(S_{ij}\right)M}\sum_{m=1}^{M}\tau^*_{mj}\,g^j\left(S_{ij}-\tau^*_{mj}\left|\tau^*_{mj}\right.\right)$. Also, $P\left(\tau_{ij} < a\left|S_{ij}\right.\right) = \dfrac{1}{k^j\left(S_{ij}\right)M}\sum_{\tau^*_{mj}<a} g^j\left(S_{ij}-\tau^*_{mj}\left|\tau^*_{mj}\right.\right)$ is

the cumulative posterior distribution of τ_{ij} which can be used to infer Bayesian confidence intervals (i.e., credible intervals). For example, the 80 percent credible interval is (*L, U*) such that $P\left(L \le \tau_{ij} \le U \left| S_{ij} \right.\right)$=0.80. Here we choose the lower- and upper-bounds corresponding to the values of *a* such that $P\left(\tau_{ij} < a \left| S_{ij} \right.\right) = 0.10$ and $P\left(\tau_{ij} < a \left| S_{ij} \right.\right) = 0.90$.[22]

4.0 An Empirical Application

We estimate the parameters in the reduced-form model employing moments defined in terms of the correlations of scores on the third- through eighth-grade New York State math and ELA tests for the cohort of New York City students who were in the third grade in the 2004-2005 school year. Students making normal grade progression were in the eighth grade in 2009-2010. The exams, developed by CTB-McGraw Hill, are aligned to the New York State learning standards and are given to all registered students, with limited accommodations and exclusions. Table 1 reports descriptive statistics for the cohort of students studied. Correlations for ELA and Math are shown below the diagonals in Tables 3 and 4.[23]

4.1 Testing Model Assumptions

[22] A common alternative is to define the credible/confidence interval to be the narrowest interval (*L, U*) for which $P\left(L \le \tau_{ij} \le U \left| S_{ij} \right.\right) = 0.80$. In the computation of $P\left(\tau_{ij} < a \left| S_{ij} \right.\right)$ we employ estimates of $\mu_j, \sigma^2_{\tau_j}$, and the parameters in the function $\sigma^2_{\eta_j}(\tau_{ij})$, but do not adjust the formula for $P\left(\tau_{ij} < a \left| S_{ij} \right.\right)$ to account for this imprecision.

[23] There are a nontrivial number of missing test scores. For example, consider the percent of students having scores in the data for a particular grade but missing score for the next grade. The percentage of missing scores in the following grade averages seven percent across grades in each subject. The extent to which this is a problem depends upon the reasons for the missing data. There is little problem if scores are missing completely at random. (See Rubin (1987) and Schafer (1997).) However, this does not appear to be the case for the NY tests. In particular, we find evidence that students having missing scores typically score relatively low in the grades where scores are present. The exception is that there are some missing scores for otherwise high-scoring students who skip the next grade. To avoid statistical problems associated with this systematic pattern of missing scores, we impute values of missing scores using SAS Proc MI. The Markov Chain Monte Carlo procedure is 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.

We first explore whether the data is consistent with the assumptions which imply that $E\left(\tau_{i,j+1} \mid \tau_{ij}\right)$ is a linear function of τ_{ij}. While the two assumptions are sufficient to assure the linearity of $E\left(\tau_{i,j+1} \mid \tau_{ij}\right)$, it is this linearity that implies the structure of correlations shown in (14). It is fortunate that we are able to assess whether $E\left(\tau_{i,j+1} \mid \tau_{ij}\right)$ is in fact linear.

The lines in Figures 1 and 2 are empirical, nonparametric estimates of the function $E\left(S_{i,j+1} \mid S_{ij}\right)$ for ELA and math, respectively, showing how the observed scores of students in the eighth grade are related to scores in the prior grade. The bubbles with white fill show the actual combinations of observed seventh- and eighth-grade scores; the area of each bubble reflects the relative number of students with that score combination.

The dark bubbles toward the bottoms of Figures 1 and 2 show the IRT standard errors of measurement (SEMs) for the seventh grade tests (in reference to the right vertical axis) reported in the test technical reports.[24] Note that the extent of measurement error associated with the test instrument is meaningfully larger for both low and high scores, reflecting the nonlinear mapping between raw and scale scores. Each point of the standard errors of measurement plot corresponds to a particular scale score as well as a corresponding raw score; movements from one dot to the next (left to right) reflect a one-point increase in the raw score (e.g., one additional question being answered correctly), with the scale-score change shown on the horizontal axis. For example, starting at an ELA scale score of 709, a one point raw-score increase corresponds to a 20 point increase in the scale score to 729. In contrast, starting from a scale score of 641, a one point increase in the raw score corresponds to a two point increase in the scale score. This varying coarseness of the raw- to scale-score mappings – reflected in the varying spacing of

[24] CTB/McGraw-Hill (2006, 2007, etc.).

points aligned in rows and columns in the bubble plot – explains why the reported scale-score standard errors of measurement are substantially higher for both low and high scores. Even if the variance were constant across the range of raw scores – as assumed in classical test theory used to produce the reliability coefficient estimates in the technical reports – the same would not be the case for scale scores.

The fitted nonparametric curves in Figures 1 and 2, as well as very similar results for other grades, provide strong evidence that $E\left(S_{i,j+1} \mid S_{ij}\right)$ is not a linear function of S_{ij}. Even so, this does not contradict our assumption that $E\left(\tau_{i,j+1} \mid \tau_{ij}\right)$ is a linear function of τ_{ij}; test measure error can explain $E\left(S_{i,j+1} \mid S_{ij}\right)$ being S-shaped even when $E\left(\tau_{i,j+1} \mid \tau_{ij}\right)$ is linear in τ_{ij}. It is not measurement error *per se* that implies $E\left(S_{i,j+1} \mid S_{ij}\right)$ will be an S-shaped function of S_{ij}; $E\left(S_{i,j+1} \mid S_{ij}\right)$ will be linear in S_{ij} if the measurement-error variance is constant (i.e., $\sigma^2_{\eta_{ij}} = \sigma^2_{\eta_{\bullet j}}, \forall i$). However, $E\left(S_{i,j+1} \mid S_{ij}\right)$ will be a S-shaped function of S_{ij} when η_{ij} is heteroskedastic with $\sigma_{\eta_{ij}} = \sigma_{\eta_j}(\tau_{ij})$ having a U-shape (e.g., the SEM patterns shown in Figures 1 and 2). The explanation and an example are included in Appendix B.

Appendix B also includes a discussion of how information regarding the pattern of test measurement error can be used to obtain consistent estimates of the parameters in a corresponding polynomial specification of $E\left(\tau_{i,j+1} \mid \tau_{ij}\right)$. We utilize this approach to eliminate the inconsistency of the parameter estimates associated with the measurement error reflected in the SEMs reported in the technical reports. Even though this does not eliminate any

inconsistency of parameter estimates resulting from other sources of measurement error, we are able to adjust for the meaningful heteroskedasticity reflected in the reported SEMs.[25]

Results from using this approach to analyze the NY test data are shown in Figures 3 and 4 for ELA and math, respectively. As an example, consider graph (d) in either figure. The thicker, S-shaped curve corresponds to the OLS estimation of S_{i5} regressed on S_{i4} using a cubic specification. We employ a third-order polynomial because it is the lowest-order specification that can capture the general features of the nonparametric estimates of $E\left(S_{i,j+1} \mid S_{ij}\right)$ in Figures 1 and 2. The dashed line is a cubic estimate of $E\left(\tau_{i,j+1} \mid \tau_{ij}\right)$ obtained using the approach described in Appendix B to avoid parameter-estimate inconsistency associated with that part of test measurement error reflected in the SEMs reported in the technical reports. For comparison, the straight line is the estimate of $E\left(\tau_{i,j+1} \mid \tau_{ij}\right)$ employing this approach and a linear specification. Across the grade-pair graphs, it is striking how close the consistent cubic estimates of $E\left(\tau_{i,j+1} \mid \tau_{ij}\right)$ are to being linear.[26] Overall, the assumption that $E\left(\tau_{i,j+1} \mid \tau_{ij}\right)$ is a linear function of τ_{ij} appears to be quite reasonable in our application.

4.2 Estimated Model

Parameter estimates for the reduced-form model and their standard errors are reported in

[25] As discussed below, how the reported SEMs vary with the level of ability is quite similar to our estimates of how the standard deviations of the measurement-error from all sources vary with ability. If true, by accounting for the heteroskedasticity in the measurement error associated with the test instrument, we are able to roughly account for the effect of heteroskedasticity, increasing our confidence in the estimated *curvature* of $E\left(\tau_{i,j+1} \mid \tau_{ij}\right)$ for each grade and subject. At the same time, not accounting for other sources of measurement error will result in the estimated cubic specification generally being flatter than $E\left(\tau_{i,j+1} \mid \tau_{ij}\right)$.

[26] The cubic estimates of $E\left(\tau_{i,j+1} \mid \tau_{ij}\right)$ in the graphs might be even closer to linear if we had accounted for all measurement error. This was not done to avoid possible circularity; one could question results where the estimates of the overall measurement-error variances are predicated maintaining linearity and the estimated variances are then used to assess whether $E\left(\tau_{i,j+1} \mid \tau_{ij}\right)$ is in fact linear.

Table 2. First consider how well the estimated models fit the observed score correlations. The empirical correlations for ELA and math, respectively, are shown below the diagonals in Tables 3 and 4. The predicted correlations implied by the estimated models are above the diagonals. To evaluate goodness of fit, consider the absolute differences between the empirical and predicted correlations. The average, and average percentage, absolute differences for ELA are 0.001 and one-fifth of one percent, respectively. For math, the differences are 0.003 and one-half of one percent. Thus, the estimated reduced-form models fits the New York data quite well.

Returning to Table 2, the estimated generalizability coefficients for math are meaningfully larger than those for ELA, and the estimates for ELA are higher in some grades compared to others. These differences are of sufficient size that one could reasonably question whether they reflect underlying differences in the extent of test measurement error. Instead the pattern could reflect estimation error or a fundamental shortcoming of our approach, or both. Fortunately, we can compare these estimates to the reliability measures reported in the technical reports for the New York tests, to see whether the reliability coefficients differ in similar ways. The top two lines in Figure 5 marked with squares show the reported reliability coefficients for math (solid line) and ELA (dashed line). The lower two lines marked with diamonds show the generalizability coefficient estimates reported in Table 2. It is not surprising that the estimated generalizability coefficient are smaller than the corresponding reported reliability coefficients, as the latter statistics do not account for all sources of measurement error. However, consistencies in the patterns are striking. The differences between the reliability and generalizability coefficients vary little across grades and subjects, averaging 0.117. Reflecting this result, the generalizability coefficient estimates for math are higher than those for ELA, mirroring corresponding difference between the reliability coefficients reported in the technical reports.

Also, in each subject the variation in the generalizability coefficient estimates across grades closely mirrors the corresponding across-grade variation in the reported reliability coefficients. This is especially noteworthy given the marked differences between math and ELA in the patterns across grades.

The primary motivation for this paper is the desire to estimate the overall extent of measurement error motivated by concern that the measurement error in total is much larger than that reported in test technical reports. The estimates of the overall extent of test measurement error on the NY math exams, on average, are over twice as large as that indicated by the reported reliability coefficients. For the NY ELA tests, the estimates of the overall extent of measurement error average 130 percent higher than that indicated by the reported reliability coefficients. The extent of measurement error from other sources appears to be at least as large as that associated with the construction of the test instrument.

Estimates of the variances in actual student achievement can be obtained employing estimates of the overall extent of test measurement error together with the test-score variances. Universe-score variance estimates for our application are reported in column (3) of Table 5. It is also possible to infer estimates of the variances of universe-score gains shown in column (6). Because these values are much smaller than the variances of test-score gains, the implied generalizability coefficient estimates in column (7) are quite small, especially for ELA.

Estimation of the overall extent of measurement error for a population of students only requires descriptive statistics of scores for each test and test-score correlations. However, additional inferences are possible when student-level test-score data are available. In particular, student-level data and the formula in Equation 16 can be used to estimate $\alpha_j(\tau_i) \equiv \sigma^2_{\eta_j}(\tau_i) \Big/ \sigma^2_{\eta_{\cdot j}}$ characterizing how the variance of measurement error varies with student ability. For example,

for grade pairs 4-5 and 6-7 we compute $\hat{\alpha}_i$ for each student and grade pair using Equation 16. Assuming that a student's mean (normalized) universe score across the grade pair, τ_i^*, is the relevant measure of ability in $\alpha_j(\tau_i)$, τ_i^* can be estimated using a student's average normalized test score for the adjacent grades, $\overline{S}_i^*$. Here we estimate $\alpha_j(\tau_i^*)$ employing a fourth-order polynomial. However, regressing $\hat{\alpha}_i$ on $\overline{S}_i^*$ would yield inconsistent parameter estimates as a result of $\overline{S}_i^*$ measuring τ_i^* with error. If $\lambda_{ik} \equiv E\left((\tau_i^*)^k \left| \overline{S}_i^* \right.\right)$, $k = 1,2,3,4$, were know for each student, consistent estimates of polynomial parameters could be obtained by regressing $\hat{\alpha}_i$ on λ_{i1}, λ_{i2}, λ_{i3}, and λ_{i4}.[27] The problem is that computation of λ_{ik} requires knowledge of $\alpha_j(\tau_i^*)$ – the function we are trying to estimate.

This circularity suggests the following iterative solution. (1) Obtain an initial estimate of the parameters in $\alpha_j(\tau_i^*)$ by regressing $\hat{\alpha}_i$ on $\overline{S}_i^*$. (2) Use the estimated function $\hat{\alpha}_j(\overline{\tau}_i^*)$ to compute estimates of the λ_{ik}, i.e., $\hat{\lambda}_{ik}$.[28] (3) Regress $\hat{\alpha}_i$ on $\hat{\lambda}_{i1}$, $\hat{\lambda}_{i2}$, $\hat{\lambda}_{i3}$, and $\hat{\lambda}_{i4}$ to obtain an updated estimate of $\hat{\alpha}_j(\overline{\tau}_i^*)$. Steps two and three can be repeated until estimates of the polynomial parameters converge (a dozen or so repetitions in our analyses). In this way we estimate how the variance of measurement error from all sources varies across the range of universe scores; $\hat{\sigma}^2_{\eta_j}(\overline{\tau}_i) = \hat{\alpha}_j(\overline{\tau}_i)\,\hat{\sigma}^2_{\eta_{\cdot j}} = \hat{\alpha}_j(\overline{\tau}_i)(1-\hat{G}_j)\,\hat{\omega}_{jj}$.

The solid lines in Figures 6 and 7 are our estimates of $\hat{\sigma}_{\eta_j}(\tau_i)$. The dashed lines show the IRT SEMs reported in the test technical reports for the grade. The shapes of the two curves in

[27] For example, see the discussion of the "structural least squares" polynomial estimator in Kukush et. al (2005).

[28] Extending the approach for computing $E\left(\tau_{ij} \left| S_{ij} \right.\right)$ discussed above,

$\hat{\lambda}_{ik} = E\left(\left(\tau_{ij}\right)^k \left| S_{ij} \right.\right) = \left[k^j\left(S_{ij}\right) M\right]^{-1} \sum_{m=1}^{M} \left(\tau_{mj}^*\right)^k \phi\left(S_{ij} - \tau_{mj}^* \big/ \sigma^2_{\eta_j}(\tau_{mj}^*)\right) \Big/ \sigma_{\eta_j}(\tau_{mj}^*)$.

each graph indicate that our estimates of how the overall measurement-error standard deviations vary over the range of universe scores is very similar to the patterns reported in the technical reports. These results together with those shown in Figure 5 are striking. Our estimates of the overall extent of measurement error, as expected, are systematically higher than those associated with the test instrument. The estimated standard deviation of the overall measure error for each test varies over the range of abilities in a way quite similar to the pattern seen for the reported IRT SEMs. In addition, as shown in Figure 5, the variation in generalizability coefficient estimates across grades mirror across-grade differences in the reliability coefficients and the differences between the math and ELA generalizability coefficient estimates mirror the corresponding differences between the reported math and ELA reliability coefficients. These similarities do not prove the accuracy of our technique for estimating the overall extent of test measurement error, but they do greatly increase our confidence that the technique is able to identify at times subtle differences in the extent of measurement error.

4.3 Inferences Regarding Universe Scores and Universe Score Gains

Observed test scores typically are used to directly estimate students' abilities and ability gains. More precise estimates of universe scores and/or universe-score gains for individual students can be obtained employing the observed scores along with the parameter estimates in Table 2 and the estimated measurement-error heteroskedasticity measured by $\hat{\sigma}_{\eta_j}(\tau_i)$. As an example, the solid S-shaped line in Figure 8(a) shows the values of $\hat{E}\left(\tau_{ij} \mid S_{ij}\right)$ for fifth-grade ELA. Results for grades five and seven math are shown in Figure 9. Results for both subjects in other grades are quite similar. Referencing the 45^{o} straight line, the estimated posterior-mean ability levels for higher-scoring students are substantially below the observed scores while predicted ability levels for low-scoring students are above the observed scores. This Bayes

"shrinkage" is largest for the highest and lowest scores due to the estimated pattern of measurement-error heteroskedasticity. The dashed lines show 80-percent Bayesian credible (confidence) bounds for ability conditional on the observed score. For example, the BLUP of the universe-score for fifth-grade students scoring 775 in ELA is 731, 44 point below the observed score. We estimate that 80 percent of students scoring 775 have universe scores in the range 717-747; $P\left(717.1<\tau_{ij}<747.2\left|S_{ij}=775\right.\right)=0.80$. In this case, the observed score is 28 points higher than the upper bound of the 80-percent credible interval. Midrange scores are somewhat more informative, reflecting the smaller standard deviation of test measurement error. For an observed score of 650, the estimated posterior mean and 80 percent Bayesian confidence interval are 652 and (639,665), respectively. The credible bounds for a 775 score is 15 percent larger than that for a score of 650.

As Figures 8 and 9 make clear, utilizing test scores to directly estimate students' abilities is problematic for high- and, to a lesser extent, low-scoring students. To explore further, consider the root of the expected mean squared errors (RMSE) associated with estimating student ability using (i) observed scores and (ii) estimated posterior mean abilities conditional on observed scores.[29] In the case of the fifth-grade math exam shown in Figure 9(a), the RMSE associated with using $\hat{E}\left(\tau_{ij}\left|S_{ij}\right.\right)$ to estimate students' abilities is 15.5 scale-score points. In contrast, the RMSE associated with using S_{ij} is 18.4, 19 percent larger. The magnitude of this difference is meaningful given that $\hat{E}\left(\tau_{ij}\left|S_{ij}\right.\right)$ differs little from S_{ij} over the range of scores for which there are relatively more students. Over the range of actual abilities between 620 and 710 in Figure

[29] The expected values are computed using Monte Carlo simulation and assuming our parameter estimates are correct.

9(a), the RMSE for $\hat{E}\left(\tau_{ij}\left|S_{ij}\right.\right)$ and S_{ij} are 15.5 and 15.6, respectively. For ability levels below 620 the RMSEs are 16.8 and 21.8, respectively, the latter being 30 percent larger. For students whose ability levels are greater than 710, the RMSE of employing $\hat{E}\left(\tau_{ij}\left|S_{ij}\right.\right)$ to estimate τ_{ij} is 14.4, smaller than the overall RMSE for this estimator. In contrast, the RMSE associated with using S_{ij} to estimate τ_{ij} is 31.3 for students whose actual abilities are greater than 710 -- over twice as large as the corresponding RMSE for $\hat{E}\left(\tau_{ij}\left|S_{ij}\right.\right)$. By estimating the overall extent and pattern of test measurement error from all sources, it is possible to compute estimates of universe scores that have statistical properties superior to those corresponding to merely using the observed scores of students as estimates of their ability levels.

Turning to the measurement of ability gains, the solid S-shaped curve in Figure 10 shows the posterior-mean universe-score change in math between grades five and six conditional on the observed score change. Again, the dashed lines show 80-percent credible bounds. For example, among students observed to have a 40-point score increase between the fifth and sixth grades, their actual universe score changes are estimated to average 13.5. Eighty-percent of all students having a 40-point score increase are estimated to have actual universe score changes falling in the interval -1.1 to 27.4. It is noteworthy that for the full range of score change shown (± 50 points), the 80-percent credible bounds include there actually being no change in ability.

Note that there are many different combinations of scores that yield a given change in observed scores; a score increase from 590 to 630 implies a 40-point change as does an increase from 710 to 750. Figure 10 corresponds to the case where one knows the score change but not

the pre- and post-scores.[30] However, for a given score change, the mean universe-score change and credible bounds will vary across known score levels because of the pattern of measurement error heteroskedasticity. For example, Figure 11 shows the posterior-mean universe score change and credible bounds conditional on particular combinations of scores that correspond to a 40-point increase. For example, students scoring 710 on the grade-five exam and 750 on the grade-six exam are estimated to have a 10.2 point universe-score increase on average, with 80 percent of such students having actual changes in ability in the interval (-12.4, 31.0). (Note that a 40 point score increase is relatively large in that the standard deviation of the score change between the fifth- and sixth-grades is 26.0.) For students having a 40-point score increase, there actually being no change in ability falls outside the credible bounds only when the score in grade five is approximately between 615 and 658.

A striking result in Figure 11 is that the posterior mean universe-score change, $\hat{E}\left(\tau_6-\tau_5\left|S_5,S_6\right.\right)=\hat{E}\left(\tau_6\left|S_5,S_6\right.\right)-\hat{E}\left(\tau_5\left|S_5,S_6\right.\right)$, is substantially smaller than the corresponding observed-score change, warranting further explanation. Again consider $\hat{E}\left(\tau_6-\tau_5\left|S_5=710,S_6=750\right.\right)=10.3$. Figure 12 illustrates why this is substantially smaller in

[30] The joint density of $\tau_{ij},\tau_{i,j+1},\eta_{ij}$, and $\eta_{i,j+1}$ is $h^j\left(\tau_{ij},\tau_{i,j+1},\eta_{ij},\eta_{i,j+1}\right)=g^j\left(\eta_{ij}\left|\tau_{ij}\right.\right)g^{j+1}\left(\eta_{i,j+1}\left|\tau_{i,j+1}\right.\right)f\left(\tau_{ij},\tau_{i,j+1}\right)$. With $\delta=\tau_{j+1}-\tau_j$ and $D=S_{j+1}-S_j=\delta+\eta_{j+1}-\eta_j$, the joint density of $\tau_{ij},\delta,\eta_{ij}$, and D is $h^j\left(\tau_{ij},\tau_{ij}+\delta,\eta_{ij},D-\delta+\eta_{ij}\right)$. Integrating over τ_{ij} and η_{ij} yields the joint density of δ and D; $z\left(\delta,D\right)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}g^{j+1}\left(D-\delta+\eta_{ij}\left|\tau_{i,j+1}\right.\right)f^2\left(\tau_{ij}+\delta\left|\tau_{ij}\right.\right)g^j\left(\eta_{ij}\left|\tau_{ij}\right.\right)f^1\left(\tau_{ij}\right)d\eta_{ij}\,d\tau_{ij}$ where $f^1\left(\tau_{ij}\right)$ is the marginal density of τ_{ij} and $f^2\left(\tau_{i,j+1}\left|\tau_{ij}\right.\right)$ is the conditional density of $\tau_{i,j+1}$ given τ_{ij}. This integral can be computed using $z\left(\delta,D\right)=\left(1/J\right)\sum_{j=1}^{J}g^{j+1}\left(D-\delta+\eta_{ij}^*\left|\tau_{ij}^*+\delta\right.\right)f^2\left(\tau_{ij}^*+\delta\left|\tau_{ij}^*\right.\right)$ where $\left(\tau_{ij}^*,\eta_{ij}^*\right),\ j=1,2,\cdots,J,$ is a sufficiently large number of draws from the joint distribution of $\left(\tau_{ij},\eta_{ij}\right)$. In turn, the density of the posterior distribution of δ given D is $z\left(\delta\left|D\right.\right)=z\left(\delta,D\right)/l\left(D\right)$ where $l\left(D\right)=\left(1/J\right)\sum_{j=1}^{J}g^{j+1}\left(D-\tau_{i,j+1}^*+\tau_{ij}^*+\eta_{ij}^*\left|\tau_{i,j+1}^*\right.\right)$ is the density of D. The cumulative posterior distribution is $P\left(\delta\le a\left|S\right.\right)=\left(1/J\,l(D)\right)\sum_{\tau_{i,j+1}^*-\tau_{ij}^*\le a}g^{j+1}\left(D-\tau_{i,j+1}^*+\tau_{ij}^*+\eta_{ij}^*\left|\tau_{i,j+1}^*\right.\right)$. Finally, the posterior mean ability given D is $E\left(\delta\left|D\right.\right)=\left(1/J\,l(D)\right)\sum_{j=1}^{J}\left(\tau_{i,j+1}^*-\tau_{ij}^*\right)\,g^{j+1}\left(D-\tau_{i,j+1}^*+\tau_{ij}^*+\eta_{ij}^*\left|\tau_{i,j+1}^*\right.\right)$.

magnitude compared to the 40-point increase in score. First, $\hat{E}(\tau_6|S_6 = 750) = 733.0$ is 17 points below the observed score due to the Bayes shrinkage toward the mean $\mu_6 = 667.8$.

$\hat{E}\left(\tau_6|S_5 = 710, S_6 = 750\right) = 729.9$ is even smaller; because S_6 is a noisy estimate of τ_6 and τ_5 is correlated with τ_6, the value of S_5 provides information regarding the distribution of τ_6 that goes beyond the information gained by observing S_6. Note that $E\left(\tau_6|S_5, S_6\right)$ would equal $E(\tau_6|S_6)$ if either $\sigma^2_{\eta_6} = 0$ or $\rho_{56} = 0$. $\hat{E}\left(\tau_6|S_5, S_6\right)$ is less than $\hat{E}(\tau_6|S_6)$ because S_5 is substantially below S_6. Similar logic holds for the fifth grade. $\hat{E}(\tau_5|S_5 = 710) = 707.5$ is less than 710 because the latter is substantially above μ_5. However, $\hat{E}(\tau_5|S_5, S_6) = 719.6$ is meaningfully larger than $\hat{E}(\tau_5|S_5) = 707.5$ and larger than $S_5 = 710$, reflecting that $S_6 = 750$ is substantially larger than S_5. In summary, among New York City students scoring 710 on the fifth-grade math exam and 40 points higher on the sixth grade exam, we estimate the mean gain in ability is little more than one-forth as large as the actual score change; $\hat{E}\left(\tau_6|S_5, S_6\right) - \hat{E}\left(\tau_5|S_5, S_6\right) = 729.9 - 719.6 = 10.3$. The importance of accounting for the estimated correlation between ability levels in grades five and six is reflected in the fact that the mean ability increase would be two and one-half times large were the ability levels uncorrelated, $\hat{E}\left(\tau_6|S_6\right) - \hat{E}\left(\tau_5|S_5\right) = 733.0 - 707.5 = 25.5$.

5.0 Conclusion

In this paper we show that there is a credible and feasible approach for estimating the total extent of test measurement error utilizing estimates of the empirical correlation or covariance matrix for three or more interval-scaled tests. The scales can differ across the tests

provided that they are linear transformations of an underlying common vertical scale that need not be known. Our approach maintains relatively unrestrictive assumptions regarding the structure of student achievement growth. We assume that academic achievement is cumulative following a first-order autoregressive process; $\tau_{ij} = \beta_{j-1}\tau_{i,j-1} + \theta_{ij}$ where there is at least some persistence (i.e., $\beta_{j-1} > 0$) and the possibility of decay ($\beta_{j-1} < 1$), which can differ across grades. Even though derivations would be more complicated, one could employ some other structure (e.g., a second-order autoregressive process). With $\tau_{ij} = \beta_{j-1}\tau_{i,j-1} + \theta_{ij}$, an additional assumption is needed regarding the stochastic properties of θ_{ij}. A reduced-form specification is employed in the paper and, to illustrate the generality of the approach, three examples of fully specified structural models are outlined in Appendix A,

Our approach is a meaningful generalization of the test-retest method, providing a useful, more generally applicable tool for estimating the extent of test measurement error from all sources. Estimation is straightforward and the key assumptions underlying the empirical model (i.e., $\tau_{i,j+1} = \beta_j\tau_{ij} + \theta_{i,j+1}$ with $\beta_j \neq 0$ and $E\left(\theta_{i,j+1}\middle|\tau_{ij}\right)$ is a linear function of τ_{ij}) appear to be quite reasonable. Furthermore, these assumptions imply that $\tau_{i,j+1}$ is a linear function of τ_{ij} which can be tested.

Estimation of the overall extent of measurement error for a population of students only requires computed test-score descriptive statistics and correlations. However, when student-level test-score data are available, one can explore the extent and pattern of measurement error heteroskedasticity. Results for New York make clear that heteroskedasticity can be important in that variation in the extent of test measurement error across ability levels is quite large.

The overall extent of test measurement error can be estimated without specifying particular functional forms for the distribution of either abilities or test measurement error. However, by maintaining standard assumptions (e.g., normality), one can make inferences regarding universe scores and universe score gains. In particular, for a student with a given score, the Bayesian posterior mean and variance of τ_{ij} given S_{ij}, $E\left(\tau_{ij}\left|S_{ij}\right.\right)$ and $V\left(\tau_{ij}\left|S_{ij}\right.\right)$ are easily computed where the former is the best linear unbiased predictor (BLUP) of the student's actual ability. Similar statistics for test score gains can also be computed. We show that using the observed score as an estimate of a student's underlying ability can be quite misleading for relatively low- or high-ability students. However, this is not the case when the posterior mean is employed.

Estimates of the overall extent of test measurement error have a variety of uses that go beyond merely assessing the reliability of various assessments. Employing $E\left(\tau_{ij}\left|S_{ij}\right.\right)$, rather than S_{ij}, to estimate τ_{ij} is one example. Another is the computation of effect sizes where the magnitudes of the effects of different causal factors can be judged relative to either the standard deviation of ability or the standard deviation of ability gains. Bloom et al. (2008) discuss the desirability of taking into account the dispersion of ability or ability gains rather than test scores or test-score gains but note that analysts often have little if any information regarding the extent of test measurement error.

As demonstrated above, the same types of data researchers often employ to estimate how various factors affect educational outcomes can be used to estimate the overall extent of test measurement error. Based on the variance estimates shown in columns (1) and (3) of Table 5, for the tests we analyze effect sizes measures relative to the standard deviation of ability will be ten to 18 percent larger than effect sizes measured relative to the standard deviation of test scores. In

cases where it is pertinent to judge the magnitudes of effects in terms of achievement gains, effect sizes measured relative to the standard deviation of ability gains will be two to over three times larger compared to those measured relative to the standard deviation of test-score gains.

Another use of the estimated extent of test measurement error pertains to the common practice of entering student test scores as right-hand-side variables in regression equations, as is often done in value-added modeling. A concern is that the measurement error associated with the prior tests can bias other coefficient estimates. However, any such problem can be avoided by employing $E(\tau_{ij} \mid S_{ij})$ rather than S_{ij} as a proxy for ability in the regression.[31] Doing so has a second benefit. Even if the dependent variable in such a regression is a linear function of ability, a problem of nonlinearity is introduced when S_{ij} is used to proxy ability. This problem arises when $E(\tau_{ij} \mid S_{ij})$ is a nonlinear function of S_{ij}, as we demonstrate in our application. In an effort to deal with this issue, one could include the square and cube of S_{ij} in the regression. However, the problem can be avoided by employing $\hat{E}(\tau_{ij} \mid S_{ij})$ rather than S_{ij} in the regression.

Finally, by estimating the extent and pattern of test measurement error one can assess the precision of a variety of measures that are computed based upon test scores. These include indicators of student proficiency (e.g., AYP), teacher- and school-effect estimates and accountability measures more generally. As we have shown, it is possible to measure the reliability of such measures as well as employ the estimated extent of test measurement error to calculate more accurate measures, information which should be employed in policy applications based on student achievement tests.

Overall, this paper has both methodological and substantive implications. Methodologically it shows that the full extent of test measurement error can be estimated without

[31] See Sullivan (2001).

employing the costly test-retest strategy. Substantively, it shows that the overall measurement error is substantially greater than reported split-test measurement error and this difference suggests that much empirical work has been underestimating the effect sizes of interventions (e.g., programs or teachers) that affect student learning.

References

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of Labor Economics* 25(1): 95–135.

Abowd, J.M. and D. Card (1989) "On the Covariance Structure of Earnings and Hours Changes," Econometrica 57(2), 411-445.

Altonji, J.G. and L.M. Segal (1996) "Small Sample Bias in GMM Estimation of Covariance Structures," *Journal of Business and Economic Statistics* 14, 353-366.

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

Baltagi, B. H. (2005) *Econometric Analysis of Panel Data*, West Sussex, England: John Wiley & Sons, Ltd.

Bloom, H.S., C.J. Hill, A.R. Black and M.W. Lipsey (2008) "Performance Trajectories and Performance Gaps as Achievement Effect-Size Benchmarks for Educational Interventions," Journal of Research on Educational Effectiveness (1) 289-328.

Brennan, R. L. (2001) *Generalizability Theory*, New York: Sprnger-Verlag.

Cameron, A.C. and P.K. Trivedi (2005) *Microeconometrics: Methods and Applications*, New York: Cambridge University Press.

Carlin, B. P. and T. A. Louis (1996) *Bayes and Empirical Bayes Methods for Data Analysis*, Boca Raton: Chapman & Hall/CRC.

Conbach, L.J., R.L. Linn, R.L. Brennan and E.H. Haertel (1997) "Generalizability Analysis for Performance Assessments of Student Achievement or School Effectiveness," Educational and Psychological Measurement, 57(3), 373-399.

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

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

Feldt, L. S. and R. L. Brennan (1989) "Reliability," in *Educational Measurement* 3rd ed., New York: American Council on Education.

Goldhaber, D. and E. Anthony (2007) "Can Teacher Quality Be Effectively Assessed? National Board Certification as Signal of Effective Teaching," *The Review of Economics and Statistics* 89(1), 134-150.

Goldhaber, D., and M. Hansen. 2010 "Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions." CALDER working paper.

Haertel, E. H. (2006) "Reliability," in *Educational Measurement*, fourth edition, R. L. Brennan, ed., Praeger.

Hill, C., H. Bloom, A. Black, M. Lipsey, "Empirical Benchmarks for Interpreting Effect Sizes in Research" MDRC Working Paper, July 2007.

Koedel, Cory, and Julian R. Betts. 2007. "Re-Examining the Role of Teacher Quality in the Educational Production Function." Working Paper 0708. University of Missouri, Department of Economics.

Kukush, A., H. Schneesweiss and R. Wolf (2005) "Relative Efficiency of Three Estimators in a Polynomial Regression with Measurement Errors," *Journal of Statistical Planning and Inference* 127, 179-203.

Lee, W. C., R. L. Brennan and M. J. Kolen (2000) "Estimators of Conditional Scale-Score Standard Errors of Measurement: A Simulation Study," *Journal of Educational Measurement* 37(1), 1-20.

McCaffrey, D. F., J. R. Lockwood, D. Koretz, T. A. Louis and L Hamilton (2004) "Models for Value-Added Modeling of Teacher Effects," *Journal of Educational and Behavioral Statistics* 29(1), 67-101.

McCaffrey, Daniel F., Tim R. Sass, J. R. Lockwood, and Kata Mihaly. 2009. "The Intertemporal Variability of Teacher Effect Estimates." *Education Finance and Policy* 4(4):572–606.

Rogosa, D.R. and J. B. Willett (1983) "Demonstrating the Reliability of Difference Scores in the Measurement of Change," *Journal of Educational Measurement* 20(4) 335-343.

Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys*, New York: J. Wiley & Sons.

Sanders, W. and J. Rivers (1996) "Cumulative and Residual Effects of Teachers on Future Student Academic Achievement," working paper, University of Tennessee Value-Added Research and Assessment Center.

Sanford, E. E. (1996) "North Carolina End-of-Grade Tests: Reading Comprehension, Mathematics," Technical Report #1. Division of Accountability/Testing, Office of Instruction and Accountability Services, North Carolina Department of Public Instruction.

Schafer, J. L. (1997) *Analysis of Incomplete Multivariate Data*, London: Chapman & Hall.

Shen, W. and T. A. Louis (1998) "Triple-goal Estimates in Two-Stage Hierarchical Models," *Journal of the Royal Statistical Society* 60(2), 455-471.

Sullivan, D. G. (2001) "A Note on the Estimation of Linear Regression Models with Heteroskedastic Measurement Errors," Federal Reserve Bank of Chicago working paper WP 2001-23.

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

Todd, P.E. and K.I. Wolpin (2003) "On the Specification and Estimation of the Production Function for Cognitive Achievement," *The Economic Journal* 113, F3-F33.

Wright, S. P. and W. L. Sanders ( ) "Decomposition of Estimates in a Layered Value-Added Assessment Model," Value-Added Assessment

Table 1
Descriptive Statistics for Cohort

	ELA		Math	
	mean	standard deviation	mean	standard deviation
Grade 3	626.8	37.3	616.5	42.3
Grade 4	657.9	39.0	665.8	36.0
Grade 5	659.3	36.1	665.7	37.5
Grade 6	658.0	28.8	667.8	37.5
Grade 7	661.7	24.4	671.0	32.5
Grade 8	660.5	26.0	672.2	31.9
	N = 67,528		N = 74,700	

Table 2 Correlation and Generalizability Coefficient Estimates, New York City

Parameters+	Math	ELA
ρ^*_{34}	0.8144	0.8369
	(0.0016)	(0.0016)
ρ_{45}	0.9581	0.9785
	(0.0012)	(0.0013)
ρ_{56}	0.9331	0.9644
	(0.0011)	(0.0012)
ρ_{67}	0.9647	0.9817
	(0.0011)	(0.0012)
ρ^*_{78}	0.8711	0.8168
	(0.0013)	(0.0013)
G_4	0.8005	0.7853
	(0.0024)	(0.0025)
G_5	0.8057	0.7169
	(0.0020)	(0.0018)
G_6	0.8227	0.7716
	(0.0019)	(0.0019)
G_7	0.8284	0.7184
	(0.0020)	(0.0019)

+ The parameter subscripts here correspond to the grade tested. For example, ρ^*_{34} is the correlation of universe scores of students in grades three and four

Table 3 Correlations of Scores on the NYS ELA Examinations in Grades Three Through Eight (Computed values below the diagonal and fitted-values above)

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Grade 3		0.7416	0.6934	0.6937	0.6571	0.6332
Grade 4	0.7416		0.7342	0.7346	0.6958	0.6705
Grade 5	0.6949	0.7328		0.7173	0.6794	0.6548
Grade 6	0.6899	0.7357	0.7198		0.7309	0.7044
Grade 7	0.6573	0.6958	0.6800	0.7303		0.6923
Grade 8	0.6356	0.6709	0.6514	0.7050	0.6923	

Table 4 Correlations of Scores on the NYS Math Examinations in Grades Three Through Eight (Computed values below the diagonal and fitted-values above)

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Grade 3		0.7286	0.7003	0.6603	0.6393	0.6119
Grade 4	0.7286		0.7694	0.7254	0.7023	0.6722
Grade 5	0.6936	0.7755		0.7597	0.7355	0.7039
Grade 6	0.6616	0.7248	0.7592		0.7964	0.7623
Grade 7	0.6480	0.6998	0.7323	0.7944		0.7929
Grade 8	0.6091	0.6685	0.7077	0.7643	0.7929	

Table 5: Variances of Test Scores, Test Measurement Error, Universe Scores, Test-Score Gains, Measurement Error for Gains, and Universe Score Gains and Generalizabiltity Coefficient for Test-Score Gain, ELA and Math

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ELA	$\sigma^2_{S_{\bullet j}}$	$\hat{\sigma}^2_{\eta_{\bullet j}}$	$\hat{\gamma}_{jj} = \hat{G}_j \sigma^2_{S_{\bullet j}}$	$\hat{\sigma}^2_{\Delta S_{\bullet j}}$	$\hat{\sigma}^2_{\Delta \eta_{\bullet j}}$	$\hat{\sigma}^2_{\Delta \tau_{\bullet j}}$	$\hat{G}^{\Delta}_j = \hat{\sigma}^2_{\Delta \tau_{\bullet j}} \Big/ \hat{\sigma}^2_{\Delta S_{\bullet j}}$
grade 7	1520.8	326.5	1194.3	763.8	695.3	68.4	0.090
grade 6	1303.0	368.8	934.2	646.2	558.9	87.3	0.135
grade 5	832.1	190.0	642.1	407.4	357.6	49.8	0.122
grade 4	595.1	167.6	427.5				
Math							
grade 7	1297.6	259.0	1038.6	661.9	532.8	129.1	0.195
grade 6	1409.5	273.8	1135.7	677.9	523.8	154.1	0.227
grade 5	1409.5	250.0	1159.5	527.8	431.0	96.8	0.183
grade 4	1054.9	181.0	873.9				

Figure 1: Nonparametric Regression of Grade 8 ELA Scores on Scores in Grade 7, Bubble Graph Showing the Joint Distribution of Scores and Standard-Error of Measurement for 7th Grade Scores

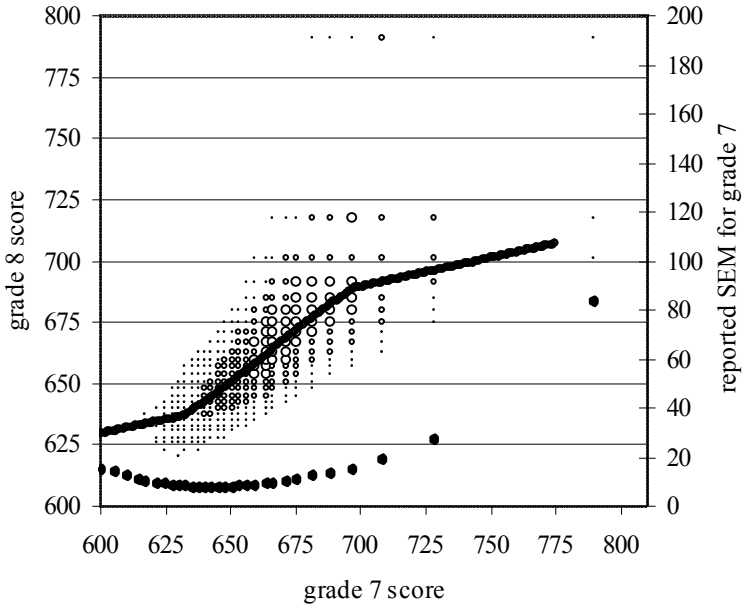


Figure 2: Nonparametric Regression of Grade 8 Math Scores on Scores in Grade 7, Bubble Graph Showing the Joint Distribution of Scores and Standard-Error of Measurement for 7th Grade Scores

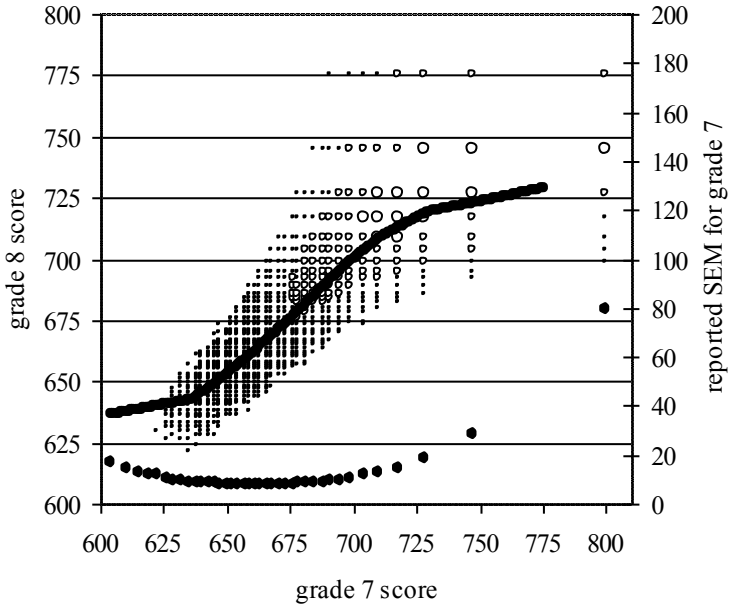


Figure 3: Cubic Regression Estimates of $E\left(S_{i,j+1} \mid S_{ij}\right)$ as well as consistent estimates of cubic and linear specifications of $E\left(\tau_{i,j+1} \mid \tau_{ij}\right)$, ELA

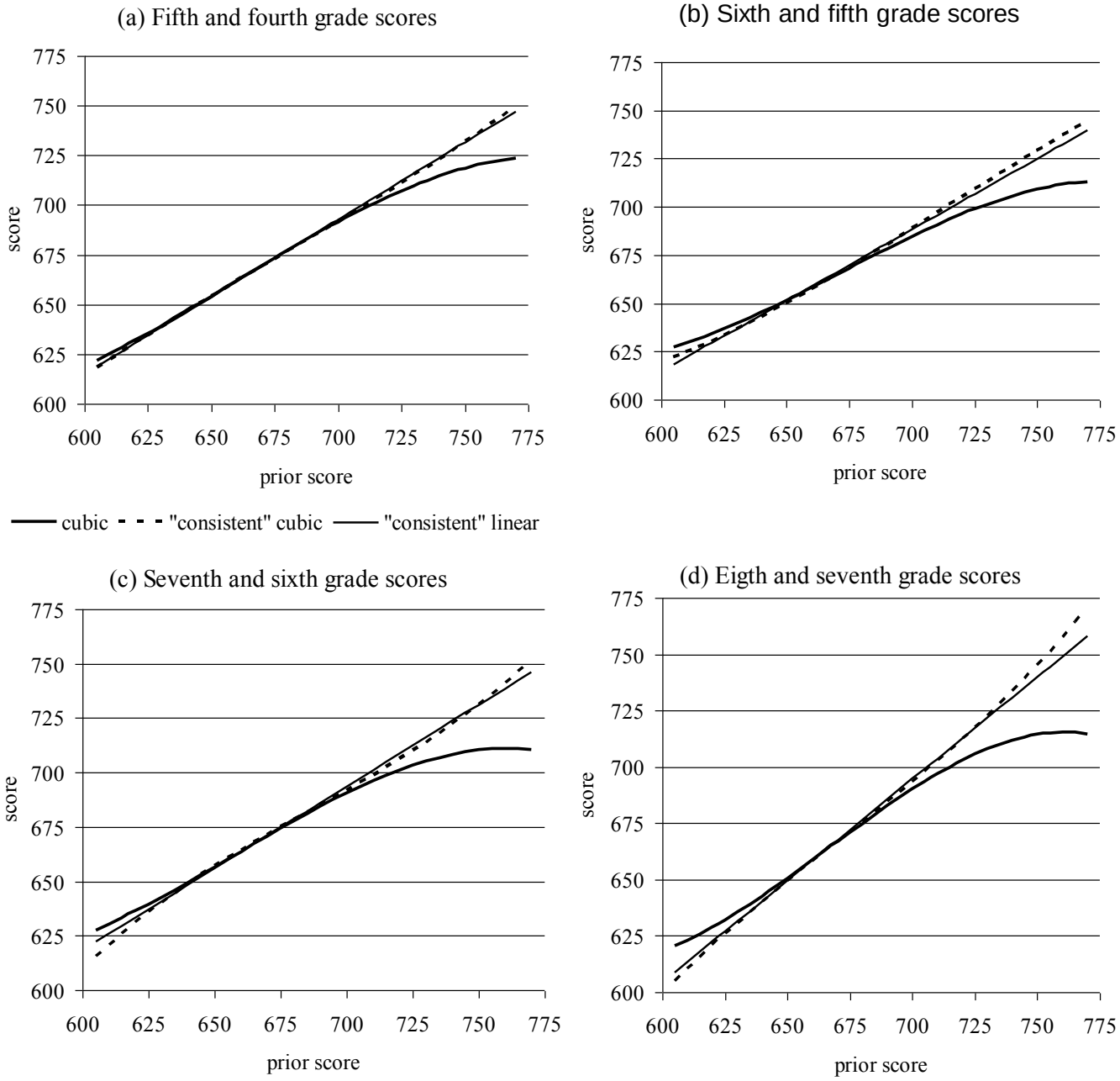


Figure 4: Cubic Regression Estimates of $E\left(S_{i,j+1} \mid S_{ij}\right)$ as well as consistent estimates of cubic and linear specifications of $E\left(\tau_{i,j+1} \mid \tau_{ij}\right)$, Math

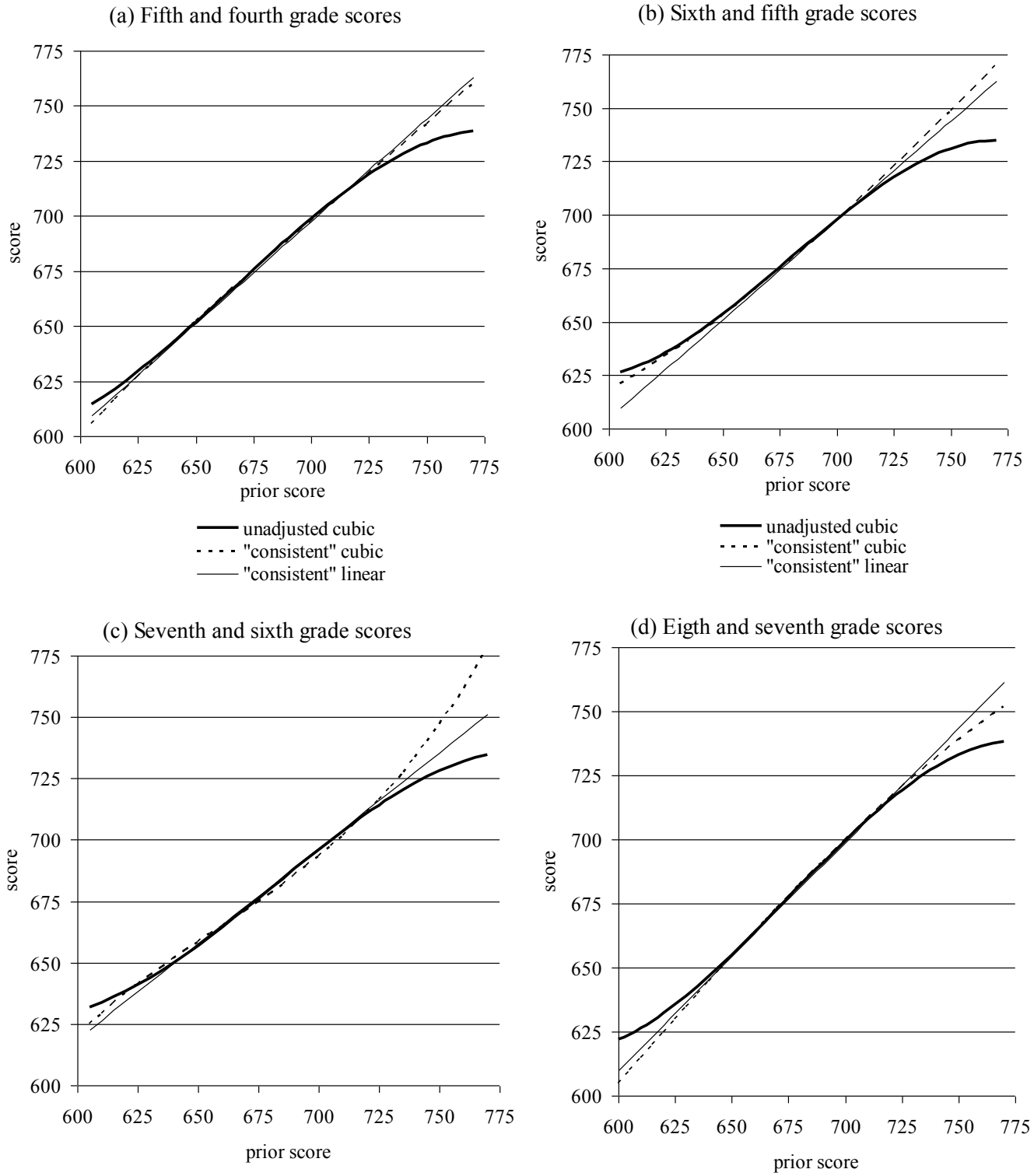


Figure 5: Generalizability and Reliability Coefficient Estimates for New York Math and ELA Exams by Grade

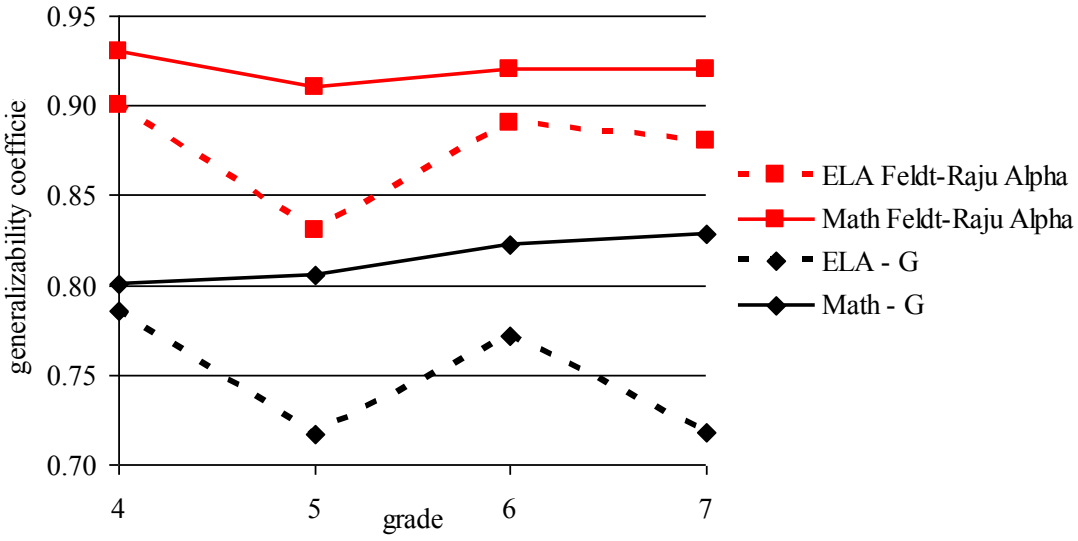


Figure 6: Standard Errors of Measurement Reported in Technical Reports (dashed lines) and Estimated Using the Reduced-Form Model, ELA by Grade Pairs

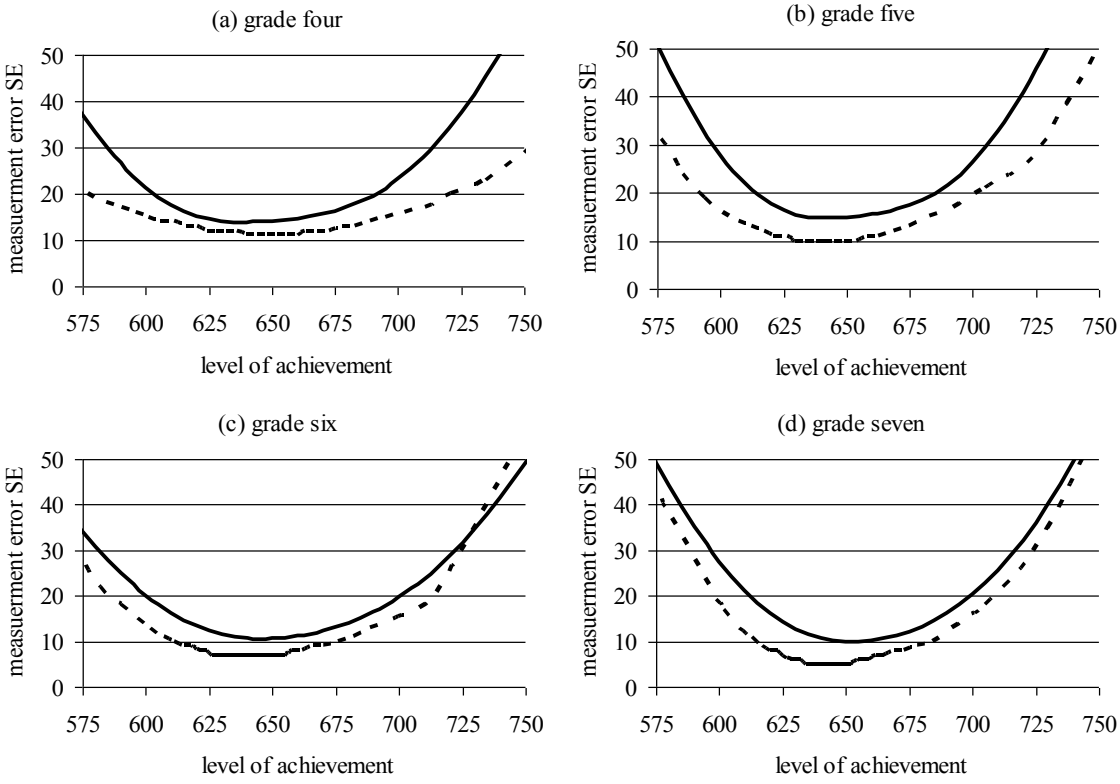


Figure 7: Standard Errors of Measurement Reported in Technical Reports (dashed lines) and Estimated Using the Reduced-Form Model, Mathematics by Grade Pairs

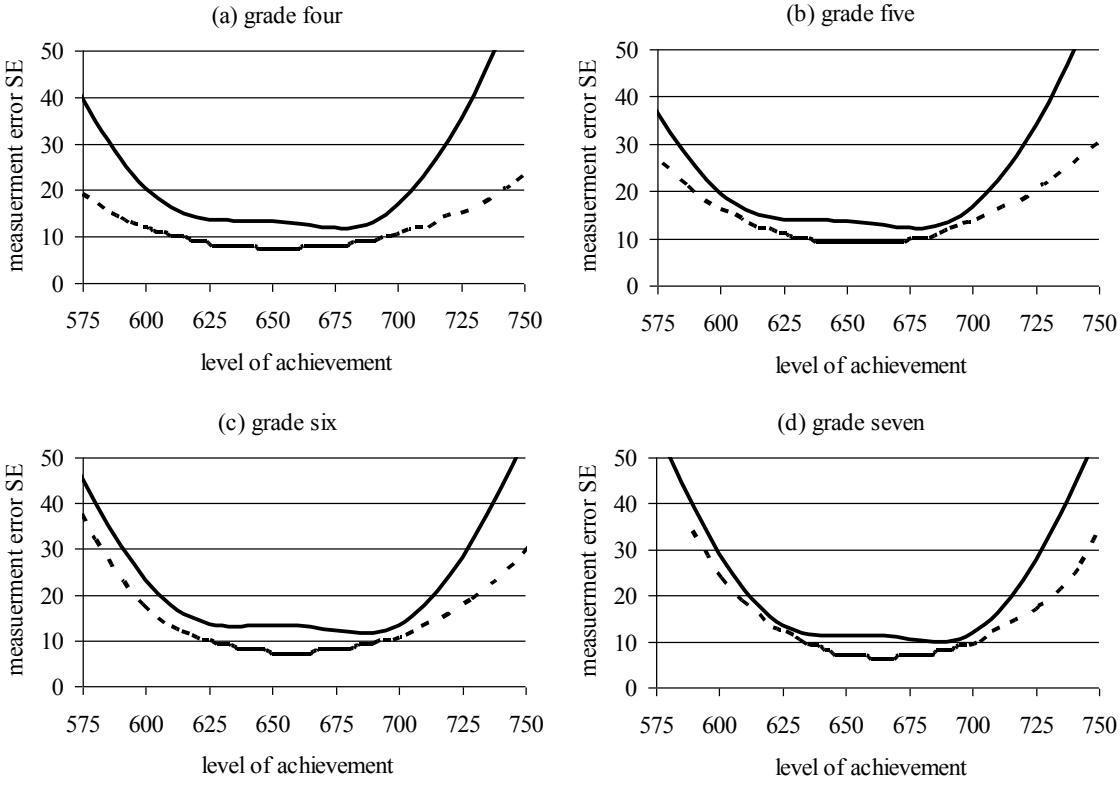


Figure 8
Estimated Posterior Mean Ability Level Given the Observed Score
and 80-Percent Bayesian Confidence Bounds, Grades 5 and 7 ELA

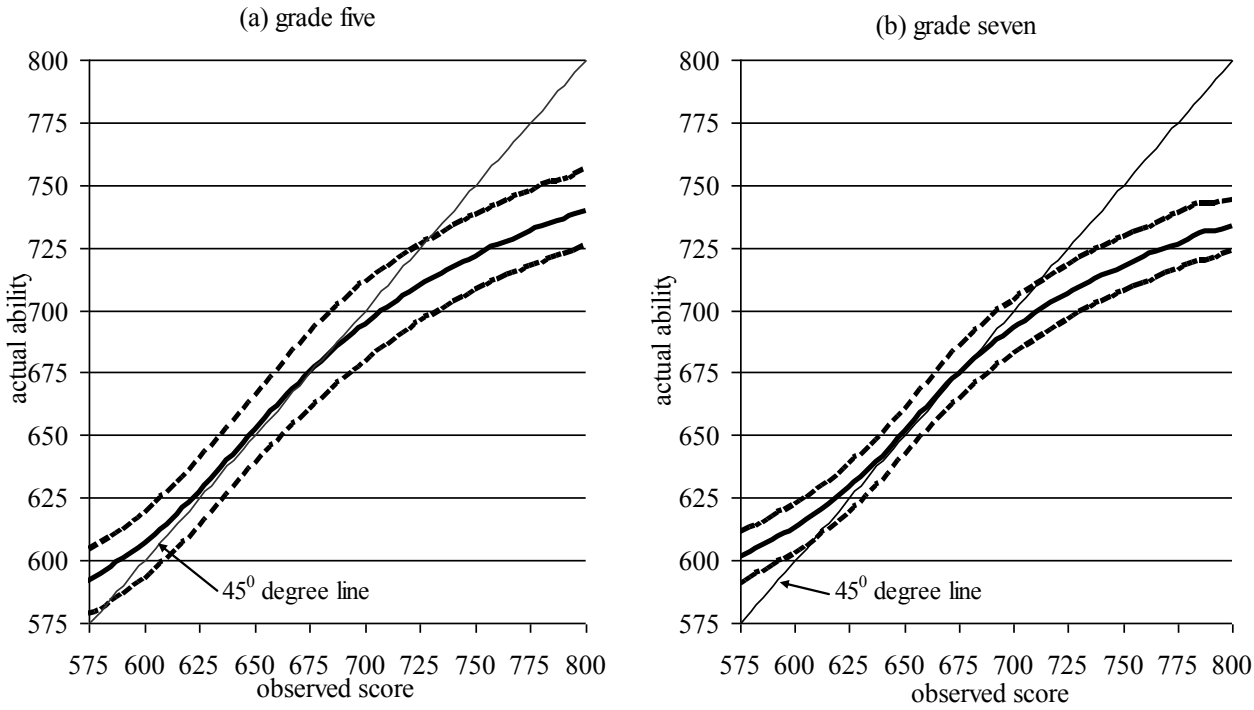


Figure 9
Estimated Posterior Mean Ability Level Given the Observed
Score and 80-Percent Bayesian Confidence Bounds, Grades 5 and 7 Math

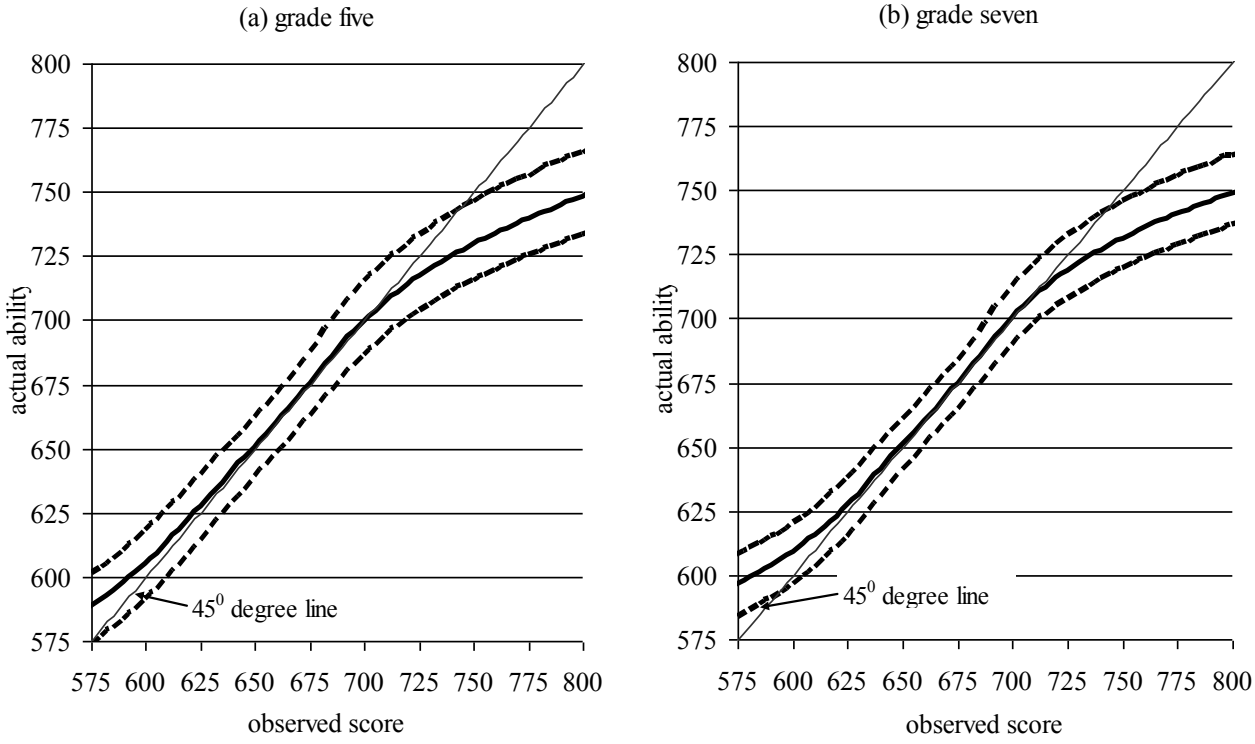


Figure 10
Estimated Posterior Mean Change in Ability Given the Change in Observed Score and 80-Percent Credible Bounds, Grades 5 and 6 Mathematics

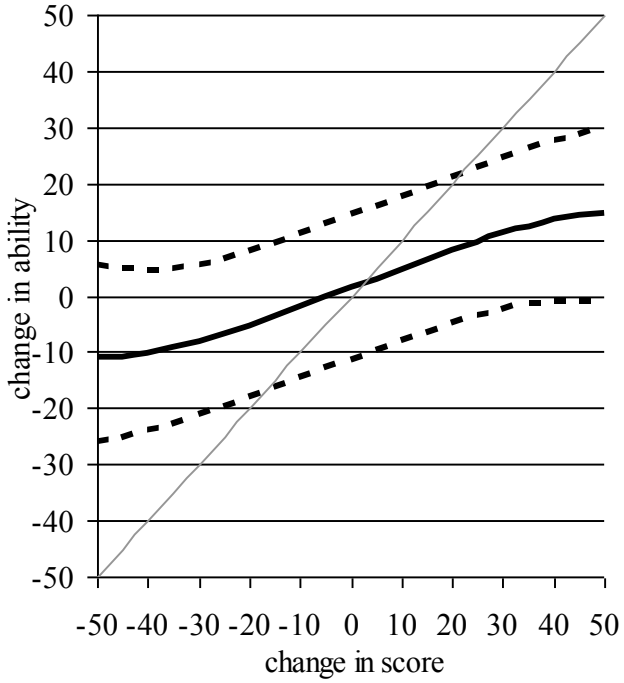


Figure 11
Estimated Posterior Mean Change in Ability for the Observed Scores in Grades Five and Six Mathematics for $S_6 - S_5 = 40$ and 80-Percent Credible Bounds

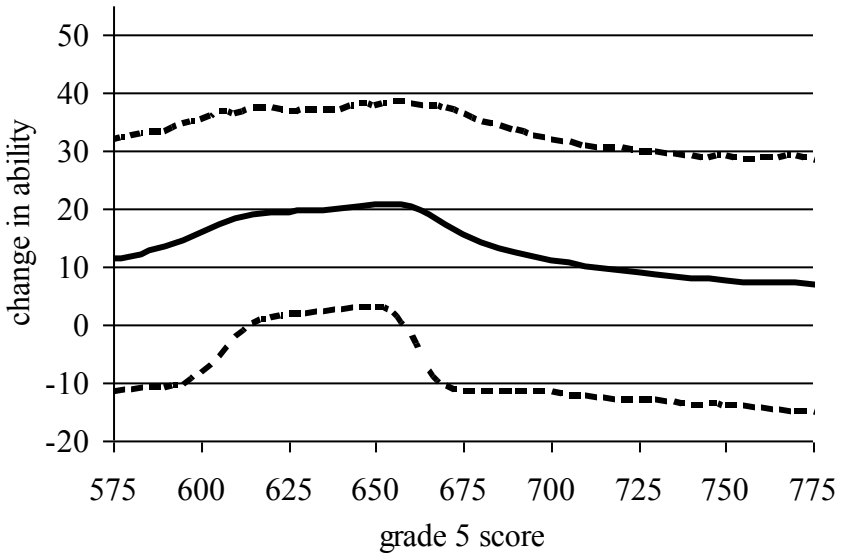
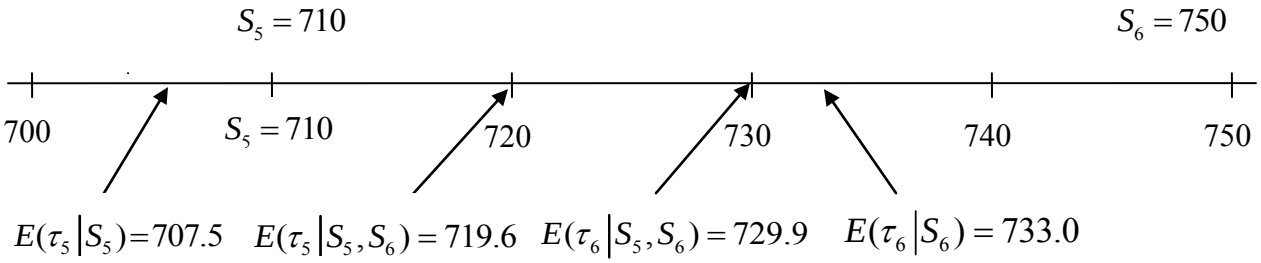


Figure 12
Example Showing Posterior means for a Forty-Point Score Increase, Grades 5 & 6 Mathematics



APPENDIX A

We illustrate the structural approach for estimating the overall extent of test measurement error using three alternative specifications for θ_{ij} in $\tau_{ij} = \beta_{j-1}\tau_{i,j-1} + \theta_{ij}$, each of which fully specifies the covariance structure of achievement gains across grades.

Before considering particular specifications, note that for any specification of θ_{ij},

$\gamma_{12} \equiv Cov(\tau_{i1}, \tau_{i2}) = Cov(\tau_{i1}, \beta_1\tau_{i1} + \theta_{i2}) = \beta_1\gamma_{11} + Cov(\tau_{i1}, \theta_{i2}) = \beta_1\gamma_{11} + \psi_{12}$ and, in general,

$\gamma_{jk} = Cov(\tau_{ij}, \tau_{ik}) = Cov(\tau_{ij}, \beta_{k-1}\tau_{i,k-1} + \theta_{ik}) = \beta_{k-1}\gamma_{j,k-1} + \psi_{jk}$, for $k > j$, where $\psi_{jk} = Cov(\tau_{ij}, \theta_{ik})$.

These recursive equations imply the structure of $\Omega_{\bullet}$ shown in Equation A1 and the moment conditions in Equation A2. Equation A3 also holds (e.g., $\gamma_{22} = \beta_1^2\gamma_{11} + 2\beta_1\psi_{12} + \sigma^2_{\theta_2}$).

$$\Omega_{\bullet} = \begin{bmatrix} \omega_{11} & \omega_{12} & \omega_{13} & \omega_{14} & \cdots \\ & \omega_{22} & \omega_{23} & \omega_{24} & \cdots \\ & & \omega_{33} & \omega_{34} & \cdots \\ & & & \omega_{44} & \cdots \\ & & & & \ddots \end{bmatrix} = \begin{bmatrix} \gamma_{11}/G_1 & \beta_1\gamma_{11} + \psi_{12} & \beta_2\gamma_{12} + \psi_{13} & \beta_3\gamma_{13} + \psi_{14} & \cdots \\ & \gamma_{22}/G_2 & \beta_2\gamma_{22} + \psi_{23} & \beta_3\gamma_{23} + \psi_{24} & \cdots \\ & & \gamma_{33}/G_3 & \beta_3\gamma_{33} + \psi_{34} & \cdots \\ & & & \gamma_{44}/G_4 & \cdots \\ & & & & \ddots \end{bmatrix} \quad \text{(A1)}$$

$$\begin{aligned}
&\omega_{11} = \gamma_{11}/G_1 \\
&\omega_{12} = \beta_1\gamma_{11} - \psi_{12} \\
&\omega_{13} = \beta_2\beta_1\gamma_{11} - \beta_2\psi_{12} - \psi_{13} \\
&\omega_{14} = \beta_3\beta_2\beta_1\gamma_{11} - \beta_3\beta_2\psi_{12} - \beta_3\psi_{13} - \psi_{14} \\
&\cdots \\
&\omega_{22} = \gamma_{22}/G_2 \\
&\omega_{23} = \beta_2\gamma_{22} - \psi_{23} \\
&\omega_{24} = \beta_3\beta_2\gamma_{22} - \beta_3\psi_{23} - \psi_{24} \\
&\cdots \\
&\omega_{33} = \gamma_{33}/G_3 \\
&\omega_{34} = \beta_3\gamma_{33} - \psi_{34} \\
&\cdots \\
&\omega_{44} = \gamma_{44}/G_4 \\
&\cdots
\end{aligned} \quad \text{(A2)}$$

$$\gamma_{jj} = V(\tau_{ij}) = V(\beta_{j-1}\tau_{i,j-1} + \theta_{ij}) = \beta^2_{j-1}\gamma_{j-1,j-1} + 2\beta_{j-1}\psi_{j-1,j} + \sigma^2_{\theta_j}. \quad \text{(A3)}$$

This covariance structure follows from only assuming $\tau_{ij} = \beta_{j-1}\tau_{i,j-1} + \theta_{ij}$. We consider three specifications for θ_{ij}, each of which implies formulae for ψ_{jk} and $\sigma^2_{\theta_j}$. In each case, we utilize a random variable ε_{ij} having the following properties: $Cov(\varepsilon_{ij}, \varepsilon_{ik}) = 0 \ \forall j \neq k$, $V(\varepsilon_{ij}) = \sigma^2_{\varepsilon}$, $Cov(\theta_{ij}, \varepsilon_{i,j+m}) = 0 \ \forall m > 0$, and $Cov(\tau_{ij}, \varepsilon_{i,j+m}) = 0 \ \forall m > 0$.[32]

Model 1 is the relatively simple, but frequently employed, specification $\theta_{ij} = \mu_i + \varepsilon_{ij}$ where μ_i is a student-level random effect with $V(\mu_i) = \sigma^2_{\mu}$. It follows that $\sigma^2_{\theta_j} = V(\theta_{ij}) = \sigma^2_{\mu} + \sigma^2_{\varepsilon}$ and $Cov(\theta_{ij}, \theta_{ik}) = Cov(\mu_i + \varepsilon_{ij}, \mu_i + \varepsilon_{ik}) = \sigma^2_{\mu}, \ \forall j \neq k$. Here the variance of student achievement gains gross of any decay is constant across grades.

Model 2 is the specification $\theta_{ij} = \alpha\theta_{i,j-1} + \varepsilon_{ij}$ where $0 \leq \alpha < 1$. Note that $Cov(\theta_{ij}, \theta_{i,j+1}) = Cov(\theta_{ij}, \alpha\theta_{ij} + \varepsilon_{i,j+1}) = \alpha\sigma^2_{\theta_j}$ and, more generally, $Cov(\theta_{ij}, \theta_{i,j+m}) = \alpha^m\sigma^2_{\theta_j} \ \forall m > 0$.

Model 3 is the moving average $\theta_{ij} = \varepsilon_{ij} + \lambda_1\varepsilon_{i,j-1} + \lambda_2\varepsilon_{i,j-2}$. It follows that $V(\theta_{ij}) = (1 + \lambda_1^2 + \lambda_2^2)\sigma^2_{\varepsilon} = \sigma^2_{\theta}$ is constant across grades. Note that $Cov(\theta_{ij}, \theta_{i,j+1}) = \lambda_1(1 + \lambda_2)\sigma^2_{\varepsilon}$, $Cov(\theta_{ij}, \theta_{i,j+2}) = \lambda_2\sigma^2_{\varepsilon}$ and $Cov(\theta_{ij}, \theta_{i,j+m}) = 0 \ \forall m \geq 3$.

The three models differ in the degree to which the achievement gains of students are persistent over time, as shown in (A4). The expression for Model 2 is obtained through iterative

$$\begin{aligned} &\text{Model 1:} \quad \theta_{ij} = \varepsilon_{ij} + \mu_i \\ &\text{Model 2:} \quad \theta_{ij} = \varepsilon_{ij} + \alpha\varepsilon_{i,j-1} + \alpha^2\varepsilon_{i,j-2} + \alpha\theta_{i,j-3} \\ &\text{Model 3:} \quad \theta_{ij} = \varepsilon_{ij} + \lambda_1\varepsilon_{i,j-1} + \lambda_2\varepsilon_{i,j-2} \end{aligned} \tag{A4}$$

[32] We allow for the possibility that the mean of $\varepsilon_{i,j}$, say $E\varepsilon_{i,j} = \nu_j$, is nonzero and varies across grades. This generalizes the specification $E\varepsilon_{i,j} = 0, \forall j$, $E\varepsilon_{i,j}, \varepsilon_{i,k} = 0 \ \forall j \neq k$, $E\mu_i, \varepsilon_{i,j} = 0 \ \forall j$, and $E\tau_{i,j}, \varepsilon_{i,k} = 0 \ \forall k > j$. Note that the value of $V(\varepsilon_{ij}) = \sigma^2_{\varepsilon}$ can vary across the three models.

substitution. There is no diminution in the persistence of the θ_{ij} over time in Model 1; $Cov(\theta_{ij},\theta_{i,j+m})=\sigma^2_{\mu_i}$ is constant regardless of the grade span (i.e., for all *m*). The covariance of θ_{ij} and $\theta_{i,j+m}$ in Model 2 diminishes as m increases; $Cov(\theta_{ij},\theta_{i,j+m})=\ \alpha^m\sigma^2_{\theta_j}\ \ \forall m>0$. Even so, $Cov(\theta_{ij},\theta_{i,j+m})$ is greater than zero for all grade spans. In Model 3 there is no memory for spans of three grades or larger; $Cov(\theta_{ij},\theta_{ij+1})=\lambda_1(1+\lambda_2)\sigma^2_\varepsilon$, $Cov(\theta_{ij},\theta_{ij+2})=\lambda_2\sigma^2_\varepsilon$ and $Cov(\theta_{ij},\theta_{i,j+m})=0,\ \forall m\geq 3$. Memory would be limited to adjacent grades if $\lambda_1>0$ but $\lambda_2=0$ and there would be no persistence if $\lambda_1=\lambda_2=0$. Even though there is no memory in Model 3 across spans exceeding two grades, any pattern is possible for the first two years, as the values of λ_1 and λ_2 are not restricted. For example, if $\lambda_1=\alpha$ and $\lambda_2=\alpha^2$, the first three terms on the right hand side of the equations for Models 2 and 3 in (A4) would be identical. As a group, the three models include a wide range of possibilities. However, the estimation strategy discussed below can be extended to other specifications for θ_{ij} as well. We largely focus on Model 1 to illustrate the structural approach to estimation.

The specification $\theta_{ij}=\mu_i+\varepsilon_{ij}$ in Model 1 implies a relatively simple structure for the ψ_{jk}. For example, $\psi_{1k}=\psi_{1\bullet},\ \forall k$ and $\psi_{2\bullet}=\beta_1\psi_{1\bullet}+\sigma^2_\mu$. In general, $\psi_{jk}\equiv Cov(\tau_{ij},\theta_{ik})$ $=Cov(\tau_{ij},\mu_i)\ \equiv\psi_{j\bullet}$ where $\psi_{j\bullet}=Cov(\tau_{ij},\theta_{ik})=\ Cov(\beta_{j-1}\tau_{i,j-1}+\mu_i+\varepsilon_{ij},\ \mu_i+\varepsilon_{ik})=$ $\beta_{j-1}\psi_{j-1\bullet}+\sigma^2_\mu$. Thus, the value of ψ_{jk} follows from the values of $\beta_1,\beta_2,\cdots,\beta_{j-1},\psi_{1\bullet}$, and σ^2_μ. This structure, $\sigma^2_{\theta_j}=\sigma^2_\mu+\sigma^2_\varepsilon$, and Equations A2 and A3 imply the moment equations in (A5) and (A6), where the formulae for $\gamma_{jj},\ j\neq 1,$ in (A6) can be used to eliminate γ_{jj} in (A5). Estimation of the remaining parameters is relatively a straightforward.

$$
\begin{aligned}
&\omega_{11} - \gamma_{11}/G_1 = 0 \\
&\omega_{12} - \beta_1 \gamma_{11} + \psi_{1\cdot} = 0 \\
&\omega_{13} - \beta_2\beta_1 \gamma_{11} + (\beta_2 + 1)\psi_{1\cdot} = 0 \\
&\omega_{14} - \beta_3\beta_2\beta_1 \gamma_{11} + (\beta_3\beta_2 + \beta_3 + 1)\psi_{1\cdot} = 0 \\
&\omega_{15} - \beta_4\beta_3\beta_2\beta_1 \gamma_{11} + (\beta_4\beta_3\beta_2 + \beta_4\beta_3 + \beta_4 + 1)\psi_{1\cdot} = 0 \\
&\cdots \\
&\omega_{22} - \gamma_{22}/G_2 = 0 \\
&\omega_{23} - \beta_2 \gamma_{22} - [\beta_1\psi_{1\cdot} + \sigma_\mu^2] = 0 \\
&\omega_{24} - \beta_3\beta_2 \gamma_{22} - (\beta_3 + 1)[\beta_1\psi_{1\cdot} + \sigma_\mu^2] = 0 \\
&\omega_{25} - \beta_4\beta_3\beta_2 \gamma_{22} - (\beta_4\beta_3 + \beta_4 + 1)[\beta_1\psi_{1\cdot} + \sigma_\mu^2] = 0 \\
&\cdots \\
&\omega_{33} - \gamma_{33}/G_3 = 0 \\
&\omega_{34} - \beta_3 \gamma_{33} - [\beta_2\beta_1\psi_{1\cdot} + (\beta_2 + 1)\sigma_\mu^2] = 0 \\
&\omega_{35} - \beta_4\beta_3 \gamma_{33} - (\beta_4 + 1)[\beta_2\beta_1\psi_{1\cdot} + (\beta_2 + 1)\sigma_\mu^2] = 0 \\
&\cdots \\
&\omega_{44} - \gamma_{44}/G_4 = 0 \\
&\omega_{45} - \beta_4 \gamma_{44} - [\beta_3\beta_2\beta_1\psi_{1\cdot} + (\beta_3\beta_2 + \beta_3 + 1)\sigma_\mu^2] = 0 \\
&\cdots \\
&\omega_{55} - \gamma_{55}/G_5 = 0 \\
&\cdots
\end{aligned}
\qquad \text{(A5)}
$$

$$
\begin{aligned}
\gamma_{22} &= \beta_1^2\gamma_{11} + 2\beta_1\psi_{1\cdot} + \sigma_\mu^2 + \sigma_\varepsilon^2 \\
\gamma_{33} &= \beta_2^2\gamma_{22} + 2\beta_2\beta_1\psi_{1\cdot} + (2\beta_2 + 1)\sigma_\mu^2 + \sigma_\varepsilon^2 \\
\gamma_{44} &= \beta_3^2\gamma_{33} + 2\beta_3\beta_2\beta_1\psi_{1\cdot} + (2\beta_3\beta_2 + 2\beta_3 + 1)\sigma_\mu^2 + \sigma_\varepsilon^2 \\
\gamma_{55} &= \beta_4^2\gamma_{44} + 2\beta_4\beta_3\beta_2\beta_1\psi_{1\cdot} + (2\beta_4\beta_3\beta_2 + 2\beta_4\beta_3 + 2\beta_4 + 1)\sigma_\mu^2 + \sigma_\varepsilon^2
\end{aligned}
\qquad (A6)
$$

Suppose that the values of $\hat{\omega}_{ij}$ have been estimated employing student test scores spanning J grades (i.e., $i, j = 1, 2, ..., J$). Let $\hat{\Delta}_c$ represent a column vector made up of the $N_m = J(J+1)/2$ moment equations following the structure in (A5) with $\hat{\omega}_{ij}$ substituted for ω_{ij} and the formulae in (A6) substituted for $\gamma_{jj},\ j \neq 1$. Here $\hat{\Delta}_c$ is a function of the $N_p = 2J + 3$

parameters in $\pi = \left[\gamma_{11} \; \psi_{1.} \; \sigma_{\mu}^{2} \; \sigma_{\varepsilon}^{2} \; \beta_1 \; \beta_2 \cdots \beta_{J-1} \; G_1 \; G_2 \cdots G_J \right]$, which can be estimated using the generalized-method-of-moments estimator described toward the end of Section 3.2.

In background analysis we estimated the three structural models and found that estimates of all three specifications yielded predicted covariance structures that fit the covariance of test scores well, with little evidence that any one was a superior specification. Important given the motivation for this paper, the estimated patterns of test measurement error are quite robust across the three structural models and the reduced-form model discussed in the paper.

Appendix B

As noted in the paper, measurement error can, but need not, result in $E\left(S_{i,j+1} \mid S_{ij}\right)$ being a nonlinear function of S_{ij} even when $E\left(\tau_{i,j+1} \mid \tau_{ij}\right)$ is linear in τ_{ij}. It is not measurement error *per se* that implies the nonlinearity, as $E\left(S_{i,j+1} \mid S_{ij}\right)$ is linear in S_{ij} if the measurement-error is homoskedastic (i.e., $\sigma^2_{\eta_{ij}} = \sigma^2_{\eta_{\bullet j}}, \forall i$). However, $E\left(S_{i,j+1} \mid S_{ij}\right)$ is nonlinear in S_{ij} when $E\left(\tau_{i,j+1} \mid \tau_{ij}\right)$ is linear but η_{ij} is heteroskedastic with the extent of measurement error varying with the ability level (i.e., $\sigma_{\eta_{ij}} = \sigma_{\eta_j}(\tau_{ij})$). When $\sigma_{\eta_j}(\tau_{ij})$ is U-shaped, as in Figures 1 and 2, $E\left(S_{i,j+1} \mid S_{ij}\right)$ is an S-shaped function of S_{ij}. An explanation and example follow.

Consider the case where $\tau_{ij} \sim N(\mu_j, \sigma^2_{\tau_j})$, $E\left(\tau_{i,j+1} \mid \tau_{ij}\right) = \beta_0 + \beta_1 \tau_{ij}$ and $S_{ij} = \tau_{ij} + \eta_{ij}$. Note that $\tau_{i,j+1} = \beta_0 + \beta_1 \tau_{ij} + \upsilon_{ij}$ with $E\tau_{ij}\,\upsilon_{ij} = 0$ so that $S_{i,j+1} = \beta_0 + \beta_1 S_{ij} - \beta_1 \eta_{ij} + \eta_{i,j+1} + \upsilon_{ij}$. In turn, $E\left(S_{i,j+1} \middle| S_{ij}\right) = \beta_0 + \beta_1 S_{ij} - \beta_1 E\left(\eta_{ij} \middle| S_{ij}\right)$ since both $E\left(\upsilon_{ij} \middle| S_{ij}\right)$ and $E\left(\eta_{i,j+1} \middle| S_{ij}\right)$ are zero. In the homoskedastic case ($\sigma^2_{\eta_{ij}} = \sigma^2_{\eta_{\bullet j}}$), η_{ij} and S_{ij} are bivariate normal, as shown in (B-1), implying that $E\left(\eta_{ij} \middle| S_{ij}\right) = \dfrac{\sigma^2_{\eta_{\bullet j}}}{\sigma^2_{\tau_j} + \sigma^2_{\eta_{\bullet j}}}\left(S_{ij} - \mu_j\right) = \left(1 - G_j\right)\left(S_{ij} - \mu_j\right)$ where $G_j \equiv \sigma^2_{\tau_j} \Big/ \left(\sigma^2_{\tau_j} + \sigma^2_{\eta_{\bullet j}}\right)$.

$$\begin{bmatrix} \eta_{ij} \\ S_{ij} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ \mu_j \end{bmatrix}, \begin{bmatrix} \sigma^2_{\eta_{\bullet j}} & \sigma^2_{\eta_{\bullet j}} \\ \sigma^2_{\eta_{\bullet j}} & \sigma^2_{\tau_j} + \sigma^2_{\eta_{\bullet j}} \end{bmatrix} \right) \quad \text{(B-1)}$$

It follows that $E\left(S_{i,j+1} \middle| S_{ij}\right) = \beta_0 + \beta_1\left(1 - G_j\right)\mu_j + \beta_1 G_j S_{ij}$ is linear in S_{ij}.

As discussed in the paper, S_{ij} and η_{ij} are not bivariate normal when the extent of measurement error varies across ability levels (i.e., $\sigma_{\eta_{ij}} = \sigma_{\eta_j}(\tau_{ij})$). Similar to the way

$E\left(\tau_{ij}\left|S_{ij}\right.\right)$ can be computed, $E\left(\eta_{ij}\left|S_{ij}\right.\right)=\frac{1}{k^{j}\left(S_{ij}\right)M}\sum_{m=1}^{M}\eta_{ij}\,\phi\left(\eta_{ij}\big/\sigma_{\eta_j}^{2}(\tau_{mj}^{*})\right)\Big/\sigma_{\eta_j}(\tau_{mj}^{*})$ where $\phi(\,)$ is the standard-normal density, $k^{j}\left(S_{ij}\right)$ is the score density and τ_{mj}^{*} is a random draw from the distribution of τ_{ij} . In this way, we can compute $E\left(S_{i,j+1}\left|S_{ij}\right.\right)=\beta_0+\beta_1 S_{ij}-\beta_1 E\left(\eta_{ij}\left|S_{ij}\right.\right)$. For example, suppose that $\tau_{ij}\sim N(670,30)$ and $\eta_{ij}\sim N\left(0,\,\sigma_{\eta}^{2}(\tau_{ij})\right)$ with $\sigma_n(\tau_{ij})=\sigma_o+\gamma(\tau_{ij}-\mu_j)^2$ and $\sigma_{n_j.}=E_{\tau_j}\sigma_n(\tau_{ij})=\sigma_o+\gamma\sigma_{\tau_j}^{2}=15$. (These assumptions are roughly consistent with the patterns found for the NYC test scores.) The three cases shown in Figure B.1 differ with respect to the degree of heteroskedasticity: the homoskedastic case ($\sigma_o=15$ and $\gamma=0$), moderate heteroskedasticity ($\sigma_o=12$ and $\gamma=0.00333\cdots$) and a more extreme heteroskedasticity ($\sigma_o=3$ and $\gamma=0.01333\cdots$). Values of $E\left(S_{i,j+1}\left|S_{ij}\right.\right)$ for the three cases are shown in Figure B.2. $E\left(S_{i,j+1}\left|S_{ij}\right.\right)$ is linear in the homoskedastic case and the degree to which $E\left(S_{i,j+1}\left|S_{ij}\right.\right)$ is S-shaped depends upon the extent of this particular type of heteroskedasticity.

Knowing that the observed S-shape patterns of $E\left(S_{i,j+1}\left|S_{ij}\right.\right)$ can be consistent with $E\left(\tau_{i,j+1}\,|\,\tau_{ij}\right)$ being linear in τ_{ij} is useful, but of greater importance is whether $E\left(\tau_{i,j+1}\,|\,\tau_{ij}\right)$ is in fact linear for the tests of interest. This can be explored employing the cubic specification $\tau_{i,j+1}=\beta_0+\beta_1\tau_{ij}+\beta_2\tau_{ij}^{2}+\beta_3\tau_{ij}^{3}+\upsilon_{i,j+1}$ where $\beta_2=\beta_3=0$ implies linearity. Substituting $S_{ij}=\tau_{ij}+\eta_{ij}$ and regressing $S_{i,j+1}$ on S_{ij} would yield biased parameter estimates. However, if $\lambda_{ik}\equiv E\left((\tau_i^{*})^{k}\left|S_i^{*}\right.\right)$, $k=1,2,3,4$, were know for each student, regressing $S_{i,j+1}$ on λ_{i1}, λ_{i2}, λ_{i3}, and λ_{i4} would yield consistent estimates.[33]

[33] See the discussion of the "structural least squares" estimator in Kukush et. al (2005) .

Computing the $\lambda_{ik} \equiv E\left((\bar{\tau}_i^*)^k \left| \bar{S}_i^* \right.\right)$, $k = 1,2,3,4$, for each student requires knowledge of the overall extent and pattern of measurement error. It is the lack of such knowledge that motives this paper. However, we are able to compute $\hat{\lambda}_{ik} = \hat{E}\left((\bar{\tau}_i^*)^k \left| \bar{S}_i^* \right.\right)$ accounting for the meaningful measurement-error heteroskedasticity reflected in the reported SEMs[34], even though this does not account for other sources of measurement error. Computation of $\hat{E}\left((\bar{\tau}_i^*)^k \left| \bar{S}_i^* \right.\right)$ also requires an estimate of $\sigma_{\tau_j}^2$ which can be obtained by solving for $\hat{\sigma}_{\tau_j}^2$ implicitly defined in $\hat{\sigma}_{\tau_j}^2 = \hat{\sigma}_{S_j}^2 - \hat{\sigma}_{\eta_{\cdot j}}^2 = \hat{\sigma}_{S_j}^2 - \int \sigma_\eta^2(\tau)\, f\left(\tau \left| \hat{\mu}_j, \hat{\sigma}_{\tau_j}^2 \right.\right) d\tau$. We solve for $\hat{\sigma}_{\tau_j}^2$ iteratively by computing $\hat{\sigma}_{S_j}^2 - \int \sigma_\eta^2(\tau)\, f\left(\tau \left| \hat{\mu}_j, \tilde{\sigma}_{\tau_j}^2 \right.\right) d\tau = \hat{\sigma}_{S_j}^2 - \frac{1}{M}\sum_m^M \sigma_\eta^2\left(\tau_{mj}^*\right)$. Here the integral $\int \sigma_\eta^2(\tau)\, f\left(\tau \left| \hat{\mu}_j, \tilde{\sigma}_{\tau_j}^2 \right.\right) d\tau$ is computed using Monte Carlo integration with importance sampling where the τ_{mj}^* are random draws from the distribution $N\left(\hat{\mu}_j, \tilde{\sigma}_{\tau_j}^2\right)$ and $\tilde{\sigma}_{\tau_j}^2$ is an initial estimate of $\sigma_{\tau_j}^2$. This yielded an updated value of $\tilde{\sigma}_{\tau_j}^2$ which can be used to repeat the prior step. Relatively few iterations are needed for converge to the fixed-point – our estimate of $\sigma_{\tau_j}^2$ The estimate $\hat{\sigma}_{\tau_j}^2$ allows us to compute values of $\hat{\lambda}_{ik}$ and, in turn, regress $S_{ij}+1$ on $\hat{\lambda}_{i1}$, $\hat{\lambda}_{i2}$, $\hat{\lambda}_{i3}$, and $\hat{\lambda}_{i4}$.

.

[34] Because SEM values are reported for a limited set of scores, a flexible functional form for $\sigma_\eta^2(\tau)$ was fit to the reported SEM. This function was then used in computation of moments.

Figure B.1
Examples Showing Different Degrees of Heteroskedastic Measurement Error

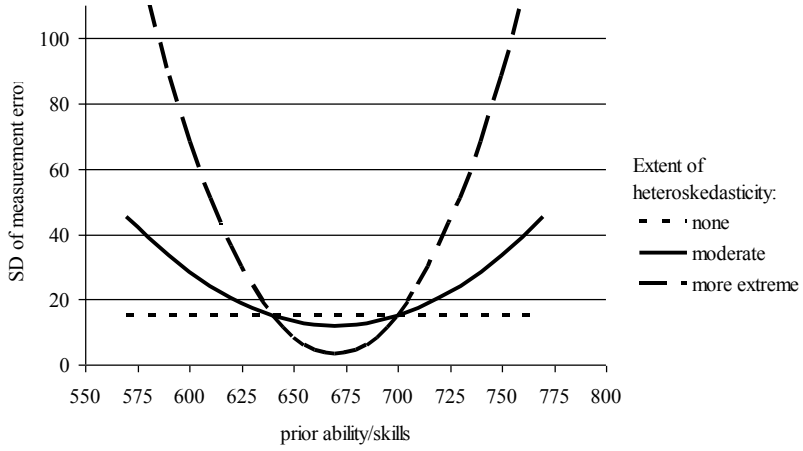


Figure B.2
How the Relationship Between $E\left(S_i^+ \middle| S_i\right)$ and S_i
Varies with the Degree of Heteroskedasticity