Measuring Test Measurement Error: A General Approach

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

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

November 27, 2012

We are grateful to the New York City Department of Education and the New York State Education Department for the data employed in this paper. We appreciate financial support from the National Science Foundation and National Center for the Analysis of Longitudinal Data in Education Research (CALDER). CALDER is supported by IES Grant R305A060018 to the American Institutes for Research. Thanks also to Dale Ballou, Ed Haertel, Dan McCaffrey, Tim Sass and Jeffery Zabel for their helpful comments. The authors are solely responsible for the content of this paper.

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 that there is a credible, low-cost approach for estimating the total measurement error that can be applied when students take three or more tests in the subject of interest (e.g., state assessments in consecutive grades). Our method generalizes the test-retest framework by allowing for (i) growth or decay in knowledge and skills between tests, (ii) tests being neither parallel nor vertically scaled, and (iii) the degree of measurement error varying 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 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 at least 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 can yield meaningful improvements in the precision of student achievement and achievement-gain estimates.

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. The estimates of reliability provided by test vendors 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 achievement tests we analyze, the measurement error is at least twice as large as that indicated in the technical reports provided by the test vendor. 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 such a technique and provides evidence of its usefulness.

Thorndike (1951) articulates a variety of factors that 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 (i.e., 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 error from other sources.

Reliability coefficients based on the test-retest approach using parallel test forms is viewed

in the psychometric literature to be 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 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 is to divide a single test into parallel parts. Such split-test reliability only accounts for the measurement error resulting from the random selection 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 being that potentially important day-to-day differences in student performance are ignored.

In this paper we show that there is a credible approach for measuring the overall extent of measurement error applicable in a wide variety of settings. Estimation is straightforward and only requires estimates of the correlations of test scores in the subject of interest at several points in time (e.g., third-, fourth- and fifth-grade math scores for a cohort of students). Student-level data is not needed, provided that estimates of test-score variances and correlations are available. Our approach generalizes the test-retest framework to allow for either growth or decay in the knowledge and skills of students between tests, allow for tests to be neither parallel nor vertically scaled and allow for the extent of measurement error to differ across tests. Utilizing test-score

[1] Feldt and Brennan (1989).

covariance or correlation estimates and maintaining minimal structure characterizing the nature of achievement growth, one can estimate the overall extent of test measurement error and decompose the test-score variance into the part attributable to real differences in achievement and the part attributable to measurement error.

In the following section we briefly introduce generalizability theory and show how the total measurement error is reflected in the covariance structure of observed test scores. In turn, we explain our statistical approach and 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 extent of test measurement error can be informative in analyses related to educational practice and policy.

1.0 Measurement Error and the Structure of Test-Score Covariances

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. Generalizability theory extends test theory to explicitly account for multiple sources of measurement error.[2] Consider the case where a student takes a test consisting of a set of tasks (e.g., questions) drawn from some universe of similar conditions of measurement with the student doing the tasks 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. Randomness in the selection of test items along with students doing especially well or poorly on particular tasks is one source of measurement error. Temporal instability in student performance is another.

Let η_{iT} represent a composite measure of the errors in measurement from all sources for a

[2] Many authors discuss classical test theory, e.g, Haertel (2006). See Cronbach et al. (1997) and Feldt and Brennan (1989) for useful introductions to Generalizability Theory and Brennan (2001) for more detail.

student taking test *T* at a particular point in time. The actual score, S_{iT}, equals $\tau_i + \eta_{iT}$ where τ_i is the student's *universe score* equaling the expected value of S_{iT} over the universe of generalization (e.g., the universes of possible tasks and occurrences). The universe score, comparable to the true score in classical test theory, measures the student's underlying academic achievement.

Generalizing the notation to allow for multiple tests, S_{ij} in $S_{ij} = \tau_{ij} + \eta_{ij}$ is the i^{th} student's score on a test for a particular subject taken in the *j*th testing period, for exposition here assumed to be one per grade.[3] τ_{ij} is the student's true achievement in a subject at the time the test is administered. We drop subscript "T' to simplify notation, but maintain that a different test in a single occurrence is given in each period. η_{ij} is the test measurement error from all sources, where $E\eta_{ij} = 0$, $E\eta_{ij}\tau_{ik} = 0$, $\forall j,k$ and $E\eta_{ij}\eta_{ik} = 0$, $\forall\ j \neq k$. Allowing for heteroskedasticity across grades and students, $E\eta_{ij}^2 = \sigma_{\eta_{ij}}^2$. Let $\sigma_{\eta_{\bullet j}}^2$ equal $\sigma_{\eta_{ij}}^2$ for all pupils in the homoskedastic case or, more generally, the mean value of $\sigma_{\eta_{ij}}^2$ for the universe of students in grade j.

Using vector notation, $S_i = \tau_i + \eta_i$ where $S_i' = \begin{bmatrix} S_{i1} & S_{i2} & \cdots & S_{iJ} \end{bmatrix}$, $\tau_i' = \begin{bmatrix} \tau_{i1} & \tau_{i2} & \cdots & \tau_{iJ} \end{bmatrix}$, and $\eta_i' = \begin{bmatrix} \eta_{i1} & \eta_{i2} & \cdots & \eta_{iJ} \end{bmatrix}$ for the first through the J^{th} tested grades[4]. Equation 1 defines $\Omega(i)$ to be the auto-covariance matrix for the i[th] student's observed test scores. H is the auto-covariance

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

$$= \begin{bmatrix} \omega_{i11} & \omega_{i12} & \cdots & \omega_{i1J} \\ \omega_{i21} & \omega_{i22} & \cdots & \omega_{i2J} \\ \vdots & \vdots & \ddots & \\ \omega_{iJ1} & \omega_{iJ2} & & \omega_{iJJ} \end{bmatrix} = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1J} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2J} \\ \vdots & \vdots & \ddots & \\ \gamma_{J1} & \gamma_{J2} & & \gamma_{JJ} \end{bmatrix} + \begin{bmatrix} \sigma_{\eta_{i1}}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{\eta_{i2}}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \sigma_{\eta_{iJ}}^2 \end{bmatrix} = \mathrm{H} + \mathrm{Z}_i \quad (1)$$

[3] In general, time intervals between tests need not be annual or 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.

[4] For example, the third grade might be the first tested grade. To simplify exposition, we often will not distinguish between the i^{th} grade and the i^{th} tested grade, even though we will mean the latter.

matrix for the universe scores in the population of students. Z_i is the diagonal matrix with the measurement-error variances for the *i*th student, $\sigma^2_{\eta_{ij}}$, on the diagonal. For the population of students, $\sigma^2_{\eta_{\bullet j}} = E\sigma^2_{\eta_{ij}}$ and $\Omega_\bullet = \mathrm{E}\Omega(i) = \mathrm{H} + \mathrm{Z}_\bullet$ 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. $\Omega(i)$ differs from $\Omega(i')$ only because of possible heteroskedastic measurement error across test-takers.

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 in the student population, γ_{jj}, and the measurement-error variance; $\omega_{jj} = \gamma_{jj} + \sigma^2_{\eta_{\bullet j}}$. The *generalizability coefficient,* $G_j \equiv \gamma_{jj} / \omega_{jj}$, measuring the portion of the variance in observed scores that is explained by the variance of universe scores, implies the characterization of $\Omega_\bullet$ shown in Equation 2. $\Omega_\bullet$ can be estimated using its empirical counterpart $\hat{\Omega}_\bullet = \sum_i \left(S_i - \bar{S}\right)\left(S_i - \bar{S}\right)' / N_S$ where N_S is the number of students with observed test scores.[5]

$$\Omega_\bullet = \begin{bmatrix} \omega_{11} & \omega_{12} & \omega_{13} & & \omega_{1J} \\ & \omega_{22} & \omega_{23} & \cdots & \omega_{2J} \\ & & \omega_{33} & & \omega_{3J} \\ & & & \ddots & \vdots \\ & & & & \omega_{JJ} \end{bmatrix} = \begin{bmatrix} \gamma_{11}/G_1 & \gamma_{12} & \gamma_{13} & & \gamma_{1J} \\ & \gamma_{22}/G_2 & \gamma_{23} & \cdots & \gamma_{2J} \\ & & \gamma_{33}/G_3 & & \gamma_{3J} \\ & & & \ddots & \vdots \\ & & & & \gamma_{JJ}/G_J \end{bmatrix} \quad (2)$$

Equation 2 and the correlation of universe scores in grades *j* and *k*, $\rho_{jk} \equiv \gamma_{jk} / \sqrt{\gamma_{jj}\gamma_{kk}}$, imply the test-score correlation matrix *R* in Equation 3. r_{jk} is the test-score correlation for grades *j* and *k*. Note that the presence of test measurement error (e.g., $G_j < 1$) implies that each correlation of

[5] 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 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.

$$R=\begin{bmatrix} 1 & r_{12} & r_{13} & r_{14} & r_{15} & \\ & 1 & r_{23} & r_{24} & r_{25} & \cdots \\ & & 1 & r_{34} & r_{35} & \\ & & & 1 & r_{35} & \\ & & & & 1 & \\ & & & & & \ddots \end{bmatrix}=\begin{bmatrix} 1 & \sqrt{G_1G_2}\rho_{12} & \sqrt{G_1G_3}\rho_{13} & \sqrt{G_1G_4}\rho_{14} & \sqrt{G_1G_5}\rho_{15} & \\ & 1 & \sqrt{G_2G_3}\rho_{23} & \sqrt{G_2G_4}\rho_{24} & \sqrt{G_2G_5}\rho_{25} & \cdots \\ & & 1 & \sqrt{G_3G_4}\rho_{34} & \sqrt{G_3G_5}\rho_{35} & \\ & & & 1 & \sqrt{G_4G_5}\rho_{45} & \\ & & & & 1 & \\ & & & & & \ddots \end{bmatrix} \quad (3)$$

test scores is smaller than the corresponding correlation of universe scores. In contrast, estimates of the off-diagonal elements of the test-score covariance matrix (i.e., $\hat{\omega}_{ij}$) directly imply estimates of the off-diagonal elements of the universe-score covariance (i.e., $\hat{\gamma}_{ij}$).

Estimates of the $\omega_{jk}=\gamma_{jk}$ alone are not sufficient to infer estimates of the γ_{jj} and G_j, the problem being that there are *J* more parameters in Equations 2, as well as 3, than there are moments.[6] However, there is a voluminous literature in which researchers employ more parsimonious covariance and correlation matrix specifications, the goal being to economize on the number of parameters one needs to estimate while retaining sufficient flexibility in the covariance structure. For a variety of such structures, one can estimate γ_{jj} and G_j. However, the reasonableness of any particular structure will be context specific.

As an example, suppose that one knew or had estimates of test-score correlations for parallel tests taken at times $t_1, t_2, \cdots, t_J$. Time intervals between consecutive tests can vary. Correlation structures that allow for changes in skills and knowledge over time typically maintain that the correlation between any two universe scores is smaller the longer is the time span between the tests, one possible specification being $\rho_{jk}=\rho^{|t_k-t_j|}$. Employing this structure and assuming $G_j=G,\ \forall j$, G and ρ are identified with three test administrations; $\rho=\left(r_{13}/r_{12}\right)^{1/|t_k-t_j|}$ and

[6] In Equation 2 there are $J(J+1)/2$ moments and $J+J(J+1)/2=J(J+3)/2$ parameters. In Equation 3 there are $J(J-1)/2$ moments and $J+J(J-1)/2=J(J+1)/2$ parameters.

$G = r_{12}r_{23}/r_{13}$.[7] If $J \geq 4$, $G_1, G_2, \ldots G_J$ and ρ are identified.

This example generalizes the congeneric model analyzed by Joreskog (1971). Tests are said to be *congeneric* if the true scores, τ_{ik}, are linear functions of a common $\tau_{i\bullet}$, implying that the true scores are perfectly correlated. For this case, Joreskog shows that G_1, G_2, and G_3 are identified, which in turn generalizes the test-retest framework where $\rho = 1$ and $G_j = G, \forall j$.

Even though the structure $\rho_{jk} = \rho^{|t_k - t_j|}$ has potential uses, it is far from general. The central contribution of this paper is to show that the overall extent of test measurement error and universe-score variances can be estimated maintaining far less restrictive universe-score covariance structures, thereby substantially generalizing the test-retest approach.The intuition is relatively straightforward. For a wide range of universe-score covariance structures, γ_{jk} in Equation 2 can be expressed as functions of γ_{jj} and γ_{kk}, which in turn can be expressed in terms of γ_{11}.[8] In such cases, estimates of the $\omega_{jk} = \gamma_{jk}, j \neq k,$ can be used to estimate γ_{jj} and $G_j = \gamma_{jj} / \omega_{jj}$.

Consider the important exception where one or more of the universe scores are multidimensional with at least one dimension of ability not correlated with any of the other universe scores, e.g., $\tau_{i2} = \tau^o_{i2} + \psi_{i2}$ with $Cov(\psi_{i2}, \tau_{ik}) = 0$ and $Cov(\tau^o_{i2}, \tau_{ik}) \neq 0, \ \forall k \neq 2$. Because $\omega_{2k} = \gamma_{2k}$ is not a function of $V(\psi_{i2})$, knowledge of ω_{jk} does not identify $V(\psi_{i2})$, $\gamma_{22} = V(\tau^o_{i2}) + V(\psi_{i2})$ or $G_2 = \left[V(\tau^o_{i2}) + V(\psi_{i2})\right] / \omega_{22}$. Thus, application of our approach in cases with

[7] The corresponding estimators of ρ and G are biased, but consistent, as they are ratios of estimated covariances. The same is true in several other examples shown below.

[8] In general $\tau_{i\,j+m} = E(\tau_{i\,j+m} | \tau_{ij}) + \delta_{i\,j+m}$ where $E\,\delta_{i\,j+m}\,\tau_{ij} = 0$. Utilizing a Taylor-series approximation for $E(\tau_{i\,j+1} | \tau_{ij})$, $\tau_{i\,j+m} = a_0^m + a_1^m(\tau_{ij} - \mu_j) + a_2^m(\tau_{ij} - \mu_j)^2 + \cdots + \delta_{i\,j+m}$ where $\mu_j = E\tau_{ij}$. Thus, $\gamma_{j\,j+m} = E(\tau_{ij} - \mu_j)(\tau_{i\,j+m} - \mu_{j+m}) = a_1^m \gamma_{jj} + a_2^m \sigma^3_{\tau_j} + \cdots$, where $\gamma_{j\,j+m}$ is a function of γ_{jj}.

multidimensional abilities requires that every skill and ability measured by each test is correlated with skills and abilities measured by the other tests. Regarding dimensionality, it is relevant to note that the IRT models used in test scoring typically maintain that each test measures ability along a single dimension, which can be tested.

Intuition also follows from Equation (3). First, note that it is always the case that $(r_{13}/r_{14})/(r_{23}/r_{24}) = (\rho_{13}/\rho_{14})/(\rho_{23}/\rho_{24})$. In general, $r_{gj}/r_{hj} \;:\; r_{gk}/r_{hk}$ as $\rho_{gj}/\rho_{hj} \;:\; \rho_{gk}/\rho_{hk}$. Also, in many circumstances it is reasonable to maintain that the universe-score correlation matrix follows some general structure, which implies functional relationships between the ρ_{jk} and simplifies expressions such as $(\rho_{13}/\rho_{14})/(\rho_{23}/\rho_{24})$ above (e.g., the ρ_{jk} are functions of a reduced number of parameters). In this way, the relative magnitudes of the r_{jk} are key in identifying the ρ_{jk} For example, the case $\rho_{jk} = \rho^{|t_k - t_j|}$ implies that $\rho = \left(r_{13}/r_{12}\right)^{1/|t_k - t_j|}$. More generally, the pattern in how the $r_{j,\,j+s}$ decline as s increases in the j^{th} row (column) relative to the pattern of decline for $r_{j'\,j'+s}$ in other rows (columns) is key in identifying ρ_{jk}.

Next note, for example, that an increase in the extent of measurement error in the second test (i.e., a decrease in G_2), other things constant, implies the same proportionate reduction in every test-score correlation in the second row and second column of R, but no change in any of the other test-score correlations, as G_2 only appears in that row and column. Whether G_2 is identified crucially depends upon whether a decrease in G_2 is the only explanation for such a proportionate decrease in $r_{2k},\ \forall k,$ with no change in $r_{mn},\ m,n \neq 2$. This is not always the case; an increase in $V(\psi_{i2})$ in the above example would imply a proportionate decline in $\rho_{2k,}\ \forall k,$ with $\rho_{mn},\ m,n \neq 2,$ unchanged. However, for many tests analysts will find it reasonable to rule out this

possibility (e.g., find it unreasonable that the universe-score correlation for the first and third tests in a series could remain unchanged even though the universe-score correlations between the first and second, as well as between the second and third, tests both increased). More generally, a variety of universe-score correlation structures rule out the possibility of a proportionate change in $\rho_{jk}, \forall k,$ with no change in every $\rho_{mn}, m,n \neq j$. In those cases, a proportionate change in the $r_{jk}, \forall k,$ with no change in $r_{mn}, m,n \neq j,$ necessarily implies an equal proportionate change in G_h.

In summary, all test-score correlations being smaller can reflect either larger measurement error or smaller universe score correlations, or a combination of both. Fortunately, it is possible to distinguish between these explanations in a variety of settings, including situations in which tests are neither parallel nor vertically scaled. In fact, the tests can measure different abilities, provided that there is no ability measured by a test that is uncorrelated with the abilities measured by the other tests. A second requirement is that one can credibly maintain minimal structure characterizing the universe-score correlations for the tests being analyzed.

Our approach draws upon the work of Abowd and Card (1989) studying the covariance structure of individual and household earnings, hours worked and other time-series variables. Especially important is the analysis of covariance structures developed by Joreskog (1978), the kernel of which can be found in Joreskog (1971).

2.0 Estimation Strategy

To estimate the overall extent of test measurement error and decompose the variance of test scores into the part attributable to real differences in achievement and the part attributable to measurement error requires estimates of test-score variances and covariances or correlations along with assumptions regarding the structure characterizing universe-score covariances or correlations. One option is to directly specify the structure of the ρ_{jk} (e.g., assume $\rho_{jk} = \rho^{|t_k - t_j|}$). An

alternative is to maintain a reasonable model of achievement growth and take advantage of structural features characterizing universe-score correlations implied by that specification.

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

$$\tau_{ij} = \beta_{j-1}\tau_{i,j-1} + \theta_{ij} \ . \qquad (4)$$

This first-order autoregressive structure models attainment in grade *j* as depending upon the level of knowledge and skills in the prior grade,[9] possibly subject to decay (if $\beta_{j-1} < 1$) that can vary across grades. A key assumption is that decay is not complete, i.e., $\beta_j \neq 0$. $\beta_j = \beta, \forall j,$ is a special case as is $\beta_j = 1$. θ_{ij} is the gain in student achievement in grade j, gross of any decay.[10]

For empirical growth models to actually measure growth in the underlying achievement of students, the test(s) must reflect a single *interval scale*, meaning that "equal-sized gains at all points on the scale represent the same increment of learning".[11] For example, the tests used to estimate $\tau_{i1}, \tau_{i2}, \cdots$ in Equation 4 could reflect a common vertical scale. Given the prevalence of questions regarding whether test scales in practice are the same across grades and years[12], it is fortunate that our approach need only employ test-score correlations. Thus, the individual tests each need to reflect an interval scale, but the scales can differ across tests.[13] Even though the extent of test measurement error for the individual tests, measured by the G_j, can be inferred even when tests are not vertically scaled, the lack of vertical scaling would have a number of undesirable consequences (e.g., certain parameters cannot be identified).

$$\tau_{ij} = \theta_{ij} + \beta_{j-1}\theta_{i,j-1} + \beta_{j-1}\beta_{j-2}\theta_{i,j-2} + \cdots + (\beta_{j-1}\beta_{j-2}\cdots\beta_{j-(s-1)}\theta_{i,j-(s-1)}) + (\beta_{j-1}\beta_{j-2}\cdots\beta_{j-s}\tau_{i,j-s}) \quad (5)$$

[9] Todd and Wolpin (2003) discuss the conditions under which this will be the case.
[10] 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_{ij} as the student's achievement gain even when $\beta_j \neq 1$.
[11] Ballou (2009).
[12] See Ballou (2009) for an informative analysis.
[13] One possibility is that even though achievement across the grades falls on a single interval scale, the tests instruments employed do not have that property.

Equation 4 can be used to infer Equation 5 that shows τ_{ij} reflects the accumulation of decayed values of prior θ_{ij}. As in other time-series models, one can assume that the sum in Equation 5 extends back to an infinite past (i.e., $s \to \infty$). A more attractive alternative in our application is to assume that the time-series for each student begins at a specified point in time (e.g., when 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 5 and the statistical structure of the θ_{ij} determine the dynamic pattern of universe scores reflected in the parameterization of Γ and $\Omega_{\bullet}$.

Two approaches can be used to characterize the statistical structure of the θ_{ij}. One is to fully specify the relationship of achievement gains across grades. One such specification is $\theta_{ij} = \mu_i + \varepsilon_{ij}$ where μ_i is a student-level random effect and ε_{ij} is white noise. Alternatives include moving averages and a first-order autoregressive process. Each such structure implies the joint distribution of $\theta_{i,j+1}$ and τ_{ij} and the conditional mean function $E(\theta_{i,j+1} \mid \gamma_{ij})$. (See Boyd, Lankford, Loeb and Wyckoff, 2012). An alternative, "reduced-form" approach is to explicitly assume that $E(\theta_{i,j+1} \mid \gamma_{ij})$ follows a particular functional form (e.g., linearity in γ_{ij}) without specifying a particular set of underlying assumptions that implies that functional form. Because of its simplicity, here we employ the reduced-form framework.

2.1 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. 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}.

Linearity implies that $\theta_{i,j+1} = a_j + b_j\,\tau_{ij} + u_{i,j+1}$ which, 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 $c_j \equiv \beta_j + b_j$. The universe score in grade *j+1* is a linear function of the universe score in the prior grade. The two components of coefficient c_j reflect (1) part of the student's proficiency in grade *j+1* already being attained in grade *j*, attenuated per Equation 4, and (2) the expected growth during year *j+1* being linearly dependence on the prior-year achievement, τ_{ij}.

The reduced-form model $\tau_{i,j+1} = a_j + c_j \tau_{ij} + u_{i,j+1}$ implies that $\rho_{j,j+1} = c_j \sqrt{\gamma_{jj} / \gamma_{j+1,j+1}}$ (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 3 implies the moment conditions in Equation 6. Because $\sqrt{G_1}$ and ρ_{12} only appear as

$$R = \begin{bmatrix} r_{12} & r_{13} & r_{14} & \cdots \\ & r_{23} & r_{24} & \cdots \\ & & r_{34} & \cdots \\ & & & \ddots \end{bmatrix} = \begin{bmatrix} \sqrt{G_1 G_2}\,\rho_{12} & \sqrt{G_1 G_3}\,\rho_{12}\rho_{23} & \sqrt{G_1 G_4}\,\rho_{12}\rho_{23}\rho_{34} & \cdots \\ & \sqrt{G_2 G_3}\,\rho_{23} & \sqrt{G_2 G_4}\,\rho_{23}\rho_{34} & \cdots \\ & & \sqrt{G_3 G_4}\,\rho_{34} & \cdots \\ & & & \ddots \end{bmatrix} \quad (6)$$

a multiplicative pair, the 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 $N_m = J\left(J-1\right)/2$ moments in Equation 6 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 $J = 3$ is sufficient for identification. For example, when $G_j = G$ estimates of the test-score correlations

for J=3 tests imply the estimators of the elemental parameters in Equation 7.

$$\hat{\rho}_{12} = \hat{r}_{13}/\hat{r}_{23} \qquad \hat{\rho}_{23} = \hat{r}_{13}/\hat{r}_{12} \qquad \hat{G} = \hat{r}_{12}\hat{r}_{23}/\hat{r}_{13} \quad (7)$$

In general, estimated test-score correlations together with assumptions regarding the structure of student achievement growth are sufficient to estimate the universe-score correlations and the relative extent of measurement error measured by the generalizability coefficients. In turn, estimates of G_j and the test score variance, ω_{jj}, imply the universe-score variance estimator $\gamma_{jj} = \omega_{jj}G_j$, measuring the dispersion in student achievement in grade *j*, as well as the estimator of the variance of test measurement error from all sources $\sigma^2_{\eta_{\bullet j}} = \omega_{jj}(1-G_j)$.

The equations in (7) illustrate the general intuition regarding identification discussed in Section 1.0. Consider the implications of $\hat{r}_{12}$, $\hat{r}_{23}$, and $\hat{r}_{13}$ being smaller. First, this does not necessarily imply an increase in the extent of test measurement error. The last equation in (7) implies that $d\hat{G}/\hat{G} = d\hat{r}_{12}/\hat{r}_{12} + d\hat{r}_{23}/\hat{r}_{23} - d\hat{r}_{13}/\hat{r}_{13}$. Thus, $\hat{G}$ is constant in cases where the sum of the proportionate changes in $\hat{r}_{12}$ and $\hat{r}_{23}$ equals the proportionate change in $\hat{r}_{13}$. In such cases, the magnitude of the proportionate reduction in $\hat{r}_{13}$ equals or exceeds the proportionate reductions in $\hat{r}_{12}$ ($\hat{r}_{23}$). With strict inequalities, $\hat{\rho}_{12}$ and $\hat{\rho}_{23}$ will decline, as shown in the first two equations. If the proportionate reduction in $\hat{r}_{13}$ equals the proportionate reductions in both $\hat{r}_{12}$ and $\hat{r}_{23}$, $\hat{\rho}_{12}$ and $\hat{\rho}_{23}$ would remain constant but $\hat{G}$ would have the same proportionate reduction. In other cases, changes in $\hat{r}_{12}$, $\hat{r}_{23}$, and $\hat{r}_{13}$ will imply changes in $\hat{G}$ as well as a change in either $\hat{\rho}_{12}$ or $\hat{\rho}_{23}$, or changes in both.

Whether the parameters are exactly identified as in Equation 7 or over-identified, 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

6, 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. π is locally identified if $B \xrightarrow{P} B_0$ and $\text{rank}[B_0\, \partial r(\pi)/\partial \pi'] \geq N_{\pi}$, $N_M \geq N_{\pi}$ being a necessary condition. With strict equalities the parameters are exactly identified where $\hat{r} = r(\hat{\pi})$ implicitly defines the estimator that is the same for all B. (See Cameron and Trivedi, 2005.) We employ the identity matrix so that $\hat{\pi}_{MD} = \text{argmin}_{\pi}\ [\hat{r}-r(\pi)]'\,[\hat{r}-r(\pi)]$.[14] 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 $\hat{\sigma}^2_{\eta_{\bullet j}} = \hat{\gamma}_{jj}\,(1-\hat{G}_j)\,\big/\hat{G}_j = (1-\hat{G}_j)\,\hat{\omega}_{jj}$. Rather than estimating γ_{jj} and $\sigma^2_{\eta_{\bullet j}}$ in a second step, the moment conditions included in $r(\pi)$ and r can be expanded to include the moments $\omega_{jj} = \gamma_{jj}\,\big/G_j$ and $\omega_{jj} = (1-G_j)\big/\sigma^2_{\eta_{\bullet j}}$, directly yielding parameter estimates and standard errors for all the parameters of interest.

2.2 Additional Points

To estimate the overall extent of measurement error for a population of students, one only needs test-score descriptive statistics and correlations, an attractive feature of our approach. Additional inferences are possible when student-level data are available. The extent and pattern of heteroskedasticity can be analyzed. The reduced-form model $\tau_{i,j+1} = a_j + c_j\tau_{ij} + u_{i,j+1}$ along with

---

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

the formula $S_{ik} = \tau_{ik} + \eta_{ik}$ implies that $\eta_{i,j+1} - c_j \eta_{ij} + u_{i,j+1} = S_{i,j+1} - a_j - c_j S_{ij}$. With the variances of the expressions before and after the equality being equal, it follows that $\sigma^2_{\eta_{i,j+1}} + c_j \sigma^2_{\eta_{ij}}$ $= V\left(S_{i,j+1} - c_j S_{ij}\right) - \sigma^2_{u_{j+1}}$. Utilizing this expression, $\sigma^2_{\eta_{ij}}$ and $\sigma^2_{\eta_{i,j+1}}$ can be estimated by specifying a functional relationship between the two variances. $\sigma^2_{\eta_{i,j+1}} = \sigma^2_{\eta_{ij}}$ is an example, but is of limited use in that it does not allow for either (i) variation in common factors affecting $\sigma^2_{\eta_{ij}}$ for all students (e.g., a decrease in $\sigma^2_{\eta_{\bullet j}} = E\sigma^2_{\eta_{ij}}$ resulting from an increase in the number of test items) or (ii) variation between $\sigma^2_{\eta_{ij}}$ and $\sigma^2_{\eta_{i,j+1}}$ for individual students holding $\sigma^2_{\eta_{\bullet j}}$ and $\sigma^2_{\eta_{\bullet j+1}}$ constant. The specification $\sigma^2_{\eta_{i,j+1}} \Big/ \sigma^2_{\eta_{\bullet j+1}} = \sigma^2_{\eta_{ij}} \Big/ \sigma^2_{\eta_{\bullet j}}$ or, equivalently, $\sigma^2_{\eta_{i,j+1}} = K_j \sigma^2_{\eta_{ij}}$ with $K_j \equiv \sigma^2_{\eta_{\bullet j+1}} \Big/ \sigma^2_{\eta_{\bullet j}}$ allows for differences across tests in the population mean measurement-error variance. To allow for variation between $\sigma^2_{\eta_{ij}}$ and $\sigma^2_{\eta_{i,j+1}}$ for individual students as well, we assume that $\sigma^2_{\eta_{i,j+1}} =$ $K_j \sigma^2_{\eta_{ij}} + \xi_{ij}$ where the random variable ξ_{ij} has zero mean. This implies the expression shown in Equation 8 where $c_j = \rho_{j,j+1} \Big/ \sqrt{\gamma_{jj} / \gamma_{j+1,j+1}} = \gamma_{j,j+1} / \gamma_{jj}$ and $\sigma^2_{u_{j+1}} = \gamma_{j+1,j+1} + c_j^2 \gamma_{jj} - 2c_j \gamma_{j,j+1}$.

$$\sigma^2_{\eta_{ij}} = \left[V\left(S_{i,j+1} - c_j S_{ij}\right) - \sigma^2_{u_{j+1}} - \xi_{ij}\right] \Big/ \left(K_j + c_j\right) \quad (8)$$

Thus, $\sigma^2_{\eta_{C\,j}} = \left(1/N_C\right) \sum_{i \in C} \left[\left(S_{i,j+1} - \hat{c}_j S_{ij}\right)^2 - \hat{\sigma}^2_{u_{j+1}}\right] \Big/ \left(\hat{K}_j + \hat{c}_j\right)$ can be used to estimate the mean measurement-error variance for a group of students such that $i \in C$. In addition, one can employ the noisy student-level estimate $\hat{\sigma}^2_{\eta_{ij}} = \left[\left(S_{i,j+1} - \hat{c}_j S_{ij}\right)^2 - \hat{\sigma}^2_{u_{j+1}}\right] \Big/ \left(\hat{K}_j + \hat{c}_j\right)$ as the dependent variable in a regression analysis estimating the extent to which $\sigma^2_{\eta_{ij}}$ varies with the level of student achievement or other variables.

The parameters entering the covariance structure also can be estimated without specifying the distributions of τ_{ij} and η_{ij}. However, additional inferences are possible when one assumes particular functional forms. 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 $E\left(\tau_{ij}\left|S_{ij}\right.\right) = (1-G_{ij})\mu_j + G_{ij}S_{ij}$ 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 best-linear-unbiased 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 in our application – the normal density of η_{ij} is $g^j\left(\eta_{ij}\left|\tau_{ij}\right.\right) = \phi\left(\eta_{ij}\Big/\sigma_{\eta_j}(\tau_{ij})\right)\Big/\sigma_{\eta_j}(\tau_{ij})$ where $\phi(\,)$ is the standard-normal density. The joint density of τ_{ij} and η_{ij} is $h^j\left(\eta_{ij},\tau_{ij}\right) = g^j\left(\eta_{ij}\left|\tau_{ij}\right.\right) f^j(\tau_{ij}) = \dfrac{1}{\sigma_{\eta_j}(\tau_{ij})\sqrt{\gamma_{jj}}}\phi\left(\eta_{ij}\Big/\sigma_{\eta_j}(\tau_{ij})\right)\phi\left((\tau_{ij}-\mu_j)\Big/\sqrt{\gamma_{jj}}\right)$ which is <u>not</u> bivariate normal, due to $\sigma_{\eta_{ij}}$ being a function of τ_{ij}. (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 $\sigma^2_{\eta_{ij}} = \sigma^2_{\eta_j}(\tau_{ij})$, the 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 $E\left(\tau_{ij}\left|S_{ij}\right.\right) = \frac{1}{k^j\left(S_{ij}\right)M}\sum_{m=1}^{M}\tau_{mj}^{*}\, g^j\left(S_{ij}-\tau_{mj}^{*}\left|\tau_{mj}^{*}\right.\right)$. Also, $P\left(\tau_{ij}<a\left|S_{ij}\right.\right) = \frac{1}{k^j\left(S_{ij}\right)M}\sum_{\tau_{mj}^{*}<a} g^j\left(S_{ij}-\tau_{mj}^{*}\left|\tau_{mj}^{*}\right.\right)$ is the cumulative posterior distribution of τ_{ij}, which can be used to infer Bayesian confidence (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 $P\left(\tau_{ij}<a\left|S_{ij}\right.\right)=0.90$.

The reduced-form model is a useful tool for estimating the overall extent of test measurement error. Estimation is straightforward and the key requirement (i.e., $E\left(\tau_{i,j+1}\left|\tau_{ij}\right.\right)$ is a linear function of τ_{ij}) will be reasonable in a variety of circumstances. However, this will not always be the case, minimum-competency exams being one possible example. Thus, in assessing the applicability of the reduced-form model in each possible application, one must assess whether the assumptions underlying the reduced-form model are likely to hold. Fortunately, whether $\tau_{i,j+1}$ is a linear function of τ_{ij} can be tested, as demonstrated below.

Finally, it is important to understand that the reduced-form model is only one of the specifications that fall within our general approach. One can carry out empirical analyses employing fully-specified statistical structures for the θ_{ij} (Boyd et al., 2012). Furthermore, rather than inferring the correlation structure based on a set of underlying assumptions, one can start with an assumed covariance structure. For example, one could assume that the universe-score correlations for a set of tests are $\rho_{jk}=\rho^{\left|t_k-t_j\right|}$ or variations on the specification shown in Equation 6. A range of other correlation structure specifications are possible. Again, the reasonableness of

any particular structure will be context specific.

3.0 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.[15]

3.1 Testing Model Assumptions

The simple structure of correlations shown in Equation 6 follows from assuming that $E\left(\tau_{i,j+1} \mid \tau_{ij}\right)$ is linear in τ_{ij}. Fortunately, test score data can be used to assess whether $E\left(\tau_{i,j+1} \mid \tau_{ij}\right)$ is close to being linear. The lines in Figures 1(a) and 1(b) are empirical, nonparametric estimates of $E\left(S_{i,j+1} \mid S_{ij}\right)$ for ELA and math, respectively, showing how eighth-grade scores are related to scores in the prior grade. The bubbles with white fill show the actual combinations of observed 7th and 8th grade scores; the area of each bubble reflects the relative number of students with that score combination.

[15] 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.

The dark bubbles toward the bottoms of Figures 1(a) and 1(b) show the IRT standard errors of measurement (SEMs) for the 7th grade tests (right vertical axis) reported in the test technical reports.[16] 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 conditional 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 correct answer), 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 points aligned in rows and columns in the bubble plot – explains why the reported scale-score conditional 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 the same would not be the case for scale scores.

The fitted nonparametric curves in Figures 1(a) and (b), 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.,

[16] As an example, see CTB/McGraw-Hill (2006).

$\sigma^2_{\eta_{ij}} = \sigma^2_{\eta_{\bullet j}}$, $\forall i$). 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 Figure 1. The explanation and an example are included in the Appendix, along with a discussion of how information regarding the pattern of test measurement error can be used to obtain consistent estimates of the parameters in a 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.[17]

Results from using this approach to analyze the NY test data are shown in Figure 2 for ELA and math, respectively. The thicker, S-shaped curves correspond to the OLS estimate of S_{i8} regressed on S_{i7} 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 Figure 1. The dashed line is a cubic estimate of $E\left(\tau_{i,j+1} \mid \tau_{ij}\right)$ obtained using the approach described in the Appendix 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. It is striking how close the consistent cubic estimates of $E\left(\tau_{i,j+1} \mid \tau_{ij}\right)$ are to being

[17] As discussed below, how the reported SEMs vary with the level of ability is 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 $E\left(\tau_{i,j+1} \mid \tau_{ij}\right)$.

linear.[18] Similar patterns were found for the other grades. 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.

3.2 Estimated Model

Parameter estimates for the reduced-form model and their standard errors are reported in 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 fit the New York data quite well.

The estimated generalizability coefficients in Table 2 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 estimation error or a fundamental shortcoming of our approach, or both, rather than underlying differences in the extent of test measurement error. 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 3 show the reported reliability coefficients for math (solid line) and ELA (dashed line). The lower two lines 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

[18] 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.

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. 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 of 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 are 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 test-score variances and correlations. Additional inferences are possible employing

student-level test-score data. In particular, such data and Equation 8 can be used to estimate $\sigma^2_{\eta_{ij}} = \sigma^2_{\eta_j}(\tau_{ij}) + \zeta_{ij}$ characterizing how the variance of measurement error varies with student ability. (ζ_{ij} is a random variable having zero mean.) Here we specify $\sigma^2_{\eta_j}(\tau_{ij})$ to be a third-order polynomial, compute $\hat{\sigma}^2_{\eta_{ij}}$ using Equation 8 and employ observed scores as estimates of τ_{ij}. Regressing $\hat{\sigma}^2_{\eta_{ij}}$ on S_{ij} would yield inconsistent parameter estimates since S_{ij} measures τ_{ij} with error. However, consistent parameter estimates can be obtained using a two-stage least-squares instrumental-variables estimator where the instruments are the scores for each student not used to compute $\hat{\sigma}^2_{\eta_{ij}}$. In the first stage S_{ij} for grade j is regressed on S_{ik}, $k \neq j, j+1,$ along with squares and cubes, yielding the fitted values $\hat{S}_{ij}$. In turn, $\hat{\sigma}^2_{\eta_{ij}}$ is regressed on $\hat{S}_{ij}$ yielding unbiased estimates of the parameters in $\sigma^2_{\eta_j}(\tau_{ij})$.

The bold solid lines in Figure 4 show $\hat{\sigma}_{\eta_j}(\tau_{ij})$. The dashed lines are the IRT SEMs reported in the test technical reports. Let $\eta_{ij} = \eta^a_{ij} + \eta^b_{ij}$ where η^a_{ij} is the test-instrument-based measurement error, η^b_{ij} is the measurement error from other sources and $\sigma^2_{\eta_{ij}} = \sigma^2_{\eta^a_{ij}} + \sigma^2_{\eta^b_{ij}} + 2\,\mathrm{cov}(\eta^a_{ij}, \eta^b_{ij})$. For a particular test, $\sqrt{\hat{\sigma}_{\eta_j}(\tau_{ij}) - \hat{\sigma}_{\eta^a_j}(\tau_{ij})}$ can be used to estimate of $\sigma_{\eta^b_j}(\tau_{ij})$, assuming that η^a_{ij} and η^b_{ij} are uncorrelated. The thin lines in Figure 4 show these "residual" estimates, $\hat{\sigma}_{\eta^b_j}(\tau_{ij})$. The range of ability levels for which $\hat{\sigma}_{\eta^b_j}(\tau_{ij})$ is shown roughly corresponds to our estimates of the ranges containing 99 percent of actual abilities. In Figure 4(b), for example, it would be the case that $P(608 \leq \tau_{i7} \leq 715) = 0.99$ if our estimates of the ability distribution were correct.

There are *a priori* explanations for why $\sigma_{\eta_j^a}(\tau_{ij})$ would be a u-shaped function for IRT-based scale-scores and would be an inverted-u-shape function in the case of raw scores. A speculative, but somewhat believable, hypothesis is that the variance of the measurement error unrelated to the test instrument is relatively constant across ability levels. However, this begs the question as to whether the relevant "ability" is measured in raw-score or scale-score units. If the raw-score measurement-error variance were constant, the nonlinear mapping from raw-scores to scale-scores would imply a u-shaped scale-score measurement-error variance – possibly explaining the u-shaped patterns of $\hat{\sigma}_{\eta_j^b}(\tau_{ij})$ in Figure 4. Whatever the explanation, values of $\hat{\sigma}_{\eta_j^a}(\tau_{ij})$ and $\hat{\sigma}_{\eta_j^b}(\tau_{ij})$ are roughly comparable in magnitude and vary similarly over a wide range of abilities. We have less confidence in the estimates of $\hat{\sigma}_{\eta_j^b}(\tau_{ij})$ for extreme ability levels. Because $\hat{\sigma}_{\eta_j^b}(\tau_{ij})$ is the square root of a residual, computed values of $\sqrt{\hat{\sigma}_{\eta_j}(\tau_{ij}) - \hat{\sigma}_{\eta_j^a}(\tau_{ij})}$ can be quite sensitive to estimation error when $\hat{\sigma}_{\eta_j}(\tau_{ij}) - \hat{\sigma}_{\eta_j^a}(\tau_{ij})$ is close to zero. Here it is relevant to note that for the case corresponding to Figure 4(a), we estimate that only 1.8 percent of students have universe scale-scores exceeding 705. In Figure 4(d), the universe-scores of slightly less than five percent of students exceed 720.

3.3 Inferences Regarding Universe Scores and Universe Score Gains

Observed scores typically are used to directly estimate students achievement and achievement 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 lines in Figure 5 show the values of $\hat{E}\left(\tau_{ij} \mid S_{ij}\right)$ for 5th and 7th grade ELA and math. Referencing the 45° 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 737, 38 point below the observed score. We estimate that 80 percent of students scoring 775 have universe scores in the range 719-757; $P\left(718.8 < \tau_{ij} < 757.2 \middle| S_{ij} = 775\right) = 0.80$. In this case, the observed score is 18 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 (638, 668), respectively. The credible bounds range for a 775 score is 30 percent larger than that for a score of 650.

Utilizing test scores to directly estimate students' abilities clearly 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.[19] For the fifth-grade math exam, the RMSE associated with using $\hat{E}\left(\tau_{ij} \middle| S_{ij}\right)$ to estimate students' abilities is 14.9 scale-score points. In contrast, the RMSE associated with using S_{ij} is 17.2, 15 percent larger. This difference is meaningful given that $\hat{E}\left(\tau_{ij} \middle| S_{ij}\right)$ 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,

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

the RMSE for $\hat{E}\left(\tau_{ij}\left|S_{ij}\right.\right)$ and S_{ij} are 14.9 and 15.1, respectively. However, for ability levels below 620 the RMSEs are 13.4 and 20.9, respectively, the latter being 57 percent larger. For students whose actual abilities are greater than 710, the RMSE associated with using S_{ij} to estimate τ_{ij} is 26.6, which is 62 percent larger than the RMSE for $\hat{E}\left(\tau_{ij}\left|S_{ij}\right.\right)$. By accounting for 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 6 shows the posterior-mean universe-score change in math between grades five and six conditional on the observed score change.[20] 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 12.7. Eighty percent of all students having a 40-point score increase are estimated to have actual universe score changes falling in the interval -2.3 to 27.0. It is noteworthy that for the full range of score changes shown (± 50 points), the 80-percent credible bounds include there actually being no change in ability.

---

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

Many combinations of scores yield a given score change; a score increase from 590 to 630 implies a 40-point change as does an increase from 710 to 750. Figure 6 corresponds to the case where one knows the score change but not the pre- and post-scores. 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 7 shows the posterior-mean universe-score change and credible bounds for various scores consistent with 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 10.3 point universe-score increase on average, with 80 percent of such students having actual changes in ability in the interval (-11.4, 31.7). (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 fifth-grade score is between 615 and 645. For students scoring at the fifth-grade proficiency cut-score (648), the average universe-score gain is 19.6 with 80 percent having actual changes in the interval (-1.15, 37.4).

A striking result in Figure 7 is that the posterior mean universe-score change, $\hat{E}\left(\tau_6 - \tau_5 \middle| S_5, S_6\right) = \hat{E}\left(\tau_6 \middle| S_5, S_6\right) - \hat{E}\left(\tau_5 \middle| S_5, S_6\right)$, is substantially smaller than the observed-score change. Consider $\hat{E}\left(\tau_6 - \tau_5 \middle| S_5 = 710, S_6 = 750\right) = 10.3$ which is substantially smaller than the 40-point score increase. First, $\hat{E}(\tau_6 \middle| S_6 = 750) = 734.0$ is 16 points below the observed score due to the Bayes shrinkage toward the mean. $\hat{E}\left(\tau_6 \middle| S_5 = 710, S_6 = 750\right) = 729.5$ 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 $E\left(\tau_6 \middle| S_5, S_6\right)$ would equal $E(\tau_6 \middle| S_6)$ if either $\sigma^2_{\eta_{.6}} = 0$ or $\rho_{56} = 0$.) Similar logic holds for the

fifth grade. $\hat{E}(\tau_5 | S_5 = 710) = 705.3$ is less than 710 because the latter is substantially above $E\tau_{i5}$. However, $\hat{E}(\tau_5 | S_5, S_6) = 719.2$ is meaningfully larger than $\hat{E}(\tau_5 | S_5) = 707.5$ and larger than $S_5 = 710$, reflecting the effect of $S_6 = 750$ being 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; $\hat{E}(\tau_6 | S_5, S_6) - \hat{E}(\tau_5 | S_5, S_6) = 729.5 - 719.2 = 10.3$. 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 as large were the ability levels uncorrelated, $\hat{E}(\tau_6 | S_6) - \hat{E}(\tau_5 | S_5) = 734.0 - 705.3 = 28.7$.

4.0 Conclusion

We show that there is a credible approach for estimating the overall extent of test measurement error using nothing more than test-score variances and non-zero correlations for three or more tests. Our approach can be used in a variety of settings and is a meaningful generalization of the test-retest method. First, substantially relaxing the requirement that the tests be parallel, our approach does not require that the tests be vertically scaled. Second, as in the case of congeneric tests analyzed by Joreskog (1971), the method allows the extent of measurement error to differ across tests. Third, the approach only requires that there is some persistence (i.e., correlation) in ability across the test administrations, a requirement far less restrictive than the requirement that ability remain constant. However, as with the test-retest framework, the applicability of our approach crucially depends upon whether a sound case can be made that the tests to be analyzed meet the necessary requirements, including assumptions regarding the structure of universe-score correlations.

As the analysis of Rogosa and Willet (1985) makes clear, commonly observed covariance

patterns can be consistent with quite different models of achievement growth; the underlying correlation structures implied by different growth models can yield universe-score correlation patterns and values that are indistinguishable. Fortunately, rather than attempting to identify the actual underlying structural model, our goal is to estimate the extent of measurement error as well as the values of the universe-score variances and correlations. We conjecture that the inability to distinguish between quite different underlying universe-score correlation structures actually is advantageous given this goal in that the estimated extent of test measurement error will be robust to a range of underlying covariance structure misspecifications. This conjecture is consistent with our finding that estimates of the extent of measurement error are quite robust across a range of model specifications. Monte Carlo simulations using a wide range of underlying covariance structures could provide more convincing evidence, but goes beyond the scope of this paper.

We illustrate the general approach employing a model of student achievement growth in which 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 ($\beta_{j-1} < 1$) that can differ across grades. An additional assumption is needed regarding the stochastic properties of θ_{ij}. In the above application, we have employed a reduced-form specification where $E\left(\tau_{i,j+1} \mid \tau_{ij}\right)$ is a linear function of τ_{ij}, an assumption that can be tested. Boyd et al. (2012) discuss three fully specified structural models which also assume that $\tau_{ij} = \beta_{j-1}\tau_{i,j-1} + \theta_{ij}$. In addition, rather than inferring the correlation structure based on a set of underlying assumptions, one can start with an assumed covariance structure where there are a range of possibilities one could employ, depending upon the tests being analyzed.

Estimation of the overall extent of measurement error for a population of students requires that one have test-score descriptive statistics and correlations; neither student-level test scores nor

assumptions regarding functional forms for the distribution of either abilities or test measurement error are needed. However, one can explore the extent and pattern of measurement error heteroskedasticity employing student-level data. Also, standard distributional assumptions (e.g., normality) allow one to 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 of the student's actual ability. Similar statistics for universe-score gains also can 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-scoring students. However, the bias is eliminated 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. Using $E\left(\tau_{ij}\left|S_{ij}\right.\right)$, rather than S_{ij}, to estimate τ_{ij} is one example. Judging the magnitudes of the effects of different causal factors relative to either the standard deviation of ability or the standard deviation of ability gains is another. Bloom et al. (2008) discuss the desirability of assessing the magnitudes of effects relative to 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 measured 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.

Estimates of the extent and pattern of test measurement error can also be used to assess the precision of a variety of measures based on test scores, including binary indicators of student proficiency, teacher- and school-effect estimates and accountability measures (e.g., AYP) more generally. 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 methodological and substantive implications. Methodologically it shows that the total measurement-error variance can be estimated without employing the limited and costly test-retest strategy. Substantively, it shows that the total measurement error is substantially greater than that measured using the split-test method, suggesting that much empirical work has been underestimating the effect sizes of interventions that affect student learning.

References

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. (2009) “Test Scaling and Value-Added Measurement,” *Education Finance and Policy* 4(4), 351-383.

Boyd, D., H. Lankford, S. Loeb and J. Wyckoff (2012) "Measuring Test Measurement Error: A General Approach", National Bureau of Economic Research working paper, W18010.

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.

Cronbach, 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.

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

Haertel, E. H. (2006) “Reliability,” in *Educational Measurement*, 4th ed., R. L. Brennan, ed., Praeger.

Joreskog, K. G. (1971) “Structural Analysis of Covariance and Correlation Matrices,” *Psychometrika* 43(4) 443-477.

Joreskog, K. G. (1971) “Statistical Analysis of Sets of Congeneric Tests,” *Psychometrika* 36 (2) 109-133.

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.

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.

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

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.

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

+ 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)
	$\sigma^2_{S_{\bullet j}}$	$\hat{\sigma}^2_{\eta_{\bullet j}}$	$\hat{\gamma}_{jj} = \hat{G}_j \sigma^2_{S_{\bullet j}}$	$\hat{\sigma}^2_{\Delta S_{\bullet j}}$	$\hat{\sigma}^2_{\Delta \eta_{\bullet j}}$	$\hat{\sigma}^2_{\Delta \tau_{\bullet j}}$	$\hat{G}^{\Delta}_j = \hat{\sigma}^2_{\Delta \tau_{\bullet j}} \Big/ \hat{\sigma}^2_{\Delta S_{\bullet j}}$
ELA							
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 Scores on Scores in Grade 7, Bubble Graph Showing the Joint Distribution of Scores and Standard-Error of Measurement for 7th Grade Scores

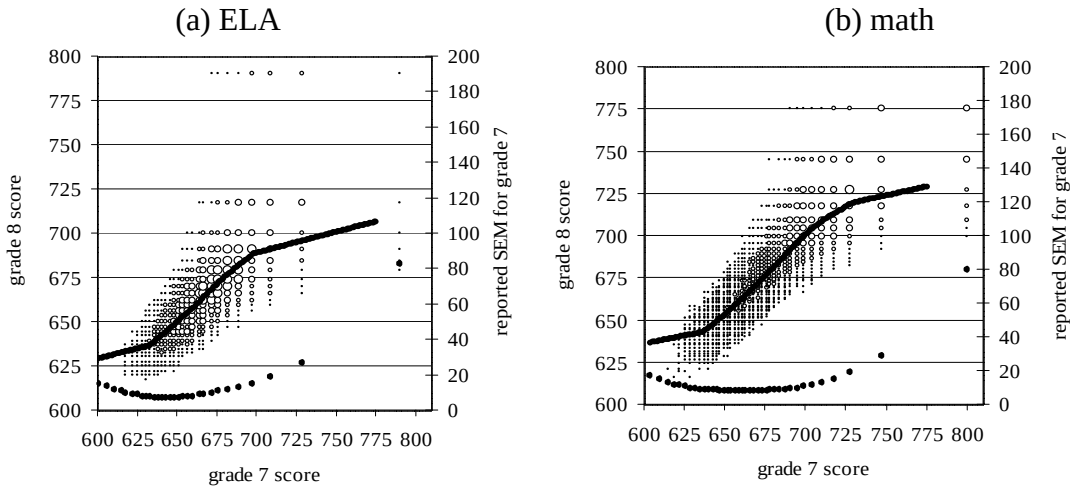


Figure 2: Cubic Regression Estimates of $E\left(S_{i,j+1} \mid S_{ij}\right)$ as well as consistent estimates of cubic and linear specifications of $E\left(\tau_{i,j+1} \mid \tau_{ij}\right)$, Grades 7 and 8

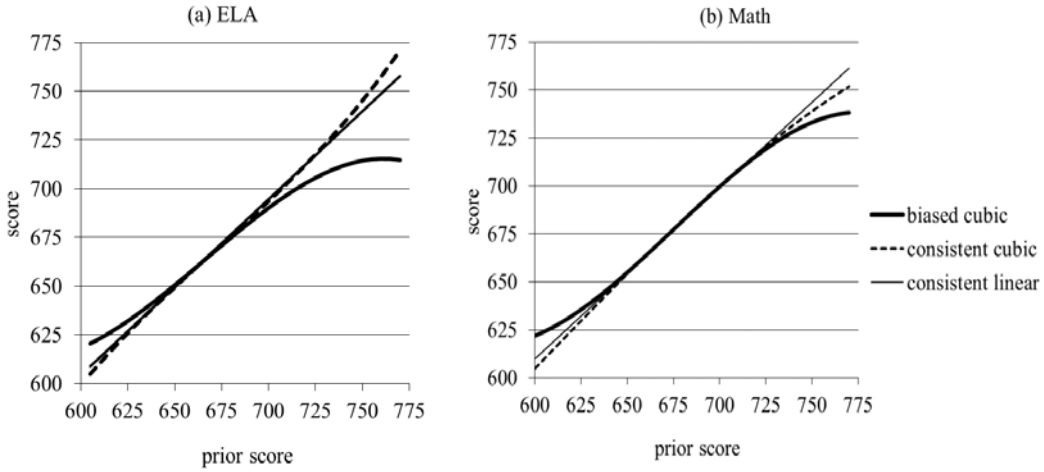


Figure 3: Generalizability and Reliability Coefficient Estimates for New York Math and ELA Exams by Grade

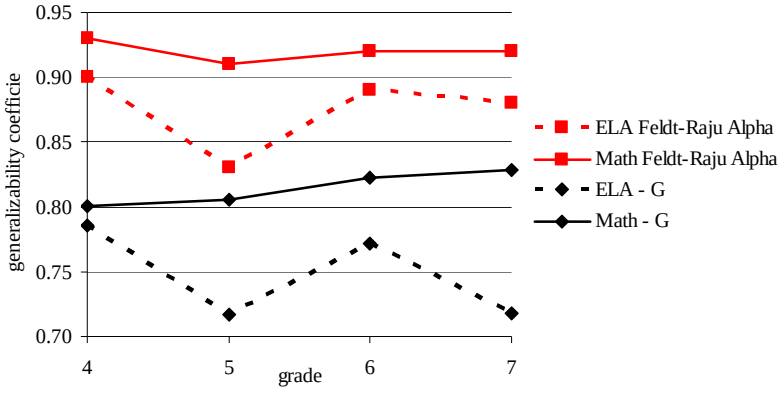


Figure 4 Estimated Standard Errors of Measurement Reported in Technical Reports, $\hat{\sigma}_{\eta_j^a}$, Estimates for the Measurement Error from All Sources, $\hat{\sigma}_{\eta_j}$, and Estimates for the Residual Measurement Error, $\hat{\sigma}_{\eta_j^b}$

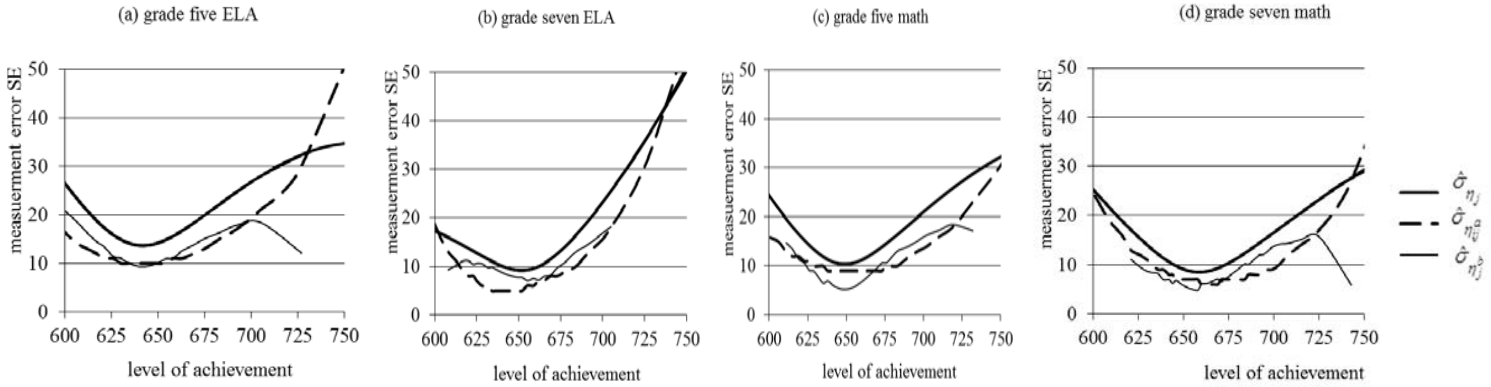


Figure 5 Estimated Posterior Mean Ability Level Given the Observed Score and 80-Percent Bayesian Confidence Bounds, Grades 5 & 7 ELA and Math

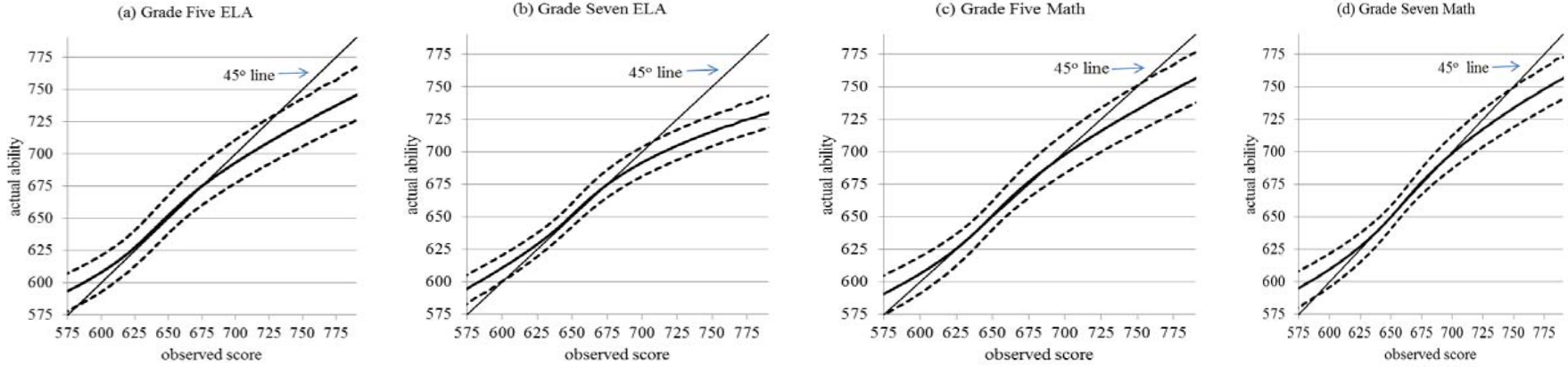


Figure 6 Estimated Posterior Mean Change in Ability Given the Score Change and 80-Percent Credible Bounds, Grades 5 and 6 Mathematics

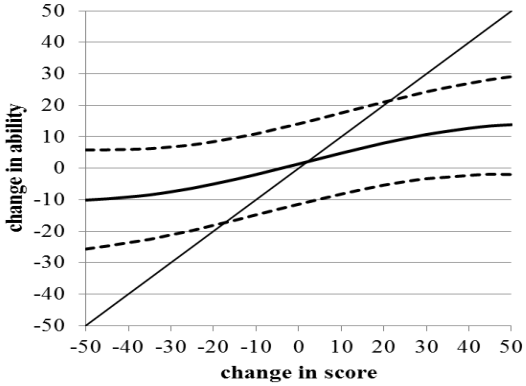
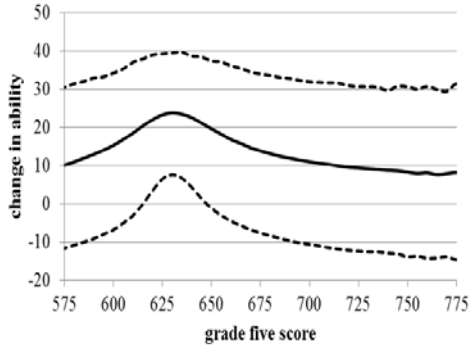


Figure 7 Estimated Posterior Mean Change in Ability for the Observed Scores in Grades Five and Six Mathematics for $S_6 - S_5 = 40$ and 80-Percent Credible Bounds



Appendix

Measurement error can 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}. $E\left(\tau_{i,j+1} \mid \tau_{ij}\right) = \beta_0 + \beta_1 \tau_{ij}$ implies that $\tau_{i,j+1} = \beta_0 + \beta_1 \tau_{ij} + \upsilon_{i,j+1}$ where $E\upsilon_{i,j+1} = 0$ and $E\tau_{ij}\,\upsilon_{i,j+1} = 0$. With $S_{i,j+1} = \tau_{i,j+1} + \eta_{i,j+1}$, $S_{i,j+1} = \beta_0 + \beta_1 \tau_{ij} + \eta_{i,j+1} + \upsilon_{i,j+1}$ and, in turn, $E\left(S_{i,j+1} \middle| S_{ij}\right) = \beta_0 + \beta_1 E\left(\tau_{ij} \middle| S_{ij}\right)$. Thus, the nonlinearity of $E\left(S_{i,j+1} \middle| S_{ij}\right)$ is determined by whether $E\left(\tau_{ij} \middle| S_{ij}\right)$ is nonlinear in S_{ij}. Consider the case where $\tau_{ij} \sim N(\mu_j, \sigma^2_{\tau_j})$ and $\eta_{ij} \sim N(0, \sigma^2_{\eta_{ij}})$ and the related discussion on page 16 in section 2.2. 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 so that $E\left(\tau_{ij} \middle| S_{ij}\right) = (1 - G_{ij})\mu_j + G_{ij} S_{ij}$, implying that $E\left(S_{i,j+1} \middle| S_{ij}\right)$ is also linear in S_{ij}. Thus, it is not measurement error *per se* that implies $E\left(S_{i,j+1} \mid S_{ij}\right)$ is nonlinear. Rather, $E\left(S_{i,j+1} \mid S_{ij}\right)$ is nonlinear in S_{ij} when 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 Figure 1, $E\left(S_{i,j+1} \mid S_{ij}\right)$ is an S-shaped function of S_{ij}, even when $E\left(\tau_{i,j+1} \mid \tau_{ij}\right)$ is linear in τ_{ij}.

When $\sigma_{\eta_{ij}} = \sigma_{\eta_j}(\tau_{ij})$, S_{ij} and τ_{ij} are not bivariate normal, $E\left(\tau_{ij} \middle| S_{ij}\right)$ can be computed using simulation, as discussed in Section 2.2. Consider an example roughly consistent with the patterns found for the NYC test scores where $\tau_{ij} \sim N(670, 900)$ and $\eta_{ij} \sim N\left(0, \sigma^2_{\eta}(\tau_{ij})\right)$ with $\sigma_n(\tau_{ij}) = \sigma_o + \gamma(\tau_{ij} - \mu_j)^2$ and $\sigma_{n_j} = E\sigma_n(\tau_{ij}) = \sigma_o + \gamma\sigma^2_{\tau_j} = 15$. The three cases in Figure B.1 differ with respect to the degree of heteroskedasticity: the homoskedastic case ($\sigma_o = 15$ and $\gamma = 0.0$), moderate heteroskedasticity ($\sigma_o = 12$ and $\gamma = 0.00333\cdots$) and more extreme

heteroskedasticity ($\sigma_o = 3$ and $\gamma = 0.01333\cdots$). The simulated values of $E\left(S_{i,j+1}\middle|S_{ij}\right) = \beta_0 + \beta_1 E\left(\tau_{ij}\middle|S_{ij}\right)$ are shown in Figure B.2 for each cases, with $\beta_0 = 0$ and $\beta_1 = 1$. $E\left(S_{i,j+1}\middle|S_{ij}\right)$ is linear in the homoskedastic case and the degree to which $E\left(S_{i,j+1}\middle|S_{ij}\right)$ is S-shaped depends upon the extent of this particular type of heteroskedasticity.

Knowing that the S-shape patterns of $E\left(S_{i,j+1}\middle|S_{ij}\right)$ in Figure 1 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 $\lambda_{ij}^k \equiv E\left(\tau_{ij}^k\middle|S_{ij}\right)$, $k = 1,2,3,4$, were known for each student, regressing $S_{i,j+1}$ on λ_{ij}^1, $\lambda_{ij}^2, \lambda_{ij}^3$, and λ_{ij}^4 would yield consistent estimates.[21]

Computing λ_{ij}^k, $k = 1,2,3,4$, 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 $\hat{\lambda}_{ij}^k = \hat{E}\left(\tau_{ij}^k\middle|S_{ij}\right)$ accounting for the meaningful measurement-error heteroskedasticity reflected in the reported SEMs[22], even though this does not account for other sources of measurement error. Computation of $\hat{E}\left(\tau_{ij}^k\middle|S_{ij}\right)$ also requires an estimate of $\sigma_{\tau_j}^2$ which can be obtained by solving for $\hat{\sigma}_{\tau_j}^2$ implicitly defined in

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

[22] 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.

$\hat{\sigma}^2_{\tau_j} = \hat{\sigma}^2_{S_j} - \hat{\sigma}^2_{\eta_{\square j}} = \hat{\sigma}^2_{S_j} - \int \sigma^2_\eta(\tau)\, f\left(\tau \middle| \hat{\mu}_j, \hat{\sigma}^2_{\tau_j}\right) d\tau = \hat{\sigma}^2_{S_j} - \frac{1}{M}\sum_m^M \sigma^2_\eta\left(\tau^*_{mj}\right)$. Here the integral is computed using Monte Carlo integration with importance sampling where the τ^*_{mj} are random draws from the distribution $N\left(\hat{\mu}_j, \tilde{\sigma}^2_{\tau_j}\right)$ and $\tilde{\sigma}^2_{\tau_j}$ is an initial estimate of $\sigma^2_{\tau_j}$. This yielded an updated value of $\tilde{\sigma}^2_{\tau_j}$ which can be used to repeat the prior step. Relatively few iterations are needed for converge to the fixed-point – our estimate of $\sigma^2_{\tau_j}$ The estimate $\hat{\sigma}^2_{\tau_j}$ 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 $\hat{\lambda}_{i4}$.

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

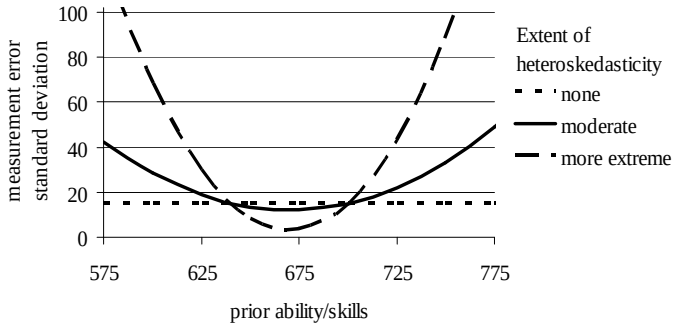


Figure B.2 How the Relationship Between $E\left(S_{i2}\middle|S_{i1}\right)$ and S_{i1} Varies with the Degree of Heteroskedasticity