The Measurement of Student Ability in Modern Assessment Systems

Brian Jacob*

University of Michigan and NBER

Jesse Rothstein†

University of California, Berkeley and NBER

January 2016

1 Introduction

Human capital is central to modern economics. In early empirical work, researchers used the number of years of education as a proxy for human capital. But recent research frequently measures variation in worker skill using cognitive test scores. These scores are used as measures of human capital in explaining wages and other employment outcomes (Neal and Johnson, 1996) and, increasingly, as outcome measures in evaluations of programs and policies aimed at improving human capital accumulation. Examples of the latter include Angrist et al. (2011) on charter schools; Jackson et al. (2014) on teachers; Krueger (1999) on class size; Lafortune et al. (2015) on school finance; and Heckman et al. (2010) and Shapiro et al. (2010) on pre-school. Since the introduction of the federal school accountability legislation No Child Left Behind in 2002, there has been a proliferation of student testing, accelerating the use of cognitive ability measures by economists.

*bajacob@umich.edu

†rothstein@berkeley.edu

Economists typically take cognitive test scores from pre-existing surveys or data sets, without much thought about how these measures are generated. We suspect that many imagine that test scores are noisy but unbiased measures of an individual's true ability - e.g., the count or fraction of items answered correctly. While this is sometimes the case, there are a wide variety of methods used to construct the test scores reported in commonly used data sets. These methods are not interchangeable, and they have important implications for the secondary use of the resulting scores.

For example, the model used to generate student test scores in the National Assessment of Educational Progress (NAEP; also known as "the Nation's Report Card") incorporates a student's background characteristics along with his or her responses to test items. Race is one of the many background variables contained in the NAEP model. This means that if a black and white student respond identically to the same set of NAEP assessment items, the reported ability for the black student will be lower than that of the white student, reflecting the fact that, on average, black students score lower than white students on this assessment. The NAEP accounts for this modeling decision when it reports summary statistics such as the black-white test score gap, but, as we explain in detail below, this modeling choice can introduce important biases into other secondary analyses involving NAEP data. Many other assessments, including those used in the longitudinal studies created by the National Center for Education Statistics (such as the Early Childhood Longitudinal Study, or ECLS), report "shrunken" scores that, if used naively, will lead secondary analysts to understate between-group gaps in achievement.

The construction of test scores also has important implications for the standard economic approaches to dealing with measurement error in explanatory variables. We show that standard errors-in-variables or instrumental variables approaches are unlikely to properly account for the measurement error contained in scores derived from modern assessments. Finally, the scaling of test scores also presents serious challenges to the unaware analyst. Economists frequently attempt to abstract from scaling issues by constructing z-scores, but, as we discuss below, z-scores constructed from differently-scaled and measured assessments may not be comparable.

Our goal in this paper is to familiarize applied economists with the construction and properties of common cognitive score measures, and with the psychometrics behind them. Information about how scores are constructed is often buried deep in technical manuals, if presented at all, and inaccessible to those without a background in psychometrics (the field concerned with the theory and methodology of psychological measurement). As a consequence, economists and other applied researchers frequently mis-use test score measures, with potentially serious consequences for their analyses.

To frame our discussion, it is useful to distinguish between several different types of tests.

- Tests designed to support research. These tests are typically administered as part of surveys conducted by a federal statistical agency, most often the National Center for Educational Statistics (NCES). They include the achievement tests given to children in the Early Childhood Longitudinal Survey (ECLS) and tests given to large samples of 4th, 8th, and 12th graders every year or two as part of the National Assessment of Educational Progress (NAEP) program. They are generally not closely tied to an individual curriculum, are not reported back to the students or their teachers or parents, and are designed with a relatively high degree of attention to psychometric and statistical details.

- School-based assessments. There are a wide variety of tests administered to students in schools as part of their regular operations. These range from state-mandated accountability tests to formative assessments administered by teachers periodically to tests administered to diagnose learning disabilities.[1] Important subcategories include annual tests used to hold students, schools and teachers accountable for their performance as, for example, mandated by the federal No Child Left Behind legislation.

  - Accountability tests. Test scores are increasingly used to hold students, schools, and teachers accountable for their performance. Tests to support this are often, in accordance with the federal No Child Left Behind Act of 2001, designed to measure whether

[1] For a more comprehensive discussion of student testing in schools, see a recent report by the Council of the Great City Schools (Hart et al., 2015).

a student passes some fixed "proficiency" standard. They vary from state to state, and are often constructed by external vendors (e.g., Pearson). Well-known tests of this sort are the Iowa Test of Basic Skills (ITBS) and the Stanford Achievement Test (known in various versions as, e.g., the SAT9 or SAT10). Scores on these tests, extracted from administrative databases, are the basis of much recent research in the economics of education (e.g., Chetty et al. 2014a,b; Figlio and Rouse 2006; Neal and Schanzenbach 2010.

  - Interim or benchmark assessments. These assessments are typically administered two to three times per year (fall, winter, and spring) in core academic subjects, most often in elementary grades. Today students often take these assessments online, and some are adaptive, meaning that students who answer the initial items correctly are given more difficult subsequent items. Examples of such assessments include the Northwest Evaluation Association's Measures of Academic Progress (NWEA-MAP), Scholastic Reading/Math Inventory (SRI/SMI), the Developmental Reading Assessment (DRA) and the Dynamic Indicators of Basic Early Literacy (DIBELS).
  - Diagnostic tests. A third category of tests is designed to identify children in need of special services, or to diagnose specific learning disabilities (e.g., dyslexia). These are widely administered, but are not commonly used in economics research.

- College entrance tests, such as the ACT and SAT.
- Occupational certification tests. As occupational certification and licensing has become more prevalent in the U.S. labor market (Kleiner and Krueger, 2013), tests have become more prominent parts of the licensing process. Exams are required for licensure in fields as diverse as cosmetology, interior design, crane operation, tree trimming, home entertainment installation, and auctioneering.[2]

[2]http://ij.org/wp-content/uploads/2015/04/licensetowork1.pdf.

We focus primarily on the first two types of tests, which while apparently similar often have very different psychometric properties. Much of our discussion applies to the others as well, however.

The remainder of the article proceeds as follows. In Section 2, we discuss the issue of assigning a quantitative scale to latent student ability. Section 3 discusses the models used to measure ability on the chosen scale. Section 4 explores the implications of the choice of measurement model for secondary analysis of the resulting scores. We conclude by describing the next generation of assessments that have been developed in conjunction with the Common Core State Standards (CCSS). We also attempt to provide applied researchers with practical guidance for working with cognitive ability measures.

2 Scaling

In this section, we discuss several issues associated with the scaling of assessment data. Test scores are potentially valuable for empirical researchers, as they quantify student traits that are otherwise hard to measure. But using them requires understanding what they represent, which involves understanding the scale assigned to the latent ability measure.[3] The commonly used assessments described above report scores on a variety of different scales. IQ tests are traditionally scored so that the distribution is centered at 100, with a standard deviation of about 15 (though see Flynn 1987, 2009). The SAT college entrance exam is scaled to have a minimum score of 200, a maximum score of 800, a mean of around 500, and a standard deviation around 100 on each subtest; its competitor, the ACT, uses integers between 1 and 36 for each of four subjects, with means around 21 and standard deviations around 6. The National Assessment of Educational Progress (NAEP) uses two scales: Scale scores range from roughly 100 to 400, with standard deviations around 30, and in addition students are also assigned to three categories (basic, proficient, and advanced).

Test scales are generally arbitrary in their locations and ranges. There is no reason that the

[3]Throughout this paper, we use "ability," "proficiency," "achievement," and "aptitude" interchangeably to refer to the latent trait that governs test performance. Although some testing discussion – particularly that pertaining to IQ tests – treats them as distinct, such distinctions (which are hotly disputed) are not important for our purposes.

College Board could not assign the lowest (highest) performing student on the SAT a score of 100 (1000) or that NAEP couldn't increase or reduce the number of gradations made. Moreover, even the distribution within the chosen range is based on arbitrary scaling decisions that vary across tests – e.g., holding the number of categories fixed, NAEP could raise or lower the thresholds among them with no less fealty to the underlying construct, and the College Board could rescale SAT scores so that the achievement level that is now scored at 500 would be scored much closer to the level currently scored at 600 than to the level currently scored at 400.

2.1 Interval or ordinal

Researchers using test scores generally treat them as an interval scale, meaning that a one unit change in a student's score at any point on the distribution reflects the same change in the underlying knowledge or skill the assessment is intended to measure. This is implicit in any analysis based on averages of scores. For example, a comparison of the mean SAT scores between a treatment and control group depends on the assumption that an improvement of one student's score from, say, 300 to 350 represents as much learning as an improvement of another student from 700 to 750. However, as many scholars have pointed out, there is generally no basis for interpreting test scales as having an interval property (Stevens, 1946; Thorndike, 1966; Bond and Lang, 2013). Like utility (and unlike income or temperature), measured achievement is best thought of as being ordinal but not cardinal.

This fact has important implications for virtually all empirical analyses that involve test scores. For the purpose of illustration, we present a very simple example from Bond and Lang (2013) to show how arbitrary scaling decisions can have dramatic effects on estimates of an oft-studied statistic in education research: the black-white test score gap. Consider a test of three, progressively difficult skills that are cumulative in the sense that a student must master skill 1 before mastering skill 2, and skill 2 before skill 3. Students can thus answer zero, one, two or all three test items correctly. Bond and Lang (2013) ask us to consider a population with two black students and two white students, in which the black students correctly answer 0 and 2 items and the white students answer 1 and 2 items correctly. The count of correct items is known in

psychometrics as the "raw score." The average raw score for black students is thus 1, while that for white students is 1.5. Hence, the gap in mean raw scores is 0.5 points, or 0.6 standard deviations.

The raw score metric assumes that each skill represents the same "amount" of knowledge (has the same marginal value), a belief that may not be true. Suppose that the skills are (in order) the the ability to recite the alphabet, the ability to recognize letters and ability to read fluently. In this case, one might consider the incremental knowledge represented by skill two to be quite small, but the steps from zero to one skills or from two to three to be extremely large. If we assume the difference between zero and one skills is much larger than that between one and two, the gap approaches 1 point, or 1.15 standard deviations. By contrast, we assume the difference in knowledge between zero and one skill is arbitrarily small and that between one and two is large, the test score gap will approach zero.

This problem is even worse if one considers changes over time. Assume that each student progresses exactly one skill level during the course of a year, so that at the end of the year the black students correctly answer 1 and 3 items correctly, and the white students answer 2 and 3 items correctly. If each skill is assumed to represent the same amount of instructional input, then the change in the raw test score gap measures the gap in inputs. This gap has remained unchanged at 0.5 points, so we might conclude that the quality of the instruction available to black and white students is the same. Using the two alternative scalings described above, however, we would conclude that the gap falls from 1.15 to 0 standard deviations, or increases from 0 to 1.15 standard deviations, with very different implications for our assessments of instructional quality. Bond and Lang (2013) show that empirical estimates of the black-white gap in achievement growth across grades are extremely sensitive to transformations of the test score, and that this sensitivity varies across test and grade level. Depending on the transformation and assessment used, they find that estimates of the *change* in the black-white test score gap between kindergarten and third grade range from 0 to 0.6 standard deviations.

To take another example, consider value-added estimates of teachers' impacts on their students' achievement. Even if we set aside questions about the causal interpretation of these esti-

mates (Rothstein, 2010, 2015; Chetty et al., 2014a,b), any comparison of the value-added across teachers with students having different baseline scores rests, implicitly, on an assumed interval scale. Without this, one cannot compare the impact of a teacher who works with very low scoring students and raises their scores by ten (scale score) points to the impact of a peer who raises by fifteen points the scores of her students with higher baseline scores, or even average the growth of one student whose score rises by ten points over the year and that of a classmate who starts at a very different place but gains only five points.

The ordinality of test scores thus poses a major challenge for those working with test score data. There are several options with regard to scaling. The first approach, favored by many psychometricians and education researchers, is to utilize the scale that comes from the model used to develop and score the assessment. The choice of model defines a scale for the achievement parameter and many psychometricians treat this scale as interval. In practice, however, most analyses of test scores as interval measures are based on an implicit assumption is that a one unit movement on the scale reflects an "equal" amount of knowledge, or an equal amount of effective instructional input, at any point in the scale. There is no particular reason to think that the functional form used to generate the assessment scores corresponds to equal measures of effective instructional inputs. Indeed, absent an operational definition of effective instructional inputs – which of course is what the test is often used to estimate – it is not even clear how one might evaluate the claim that a proposed scale has an interval property.

A second approach, advocated byBond and Lang (2013), is to simply treat test scores as ordinal measures, and limit conclusions to those that are robust to arbitrary monotonic transformations of the test score. For example, one way to describe the differences in two distributions in a metric-free way is to construct percentile-percentile (P-P) plots, which plot the cumulative distribution of group A (e.g., black students) against the cumulative distribution of group B (e.g., white students). If the two distributions are identical, the plot will lie on the 45-degree line. The extent to which the plot deviates from this line can be used to construct a scale-free measure of the difference in performance across the two populations. For example, Reardon (2008) calculates the probabil-

ity that a randomly chosen black student will have a test score higher than the randomly chosen white student, which equals the percentile in the white distribution where the average black student would fall. Ho (2009) and Ho and Haertel (2006) describe how this information can be converted to a standardized metric-free gap measure (though even here this may or may not correspond to the gap in educational inputs received).

This reliance on the ordinal property of test scores underlies a popular approach to calculating teacher value-added known as the Colorado Growth Model (also referred to as the student growth percentile model). Here the researcher assigns student i a "growth percentile" which corresponds to her percentile in the distribution of test scores in year t among the sample of students who had the same test score as student i in year t-1. The median growth percentile of a teacher's students provides the measure of teacher effectiveness, though again there is no assurance that a given increment to a teacher's median growth percentile is equally easy to achieve at all points in the teacher or student distribution. Barlevy and Neal (2012) propose a similar approach as part of a teacher accountability and compensation system.

While a focus on the ordinal nature of test scores is clearly more defensible from a psychometric perspective, it comes with important costs as it limits the questions that can be answered in research and policy evaluation. A third approach is to translate scores into units of another measure that we *are* willing to assume is interval, such as adult earnings or educational attainment (Cunha and Heckman, 2006; Cunha et al., 2010; Bond and Lang, 2015). Thus, an attainment-scaled test score would simply be the average eventual educational attainment of all students with a particular score:

$$y(q) = \frac{\sum_i \mathbf{1}(t_i = q) Y_i}{\sum_i \mathbf{1}(t_i = q)} \tag{1}$$

where Y_i is the adult attainment of student i with earlier test score t_i and $\mathbf{1}(t_i = q)$ is an indicator for student i's score equalling q. Bond and Lang (2015) use this approach to measure the black-white gap in years-of-education-scaled test scores at various grades. Scaled this way, the reading gap is roughly constant from K through grade 7 at around 0.7 years of predicted educational attainment, while the math gap is close to a full year. These are larger than the realized black-white gaps in

eventual educational attainment, as black students tend to have higher educational attainment than do white students with the same early grade test scores.

This forward-linking approach does yield a scale that can be interpreted in a meaningful way, and that is plausibly interval. On the other hand, the choice of a specific outcome to link to, and of a linking function, implies a value judgment about the weight we put on different levels of achievement. There is no assurance that the particular interval scale defined by educational attainment will correspond to that defined by another outcome (such as earnings), nor that either corresponds to the interval scale representing units of effective educational inputs. It might simply require more inputs to move a student from 9 to 10 years of education than from 11 to 12 or 15 to 16.

This kind of future-linked scale has other undesirable properties as well. For example, consider two tests that are identical except that one is shorter, so the fraction correct for examinee i has more sampling error on the short test. The $y(q)$ mapping from the fraction correct to average later outcomes will be flatter on the shorter test, so the black-white gap will appear to be smaller. Even worse, measured performance on a forward-linked scale captures not only inputs prior to the test, but also inputs that students will receive later. Thus, for example, the existence of an effective intervention program for low-scoring adolescents will raise the average educational attainment of children who scored poorly as kindergarteners, and thus compress the left tail of forward-linked kindergarten scores. This is contrary to the standard education production function approach in which a student's ability at time t is a function of all inputs the student has received up to (but not following) time t. Thus, while this sort of approach is promising, we regard it as underdeveloped and not yet ready for broad application.

Finally, it might be possible to narrow the class of scale transformations that we are willing to consider, by in effect assuming that straightforward scales such as the raw score are partially but not fully interval. For example, we might be willing to assume that the difference between SAT scores of 1500 and 1000 is larger than that between 1000 and 990, even if we aren't willing to assume that it is fifty times as large. The challenge then is to parameterize and define this no-

tion that only *some* monotonic transformations (or, alternatively, some distributions of underlying achievement) are legitimate. Nielsen (2015) provides a valuable first step in this direction. His empirical results, like those of Bond and Lang (2013), suggest that cross-sectional achievement gap estimates (e.g., for black/white and high-/low-income) are quite robust to scale misspecification, but that achievement gap change estimates are considerably more sensitive to the choice of scale.

2.2 Scaling considerations and common practice

In cases where the outcome metric is well known (e.g., SAT points), researchers will often analyze scores using the reported scale. But when the scale is not familiar, economists frequently convert individual scores to a known scale. In practice, three transformations are used most often: percentiles, z-scores and Normal Curve Equivalents (NCEs). Percentile scores are computed as the percentile of the examinee's score relative to others. Z-scores are the difference between the examinee's scale score and the mean scale score, divided by the scale score standard deviation. Normal Curve Equivalents (NCEs) are obtained by applying the inverse distribution function of the standard normal distribution to the percentile score. As the above discussion makes clear, there is little basis for saying that these ad hoc transformations yield scales that are any more or less correct. Nevertheless, even when researchers are willing to implicitly assume that one of these transformations yields an interval scale, there are several specific hazards of which one should be aware.

First, it is important to keep in mind the population against which the assessment has been normed. All test measures are defined by reference to some norming population, which in practice can be quite small and non-representative. To the extent that one's analysis is focused entirely on a single data set or well understood population, this might be fine. But any sort of comparability across assessments depends on the use of comparable norming populations, and there is no assurance that the distribution of interval-scaled ability is constant across *different* populations (e.g., states, ages or cohorts). Suppose, for example, one is interested in comparing the impact of an intervention in two states that administered different exams. Constructing z-scores from samples from the two states assumes that both the mean and standard deviation of latent achievement, if

measured on the same scale, would be identical in the two states. This assumption has little foundation. Thus, comparison of NCE, z, or percentile scores across groups is inherently fraught when they are normed to different populations.

Cascio and Staiger (2012) provide an excellent example of this concern. They ask whether the common empirical result that interventions aimed at younger children tend to have larger effects on standardized test scores (z-scores) than do those aimed at older children could be attributable to the standardization process rather than an indication that achievement becomes less malleable as children age. Consider a simple evaluation of the effect of an intervention (represented by an indicator variable T) on the ability of student i tested at age t:

$$\theta_{it} = T_i\beta_t + u_{it}. \tag{2}$$

Common practice among economists is to standardize scores separately by age. Thus, the regression that is actually estimated is

$$\frac{\hat{\theta}_{it}}{\sigma_t} = T_i\frac{\beta_t}{\sigma_t} + \frac{u_{it}}{\sigma_t}, \tag{3}$$

where σ_t is the standard deviation of measured scores $\hat{\theta}_{it}$ (either overall or conditional on T) among age-t students. As Cascio and Staiger note, variation in the coefficient of this regression across ages could be driven either by β_t or by σ_t. There may be reason to suspect that σ_t grows with age, as older students have been exposed to more out-of-school influences whose effects may accumulate. If so, this could explain the observed pattern of declining coefficients with age. To gain traction on this problem, Cascio and Staiger (2012) adopt a parametric, additive model of student test scores as depending on a permanent child ability, long-term knowledge that decays at a constant, geometric rate, and a fully transitory component that combines what they refer to as "short-term knowledge" with pure measurement error on the test. They use this model to estimate the extent to which the observed decline in treatment effects with child age might be due to increases in the variability of long-term knowledge, concluding that while variance does increase with age, it cannot fully explain the smaller treatment effects among older populations.

Second, even if interval-scaled ability *is* similarly distributed across groups, *measured* ability may not be. Consider again the common practice of standardizing scores separately by age or grade. A test of kindergarteners may have more measurement error than a similar test of fifth graders. The grade-specific standard deviation σ_t combines the true variability of ability among grade-t children with measurement error in the grade-t test. Thus, even if the former is assumed to be invariant to t, if the latter declines with age, so will σ_t. The result of dividing by the composite will be to shrink the true ability variation differentially across grades. This would make between-group differences (e.g., the black-white gap) in z-scores larger for 5th graders than for kindergarteners, even if latent ability has identical distributions in the two grades.

As we describe in more detail below, the measurement error in reported test scores is a function of the way in which the score is constructed. In the simplest case, when the individual ability estimate contains classical measurement error, the across-student standard deviation the researcher calculates will be larger than the true standard deviation of ability in the population. If a z-score constructed from such a test score is used as an outcome in a regression, measurement error will attenuate the coefficients relative to what one would obtain with the correct variance measure. On the other hand, as described more below, the individual ability measure reported in many modern assessments has a variance *smaller* than the true variance of ability, which will lead to the opposite bias. Standardization must take account of these measurement issues if results are to be informative.

3 Measurement

Scaling involves the conversion of some "raw" ability measure into scale scores with a desired distribution. In this section, we discuss how test-makers obtain those raw ability estimates.[4]

Until relatively recently, psychometricians relied on what is known as "Classical Test Theory" to construct ability measures. In this framework, a raw score such as the fraction of items

[4]In the psychometrics literature, proficiency is often used to refer to an individual's latent ability. This should not be confused with the binary, criterion-referenced notion of proficiency that is common in assessing school performance today. As noted above, we use the terms ability and proficiency interchangeably.

answered correctly is viewed as the sum of an individual's true ability and classical measurement error. While this approach is simple and transparent, it has several limitations. First, items of moderate difficulty provide more information about a student's latent proficiency than do items that are very easy or very hard for her, and even holding difficulty constant some items may be better or worse at discriminating the dimension of proficiency of interest. (For example, a test item about baseball statistics may measure knowledge of the sport better than it does statistical proficiency.) Second, in many circumstances it is desirable to be able to compare the proficiency of students who were not all given the same test, either because the test is distributed in multiple forms or because – increasingly common with computer administration – it is adaptive, with items assigned based on the student's performance on past items. When different students are given items that (may) differ in their difficulty, the raw score should be adjusted.

Accordingly, modern psychometrics generally uses what are known as "Item Response Theory" (IRT) models that view the probability a student answers each item correctly as a function of the student's ability and some characteristics of the item (van der Linden and Hambleton, 1997). We discuss these models briefly in Section 3.1, then discuss their estimation afterward.

3.1 Item response theory

Let θ_i represent the latent, unobserved ability of examinee i, and let $R_i = \left\{r_{ij}\right\}_{j=1}^{J_i}$ be i's scored responses to test items $j = 1, ..., J_i$. For simplicity, we assume that r_{ij} is binary, with 1 representing a correct response and zero an incorrect response.[5] An IRT model specifies the probability that $r_{ij} = 1$ as a function of i's latent ability θ_i and a vector of item parameters ψ_j: $Pr\left(r_{ij} = 1\right) = F\left(\theta_i; \psi_j\right)$.

In many common cases, F is specified as a logistic function of some term that is linear in θ_i. The simplest IRT model is the Rasch model (also known as the one-parameter logistic, or 1PL).

[5]Commonly, students will fail to respond to some items. Item nonresponse is often treated as ignorable, with the relevant terms omitted from the likelihood, defined below. On so-called "speeded" tests, however, the number of items responded to is informative about student ability, and in some cases missing items are scored as incorrect. Modeling informative missingness (e.g., Glas and Pimentel, 2008 and Holman and Glas, 2005) is beyond the scope of our discussion. For ease of exposition we will treat R as complete.

It specifies ψ_j as a scalar (often written as b_j), representing the difficulty of item j, and:

$$Pr\left(r_{ij}=1\right)=F\left(\theta_i;b_j\right)=\frac{e^{\theta_i-b_j}}{1+e^{\theta_i-b_j}}, \tag{4}$$

or, alternatively,

$$\ln\frac{F\left(\theta_i;b_j\right)}{1-F\left(\theta_i;b_j\right)}=\theta_i-b_j. \tag{5}$$

In this model, items vary in their difficulty but the functions $F_j(\theta)\equiv F\left(\theta;b_j\right)$ (known as "item characteristic curves," or ICCs) are merely horizontal shifts of one another. Figure 1 illustrates the assumed ICCs for two items, one with $b_j=0$ and the other more difficult with $b_j=1$. Under the Rasch model (but not under other IRT models), for any fixed set of items the maximum likelihood estimate of θ_i is a monotonic transformation of the raw score (overall fraction correct) and does not depend on the specific pattern of correct and incorrect answers.

Other IRT models add additional item parameters. The most common IRT model is the three-parameter logistic, or 3PL: $\psi_j=\left\{a_j,b_j,c_j\right\}$ and

$$Pr\left(r_{ij}=1\right)=F\left(\theta_i;\psi_j\right)=c_j+\left(1-c_j\right)\frac{e^{a_j\left(\theta_i-b_j\right)}}{1+e^{a_j\left(\theta_i-b_j\right)}}. \tag{6}$$

b_j represents item difficulty, as in the Rasch model. The a_j parameter is known as the "discrimination" of item j. It can stretch or compress the item response curve horizontally, with steeper curves for items with higher a_j. c_j, known as the "guessability" of the item, allows for a non-zero probability of a correct answer even when θ_i is very low; it is often constrained to be constant across items. The 3PL model reduces to the 1PL model when $a_j=1$ and $c_j=0$. When c_j is constrained to 0 but a_j is unrestricted, one obtains the two-parameter logistic, or 2PL model. Figure 2 illustrates ICCs for three 3PL items. One fits the Rasch model, with $a_j=1$, $c_j=0$, and $b_j=0$. A second item is more discriminating, with $a_j=2$. This does a better job of distinguishing students with θ_i near b_j, but provides little information about students with θ_i far below or far above b_j (who get the item wrong or right, respectively, with very high probability). The third item shown returns to $a_j=1$, but is guessable, with $c_j=0.2$. On this item, even the very lowest proficiency

students have a positive probability of guessing the correct answer.

These models all assume that the item response is binary. There are also IRT models for scored items (e.g., essay questions or open response math items) where more than two scores are available, and the most sophisticated assessments and the corresponding IRT models often combine both binary and polytomous choice items. For a more complete discussion of IRT models, see van der Linden and Hambleton (1997). Embretson and Reise (2000) provide a readable introduction to the field for non-psychometricians.

Across all IRT specifications, item responses are generally assumed to be independent of all other observables conditional on the latent proficiency θ_i. They are also assumed to be independent across both students and items. This means that the likelihood function for examinee i can be written as:

$$Pr\left(R_i|\theta_i;\psi_1,...\psi_{J_i}\right)=\prod_j Pr\left(r_{ij}|\theta_i;\psi_j\right)=\prod_j\left(F\left(\theta_i;\psi_j\right)\right)^{r_{ij}}\left(1-F\left(\theta_i;\psi_j\right)\right)^{1-r_{ij}}. \qquad (7)$$

In the discussion below, we refer to (7), with the accompanying specification of $F_j(\theta)$, as the "measurement" model.

3.2 Measuring student ability in test scoring

The IRT model specification is used to extract information about examinees' latent ability from observed item response data. Given an IRT specification, item parameters ψ_j are identified as the number of examinees goes to infinity, while student $i's$ proficiency θ_i is identified as the number of test items goes to infinity. In practice, tests are commonly kept short to minimize respondent burden, while the number of examinees is relatively large ($N \gg J$). This means that item parameters are consistently estimated, but student proficiency may not be.[6] As the latter is the parameter of interest – we give tests to learn about test-takers, not about the tests – this poses a challenge. There are three general ways that student ability estimates are generated in modern assessment systems.

[6]Normalizations are required for identification. For example, some tests normalize $E[\theta_i]=0$ and $1/J\sum_j a_j = 1.7$ (chosen because this makes the ICC resemble a normal CDF).

In the first approach, one treats student ability as a fixed effect and measures the examinee's performance as the maximum likelihood (ML) estimate of θ_i from (7). The result, $\hat{\theta}_i^{ML}$, is an (asymptotically, as $J \to \infty$) unbiased estimator of θ_i, though the estimates can be quite noisy when J is small. Perhaps more importantly, $\hat{\theta}_i^{ML}$ is not defined for examinees who get all questions correct or all incorrect. Ad-hoc measures are used to assign scores to students whose score would otherwise be undefined. For example, the former state-mandated test in Michigan (the MEAP) assigned students who answered all (no) items correctly a score 10 percent higher (lower) than what was otherwise possible. Other tests simply set minimum and maximum scores, and assign students with missing ML estimates as well as those with ML estimates outside those range to the endpoints.

A second approach is to treat θ_i as a random effect, augmenting (7) with a model for the distribution of θ_i. The parameters of the θ distribution are estimated via ML; once this is done, a posterior distribution can be computed for each examinee, taking the population distribution of θ as the prior and the item responses as data. An examinee's performance is often measured as the mean of this posterior distribution, known as the posterior mean or "expected *a posteriori*" (EAP) score, $\hat{\theta}_i^{EAP}$. Most of the longitudinal databases created and distributed by the National Center for Education Statistics (NCES), including the National Educational Longitudinal Study of 1988 (NELS:88), the Educational Longitudinal Study (ELS), the High School Longitudinal Study (HSLS), and the Early Childhood Longitudinal Study (ECLS), report scores constructed from posterior means.[7]

Posterior mean scores can be interpreted as a Empirical Bayes (EB) estimates of students' latent ability (Morris, 1983). EB estimates are often referred to as "shrinkage" estimators because the estimate (at least in simple cases) can be seen as a weighted average of the individual's own score (i.e., the ML estimate describe above) and the population mean ability, where the weight (or

[7]One sometimes also sees "maximum *a posteriori*" (MAP) scores, which are posterior modes. An example is the the ASVAB scores reported in the 1997 wave of the National Longitudinal Study of Youth (NLSY97).

shrinkage factor) is a function of the precision with which the individual's own score is estimated:

$$\hat{\theta}_i^{EAP} = (1-\lambda)\,E[\theta] + \lambda\,\hat{\theta}_i^{ML} = E[\theta] + \lambda\left(\hat{\theta}_i^{ML} - E[\theta]\right) \tag{8}$$

with $\lambda = \frac{V(\theta_i)}{V(\theta_i)+V\left(\hat{\theta}_i^{ML}-\theta_i\right)}$. As this expression indicates, a student's EAP score will generally be closer to $E[\theta_i]$ than is her ML estimate. Importantly, posterior means are unbiased *predictors* of the unknown parameter, but not unbiased *estimators*: $E\left[\theta_i|\hat{\theta}_i^{EAP}\right] = \hat{\theta}_i^{EAP}$, but $E\left[\hat{\theta}_i^{EAP}|\theta_i\right] \neq \theta_i$. As we discuss in Section 4, this has important implications for the secondary analysis of EAP scores.

In IRT models, ability is estimated most precisely for individuals with θ's near the middle of the measured ability distribution. This is because items are most discriminating (i,e, their characteristic curves are steepest, and thus the examinee's response provides the most information in the Fisher sense) when $\theta_i = b_j$.[8] For this reason, the reported ability measure in this framework will be shrunk more for students that score extremely high or low on the exam.

A third measurement approach, taken by several major assessments including the National Assessment of Educational Progress (NAEP), is to report what is known as "plausible values" (PVs). Each plausible value, $\hat{\theta}_i^{PV}$, is a random draw from the examinee's posterior distribution. Tests typically report several draws for each examinee, and recommend that secondary analysts use the across-draw variability to model the contribution of these random draws to the sampling variability of their estimates. [9]Plausible values are closely related to multiple imputation for missing data, and indeed both derive from Rubin's (1987; 1996) work on the topic. They are neither unbiased estimators nor unbiased predictors of individual ability. However, the across-examinee

[8]Interestingly, the standard errors of raw scores are *largest* at this point, and smaller in the tails: The variance of the fraction correct, p, is $p(1-p)$, and this is highest when p is close to 0.5. Intuitively, random chance has the biggest role when the probability of getting an item correct is close to 50%.

[9]NAEP instructs researchers to conduct their analysis separately using each of the $K = 5$ plausible values for a student, thus generating K statistics $\hat{\beta}_k$ $(k = 1,\ldots,K)$. These estimates are then averaged to obtain the point estimate $\hat{\beta}^{PV} = \frac{1}{K}\sum_1^K \hat{\beta}_k$. The variance of the point estimate is computed as the average of the five estimated variances plus the variance of the average (computed from the across-PV variability): $V\left(\hat{\beta}^{PV}\right) = \frac{1}{K}\sum_{k=1}^K V\left(\hat{\beta}_k\right) + \frac{1}{K(K-1)}\sum_{k=1}^K\left(\hat{\beta}_k - \hat{\beta}^{PV}\right)^2$.

distribution of PVs should reproduce the univariate distribution of θ_i, where the variance of maximum likelihood estimates will be larger than that of the underlying parameter and the variance of posterior means will be smaller. Indeed, the primary benefit of using PVs is to obtain consistent variance estimates for use in producing summary statistics.

Figures 3-5 use simulated data on 10,000 individuals with $\theta_i \sim \mathbb{N}(0,1)$ to illustrate these three approaches. In Figure 3, each individual is administered an exam with 15 items where response probabilities follow a simple Rasch model and the difficulty parameters are uniformly distributed on $[-2,2]$. For each examinee, we computed direct the maximum likelihood (ML) estimate of θ_i, the posterior mean, and one plausible value.[10] The dotted line shows the true proficiency distribution, while the other lines show kernel density estimates of the other distributions: dash-dot for ML, solid for posterior means, and dashed for plausible values. All three score measures are centered around zero, the mean of the true θ distribution. Their variances differ, however: The MLE distribution has a variance much larger than $V(\theta_i)$, while the posterior mean distribution has a smaller variance than does true proficiency.[11] By contrast, the plausible values have a distribution very similar to that of the true θ_i.

The test used in Figure 3 is relatively short at only 15 items (about half the length of the NAEP test), but is also well designed to discriminate among students. Figure 4 repeats the exercise using a much easier test, with item difficulty parameters distributed as $\mathbb{N}\left(-1, 0.25^2\right)$. We again see too little variability of the posterior means and too much of the ML estimates. The mismeasurement is particularly severe on the right tail of the θ_i distribution: Over 6% of students earn perfect scores, so receive ML scores of 4.0, but this in large part reflects the ease of the

[10] In the ML estimation, scores are censored from below at -4 and from above at +4, and these scores are assigned to students who got all items wrong or right, respectively. Each examinee's items are drawn independently from the $\mathbb{U}[-2,2]$ distribution, as with common items the ML score distribution would have only 16 points of support. Test scoring assumes that item parameters are known.

[11] The reduced variance of the posterior means is not because we have under-estimated $V(\theta_i)$: A correctly specified random effect model of latent ability yields a consistent estimate of the variance of true ability. But even so, the estimated posterior means for individuals have too little variation relative to that in latent ability.

questions; the highest posterior mean score is just over 1.5.

Figure 5 illustrates a longer test, with 50 items drawn from the same distribution as in Figure 3. With this many items, and this wide a spread of item difficulty, measurement choices are less important. ML scores are still too variable and posterior means not variable enough, but the differences are much smaller.

3.2.1 Incorporating conditioning variables into the generation of latent student ability measures

As Figures 3 and 4 illustrate, short tests do not accurately pin down θ_i, resulting in quite wide posterior distributions. In order to increase the precision of ability estimates, some assessments – including the premier U.S. and international assessment systems, the National Assessment of Educational Progress (NAEP) and the Program on International Student Assessment (PISA), respectively – use student background characteristics to generate more informative priors. Under this approach, the prior distribution of θ_i for student i with characteristics Z_i is specified as $p_c(\theta_i|Z_i;\pi)$, where π represents parameters (e.g., coefficients of an unknown regression of θ_i on Z_i). We refer to this as the "conditioning model" and to the vector Z_i as the "conditioning variables." Together, the conditioning model and the measurement model (equation 7) yield an expression for the likelihood of observed item responses R_i given the IRT item parameters Ψ and the conditioning model parameters π:

$$Pr(R_i|Z_i;\pi,\Psi) = \int P_{IRT}(R_i|\theta;\Psi)\, p_c(\theta|Z_i;\pi)\, d\theta. \tag{9}$$

The test-maker estimates ψ and π via maximum likelihood applied to (9), using numerical methods (e.g., quadrature) to handle the integration. Posterior distributions are then calculated using $p_c(\theta_i|Z_i;\hat{\pi})$ as the prior for student i's ability and Bayes' rule:

$$p(\theta_i|R_i,Z_i) \quad \propto \quad p_{IRT}(R_i|\theta;\hat{\Psi})p_c(\theta_i|Z_i;\hat{\pi}). \tag{10}$$

As above, the posterior mean can be viewed as an Empirical Bayes (EB) or shrinkage estimator.

Instead of being shrunk toward the unconditional population mean, however, a student's performance on the items she answered is shrunk toward the mean of all students with the same values of any included conditioning variables:[12]

$$\hat{\theta}_i^{EAP|Z} = E\left[\theta_i|Z_i\right] + \lambda_i\left(\hat{\theta}_i^{ML} - E\left[\theta_i|Z_i\right]\right). \tag{11}$$

As a result, the posterior distribution for student i varies both with the student's test responses R_i and with her background characteristics Z_i. Suppose, for example, that race is one of the background variables contained within Z (as indeed it is in NAEP), and that on average black students have lower proficiency than white students. Now consider two students, one black and one white but otherwise identical in their Zs, who also respond identically to the NAEP assessment items. Our two students' performance is "shrunken" toward different group averages. As a result, the white student's posterior distribution will stochastically dominate that of the black student, leading to gaps in their posterior means and plausible values. This creates biases in some, but not all, secondary analyses. We turn to that issue next.

4 Secondary analysis with latent ability measures

In this Section, we discuss the implications of measurement and scaling decisions for secondary analysis of the scores. For simplicity, we focus on linear OLS regressions with the test score used either as a dependent or an independent variable. In order to focus specifically on issues arising from scaling and measurement, we ignore both sampling variability (assuming that the number of examinees is large) and omitted variable bias (assuming that linear projections are of interest, perhaps because the research design supports their causal interpretation). It turns out that the existence and nature of bias depends on the type of ability measure used (i.e., a direct maximum likelihood estimate, a posterior mean or a plausible value) and whether ability measure is a dependent or independent variable.

[12]The shrinkage factor here differs from that in equation (8), as it depends on the *conditional* variance of θ, $V\left(\theta|Z\right)$.

4.1 Ability as the dependent variable

Consider the following research model:

$$\theta_i = X_i\beta + \varepsilon_i. \tag{12}$$

where θ_i is the latent student ability (measured on a scale that we assume, for now, is appropriate and interval), X_i is a vector of observed covariates including a constant, and ε_i is a residual term that has mean zero and is uncorrelated with X_i. Because we only observe an imperfect measure of latent ability, $\hat{\theta}_i \equiv \theta_i + u_i$, the feasible regression is

$$\hat{\theta}_i = X_i\beta' + \varepsilon_i'. \tag{13}$$

We are interested in the relationship between the feasible coefficient β' and the "true" coefficient β. This depends on the measurement and scaling of $\hat{\theta}_i$. We discuss measurement considerations first, assuming that $\hat{\theta}_i$ is measured on the appropriate scale, then turn to scaling issues.

4.1.1 Ability measures without conditioning variables

The simplest case is that where the ability measure is a maximum likelihood estimate, $\hat{\theta}_i^{ML}$, that can be viewed as an approximately unbiased estimate of θ_i. The error component u_i derives from the student's luck in responding to test items, and can be expected to be orthogonal to X_i. Because orthogonal measurement error in a dependent variable does not bias regression coefficients, use of $\hat{\theta}_i^{ML}$ will generate consistent (but possibly less precise) estimates of β.

In contrast, ability measures based on the posterior distribution of θ_i are not generally unbiased estimates of individual ability. As discussed above, in the simplest case $\hat{\theta}_i^{EAP}$ is a shrunken version of the maximum likelihood estimate, with a uniform shrinkage factor λ. Using this as a dependent variable will attenuate the estimated β coefficients (other than the constant) by a factor λ. For example, consider the regression of posterior mean scores (generated without conditioning variables) on a binary variable indicating if the student is poor. If poor children have lower than

average scores and non-poor children have higher than average scores, then use of posterior mean will overstate (understate) the average ability of poor (non-poor) children, thus understating the poverty achievement gap. We present an empirical illustration of this below, drawn from Briggs (2008).

It is unrealistic, however, to assume that the estimation error in $\hat{\theta}_i$ is homoskedastic. If it is not, the shrinkage factor will vary across observations. This complicates the interpretation of estimates, which may not simply be attenuated. For example, because most IRT-based tests yield more precise estimates of θ_i for students in the middle of the distribution than for those at the tails, the scores of students with very high or very low $X_i\beta$ are likely to be shrunken by more. The bias this generates is difficult to characterize in general. Similar bias arises in other contexts where shrinkage estimators are used as dependent variables. For example, Chetty et al. (2014a) assess inequities in access to good teachers by regressing what amount to Empirical Bayes estimates of teacher value-added (i.e., posterior means) on observable student characteristics. Recognizing that the resulting coefficients are attenuated relative to what would be obtained if true value-added were used as the dependent variable, Chetty et al. (2014a) multiply all coefficients by the inverse of the average shrinkage factor. However, because experienced teachers have more years of available data, their value-added scores are shrunk by less, and applying a uniform correction may not recover the correct estimates of β.

What about tests that report plausible values? A plausible value is merely the sum of the posterior mean plus a deviation that is randomly generated as part of the test scoring process. This deviation is by construction orthogonal to X_i, so contributes nothing to the estimated regression coefficient in expectation. Hence, PVs generate estimates of β that are biased in the same way as those obtained from posterior means, though less efficient.

NAEP and other assessments that report plausible values provide users with detailed guidance on how to conduct analysis using these measures. This guidance is primarily intended to ensure the researcher uses the correct variance measures when conducting inference. In practice, economists often do not follow the recommended guidelines for using PVs, instead using the first

PV or a simple average of the PVs as the individual ability measure (see, for example, Blau and Kahn, 2005, or Ferrer et al., 2006). While this is a source of frustration to psychometricians,[13] heteroskedastic-robust inferential methods should be consistent even without explicit adjustment for the random choice of PVs from each examinee's posterior distribution. But neither robust standard errors nor the recommended procedures eliminate bias in coefficients from the use of shrunken ability estimates.

4.1.2 Ability measures that incorporate conditioning variables

When $\hat{\theta}_i$ is a posterior mean or plausible value computed using conditioning variables, the potential biases are more complicated. As described above, the inclusion of conditioning variables can be thought of as shrinking a student's individual performance on given test items toward her group (Z-) specific mean (equation 11).

If all of the independent variables X_i in the research model are included in the set of conditioning variables Z_i, $X \subseteq Z$, then the second term on the right-hand side of (11) is orthogonal to X_i in expectation, and both posterior mean-based and PV-based coefficient estimates are consistent (though PVs are inefficient relative to posterior means). As intuition, note that only the portion of achievement that is not predicted by the conditioning variables is shrunken. Thus, for example, if X is a set of race indicators, and if Z also includes these indicators (perhaps along with other variables), then the racial gap in posterior means or PVs will equal the racial gap in latent proficiency.

The same intuition implies that when the X variables are *not* included in Z, secondary analysis using the posterior mean or plausible values will lead to inconsistent estimates in general, as the portion of the variation in θ_i associated with X but not Z is shrunken. The (asymptotic) bias is most severe when the explanatory variables of interest are poorly proxied for by those in the conditioning set and when the regression also includes correlated explanatory variables that *were* in (or well proxied by) the conditioning set (see, e.g., Mislevy, 1991).

Importantly, if the conditions for consistency are not met, the bias is not confined to par-

[13] For excellent recent summaries, see von Davier et al. (2009) and Carstens and Hastedt (2010).

ticular coefficients. All other coefficients may be biased as well, in ways that would be hard for a secondary researcher to predict given the limited documentation of the conditioning model in most test data sets. For example, consider a regression of test scores on student background and school policy measures, where the conditioning variables include the former but not the latter. Then the policy coefficients are likely to be quite attenuated, and the student background coefficients may also be biased if the background and policy measures are correlated.

It is not clear how severe this bias is in practice. Mislevy (1991) reports results of a reanalysis of the 1984 NAEP Long Term Trend reading assessment. He finds that biases in coefficients on X variables included in Z are small, but that coefficients on X variables not included in Z are substantial. But psychometric methodology and computational capacities have advanced considerably since 1984. Most testing systems with conditioning models use very high-dimensional Zs in hopes that any X considered by subsequent researchers will have been included directly or by proxy in Z.[14] Recent NAEP administrations use hundreds of variables in the conditioning model, including student demographics (e.g., race, gender and age), family background characteristics (e.g., parental employment, parental education), school characteristics (e.g., racial composition of the school, urbanicity of school location), student self-reports of study habits and school performance (e.g., overall grades, expected educational attainment, time spent on homework), and teacher reports of aspects of the curriculum and of school policies. Given the large amount of information, the excluded elements of X may be quite predictable from the information in Z, which would permit unbiased estimation of β. Indeed, a reader might legitimately ask whether there are any interesting analyses that involve excluded variables.

We are nevertheless concerned that even NAEP-style conditioning models may produce serious bias in interesting secondary research models. While the NAEP conditioning model includes an extensive list of student and school characteristics, it has few variables that are likely to be of

[14]This is a form of bias-efficiency tradeoff: A more efficient estimate of β might be obtainable with a lower-dimensional Z, so long as $X \subseteq Z$, but because the institution generating the PVs does not know what Xs the secondary analyst will choose it accepts less efficiency in order to permit a wider range of analyses to be unbiased.

interest for secondary policy evaluations. For example, the NAEP conditioning model does not include measures of whether the school offers performance pay to its teachers, the type of school accountability system in place in the state, or the form of the state school finance formula, and none of these are likely to be very well proxied by the characteristics that are included. If not, the estimated effects of these Xs from analyses using NAEP PVs as the outcome measure – even with an identification strategy that meets the usual criteria – may be importantly attenuated.

4.1.3 Simulation Results

To illustrate the biases described above, we present a simple simulation. We assume a data set on 10,000 students with two potential predictors of student proficiency, which we label X and W and think of here as parental education and neighborhood poverty respectively, plus an estimated proficiency measure. We assume the following data generating process:

$$\theta_i = X_i + e_i;$$

$$W_i = X_i + u_i;$$

$$\{X_i, e_i, u_i\} \overset{iid}{\sim} \mathbb{N}\left(0, 0.5^2 * \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.5 \\ 0 & 0.5 & 1 \end{pmatrix}\right).$$

Note that under this DGP, $\theta_i \sim \mathbb{N}(0,1)$, $\beta = 1$, and $\theta_i | X_i \sim \mathbb{N}\left(0, 0.5^2\right)$. We assume that students are administered a short, 20-question multiple choice test, where each item's characteristic curve follows the Rasch specification with difficulty zero (see Figure 1). We consider several different measurement models for scoring the test: Maximum Likelihood scores (with minimum and maximum scores set to -2 and +2), posterior means, and plausible values, in the latter cases with varying conditioning sets. Table 1 lays out the DGP.

Table 2 shows estimates of the coefficient from a regression of $\hat{\theta}_i$ on X_i, under several different ways of scoring the test. The first row shows results from the ideal regression, using

actual θ_is. By construction, this coefficient equals 1. The second row shows results when θ_i is not observed but the ML estimate based on the student's test responses is reported. Because this is an unbiased estimator of θ_i, it permits unbiased estimation of β ($\beta' = 1.016$, with standard error 0.012). In rows 3 and 4, the first column shows results when the test scores are reported as posterior means or plausible values, respectively, without conditioning variables. In these cases, the feasible regression yields a coefficient that is attenuated by approximately 20 percent. Columns 2-4 explore how the choice of conditioning variables for the Bayesian scoring models influences the estimated coefficients in our research model. Because X and W are correlated, the inclusion of W alone in the conditioning set reduces the bias in β', though the estimate is still noticeably attenuated. By contrast, if X is included in the conditioning set, either alone or with W, then $E\left[\hat{\theta}|X\right] = E\left[\theta|X\right]$, and β' is unbiased.

4.2 Ability as an independent variable

Now consider a research model in which latent ability is the independent variable, as in the regression

$$Y_i = X_i\gamma + \theta_i\delta + \varepsilon_i, \tag{14}$$

where Y_i is some outcome measure, X_i a vector of observed covariates (including a constant), θ_i the latent student ability, and ε_i a residual that by construction has zero mean and zero correlation with X_i and θ_i. As before, we observe only $\hat{\theta}_i \equiv \theta_i + u_i$, so the feasible regression is

$$Y_i = X_i\gamma' + \hat{\theta}_i\delta' + \varepsilon_i', \tag{15}$$

We are again interested in the relationship between the parameters of the ideal regression (14) and those of the feasible model (15).

Economists are generally familiar with regressions with measurement error in an explanatory variable. Under classical measurement error assumptions, standard errors-in-variables results imply that δ' will be attenuated by a factor equal to the reliability ratio of $\hat{\theta}_i$ conditional on X_i, $R_{\hat{\theta}|X} = \frac{V(\theta_i|X_i)}{V(\theta_i|X_i)+V(u_i)}$, and that the γ' coefficients will be biased by a factor $R_{\hat{\theta}|X}\delta\omega$, where ω rep-

resents the coefficient vector of an auxiliary regression of θ_i on X_i.[15] Because the elements of ω can be positive or negative, the bias in γ' cannot be signed in general, but ω is straightforward to estimate from the available data so the sign of biases is easily knowable.

Unfortunately, the classical measurement error assumptions only apply when individual ability is reported as a direct maximum likelihood estimate $\hat{\theta}^{ML}$. By contrast, testing systems that report Bayesian estimates of individual proficiency, either posterior means or plausible values, do not yield scores that fit the classical measurement error model. The measurement error in these scores is correlated (generally negatively) with the student's true ability θ_i. Intuitively, "shrinkage" estimators pull an examinee's reported score more toward the mean the further is her true score from the mean. The relationship between $\{\gamma', \delta'\}$ and $\{\gamma, \delta\}$ depends on the specific $\hat{\theta}_i$ measure used and on the conditioning model (if any) used by the test-maker in constructing this measure. Unlike in the dependent variable case, results are quite different for posterior means (expected *a posteriori* scores) than for plausible values.

4.2.1 Reported scores are posterior means

It is useful to start with a case where the feasible model recovers the coefficients of the ideal model. This occurs when the test-maker uses a conditioning model, as in (9), where the conditioning set Z_i includes all of the covariates X_i from the research model but does *not* include *any* additional elements that are correlated with ε_i. In this case, the posterior mean $\hat{\theta}_i^{EAP|Z}$ is an unbiased predictor of θ_i conditional on X_i: $E\left[\theta_i|\hat{\theta}_i^{EAP|Z}, X_i\right] = \hat{\theta}_i^{EAP|Z}$. As a result,

$$E^*\left[Y_i|\hat{\theta}_i^{EAP|Z}, X_i\right] = X_i\gamma + E^*\left[\theta_i|\hat{\theta}_i^{EAP|Z}, X_i\right]\delta = X_i\gamma + \hat{\theta}_i^{EAP|Z}\delta, \tag{16}$$

where E^* denotes a linear projection. Thus, the feasible regression is unbiased for $\{\gamma, \delta\}$.

Unfortunately, this case is unlikely to occur in practice. Few testing systems include conditioning variables at all, and those that do typically include a very large number of measures in

[15] This can be seen as a type of omitted variables bias, where θ_i is only partly omitted – as $R_{\hat{\theta}|X}$ asymptotes toward zero, the omission is more and more complete, and the bias in γ' approaches the omitted variables bias $\delta\omega$.

Z_i in the hope of spanning all of the explanatory factors that secondary researchers might hope to examine. (As we discussed above, this is a helpful strategy when $\hat{\theta}_i^{EAP|Z}$ is to be used as a dependent variable.) For example, recent NAEP administrations include in Z_i hundreds of principal components computed from thousands of student background and school composition and policy variables. Most secondary researchers are likely to estimate more limited models, in which many of the characteristics included in the test-makers' Z_i are left in the error term ε_i.

When the test-maker does not use a conditioning model (that is, when the posterior means are based only on item responses) the feasible coefficients are not unbiased. Here, δ' is attenuated, and γ' is biased toward the coefficients of a regression of Y_i on X_i without ability controls. The direction (but not magnitude) of these biases is the same as in the classical measurement error case considered above, though the reasoning is somewhat different. To see this, recall that the posterior mean conditional on Z_i is a "shrinkage' estimator:

$$\hat{\theta}_i^{EAP|Z} = \lambda_i \hat{\theta}_i^{ML} + (1-\lambda_i)\, E\left[\theta | Z_i\right], \tag{17}$$

where the shrinkage factor λ_i equals the reliability of $\hat{\theta}_i^{ML}$ conditional on Z_i

$$\lambda_i = \frac{V\left(\theta_i | Z_i\right)}{V\left(\theta_i | Z_i\right) + V\left(\hat{\theta}_i^{ML} - \theta_i | Z_i\right)}. \tag{18}$$

The conditional variance of true ability, $V\left(\theta_i | Z_i\right)$, shrinks as the conditioning set Z_i grows. This pulls λ_i further from one, leading the examinee's performance to be given less and less weight. By contrast, when Z_i is the empty set, λ_i is at its largest. Thus, the unconditional posterior mean is under-shrunken relative to the posterior mean conditional on X, which as we established above allows for unbiased estimation. This implies that when the unconditional posterior mean is used, δ' is attenuated. Biases are similar when Z_i is non-empty but does not include all of the variables in X_i.

What about when the conditioning model is too large, including not just X_i but also additional variables that are correlated with the residual in the research model? This creates a corre-

lation between $\hat{\theta}_i^{EAP|Z}$ and ε_i even if the latter is orthogonal to true proficiency θ_i. Suppose, for example, that both parental education and neighborhood poverty are included in the conditioning set. If the research model only includes parental education as an observed covariate, neighborhood poverty is implicitly included in the error term. The child of high school dropouts who answers a very high fraction of test items correctly will have her estimated ability pulled down if she lives in a poor neighborhood, generating a negative correction between $\hat{\theta}_i^{EAP|Z}$ and ε_i. The resulting bias cannot be signed in general since it depends on the direction of the correlations among the conditioning variables Z_i and the research model residual ε_i.

4.2.2 Reported scores are plausible values

The above discussion established that when the test-maker reports posterior means using conditioning variables Z_i but the researcher uses them as explanatory variables in a research regression that conditions on X_i, the resulting coefficients are unbiased when $Z_i = X_i$ but are likely to be biased when Z_i is either larger or smaller than X_i. This is *not* the case when the reported test scores are instead plausible values. Here, unbiasedness requires that Z_i be substantially larger; it must include not only X_i, but also the *dependent* variable from the research model, Y_i. Moreover, even here rather strong functional form restrictions are required.

It is useful to start again with the case where the test-maker's conditioning set Z_i equals the secondary analyst's covariate vector X_i. Recall that posterior means based on this conditioning set can produce unbiased estimates of the coefficients of the research model. But plausible values can be seen as posterior means plus randomly generated noise. This noise is classical measurement error, and attenuates the δ' coefficient.

So clearly the conditioning set needs to be larger than X to permit unbiased estimation using PVs. Schofield et al. (2015) show when Z_i includes both X_i *and the research model's dependent variable* Y_i, PV-based regressions are unbiased. To see this, note that PVs based on conditioning variables Z_i accurately reproduce the joint distribution of $\{\theta_i, Z_i\}$, or at least its first and second moments. When Z_i includes X_i and Y_i, then, the first and second moments of $\left\{\hat{\theta}_i^{PV|Z}, X_i, Y_i\right\}$ equal those of $\{\theta_i, X_i, Y_i\}$. This means that the coefficients of the feasible regression based on the former

are identical to those of the ideal regression based on the latter.

However,Schofield et al. (2015) emphasize another required assumption here. Specifically, the parameterization of the conditioning model requires, in effect, imposing assumptions about the shape of the joint distribution of $\{\theta, X, Y\}$, and these assumptions in turn determine the functional form of $E[Y|X,\theta]$. If the latter is not consistent with the linear form imposed by the research model, even the correct choice of conditioning variables may not permit unbiased inference. Thus, for example, if Y_i enters the conditioning model linearly, it may be possible to use the plausible values to estimate a linear regression of Y on θ without bias, but a log-linear regression would be likely to be biased (and vice versa). In other words, the functional form decisions of the test-maker constrain the set of regressions that can be estimated without bias by secondary researchers.

Schofield et al. (2015) argue that the required alignment between the conditioning and research models is unlikely to occur in practice. Test makers generally look to potential explanatory factors for θ_i in choosing their conditioning sets, and are unlikely to include variables that might be *caused by* θ_i. But in secondary research, Y_i is typically some outcome variable that might be affected by student ability and is measured subsequent to the test administration used to measure $\hat{\theta}_i$ – e.g., in Neal and Johnson (1996), Y_i is adult wages where $\hat{\theta}_i$ is computed from the AFQT test given to subjects in adolescence. In the most common case where $\hat{\theta}_i$ is a set of PVs and the conditioning model includes a wide range of predictors of θ_i but does not include Y_i itself, we are aware of no general results on the sign or magnitude of bias.

4.2.3 Simulation Results

We can extend our earlier simulation to illustrate the independent variable case as well. To the DGP introduced in Section 4.1.3, we add a new variable,

$$Y = 4 + X_i + \theta_i + v_i, \tag{19}$$

with $v_i \sim \mathbb{N}(0, 0.5)$, $corr(u,v) = 0.5$, and $corr(X,v) = corr(e,v) = 0$. (14) is our research model, with X_i and θ_i as observed explanatory variables but their coefficients (equal to 1 in the DGP)

unknown. As before, we assume examinees are presented with 20 Rasch items, all with difficulty zero, and we consider different measurement models for score reporting.

Panel A of Table 3 presents the ideal model coefficients γ and δ obtained by using the true θ_is in the regression. Because the ML estimate of θ is not an unbiased predictor, when we use it as an independent variable in Panel B, the θ coefficient is attenuated (0.78), and because θ and X are positively correlated, the X coefficient is biased upward (1.13). The use of posterior means based on no conditioning variables (Panel C, column 1) only partially reduces the biases in γ' and δ'. If the posterior mean is calculated from a model that uses X alone as a conditioning variable (column 2), one obtains unbiased estimates of γ and δ. Unlike the dependent variable case, however, posterior mean calculations that include additional conditioning variables (columns 3-6) lead one to overstate the effect of θ and thus understate the impact of X.

The final panel (D) of Table 3 presents results with plausible values. Here, the model that conditions just on X is quite badly biased, but adding Y to the conditioning set eliminates nearly all of the bias. Functional form matters, however: When the conditioning model specifies $E[\theta|X,Y]$ as linear in X and $\ln(Y)$,[16] the PV-based regression is meaningfully biased.

4.3 Options for unbiased estimation

Given the potential biases outlined above, an important question is whether there are options available to the secondary researcher that permit unbiased estimation. Fortunately, there are options in many cases; unfortunately, all require access to additional information beyond the reported test score itself.

First, in some cases it may be possible to reverse-engineer the above bias formulas. For example, if an ML estimate of θ is to be used as an independent variable and estimates of the reliability of this measure are available (e.g., the average standard error of measurement (SEM) across examinees), the errors-in-variables results can be used to back out estimates that would be obtained were correctly measured ability available for use in the regression. Similarly, where

[16]For computation of the log, Y_i is left-censored at 0.5 in the small number of cases that fall below that.

ability is to be used as a dependent variable and the test only reports posterior means, estimates of the SEM as a function of θ can be used to "un-shrink" the test scores.

More general solutions typically require access to item-level data. If these data are available, it is possible to generate consistent and efficient estimates of the coefficients in the research model by integrating the measurement and research models into a single system of equations. In the dependent variable case (Section 4.1), the researcher specifies a new conditioning model with only the X_i variables included in the conditioning set. This model specifies the distribution of θ as a function of X, for example as $\theta_i|X_i \sim \mathbb{N}\left(X_i\beta, \sigma^2\right)$, so nests the research model (12). This is combined with the IRT measurement model to yield a likelihood for the observed item responses in terms of the item parameters ψ and the coefficients of interest β, as in (9). For example, if the conditioning model specifies that $\theta_i|X_i \sim N\left(X_i\beta, \sigma^2\right)$, then the likelihood is

$$L\left(R_i|X_i;\beta,\sigma^2;\Psi\right) = \int P_{IRT}\left(R_i|\theta;\Psi\right)\phi\left(\frac{\theta - X_i\beta}{\sigma}\right)d\theta.$$

β is then identified by maximizing this likelihood, without the intermediate step of generating posterior distributions for each examinee's θ_i and relating these to X_i. This is known as Marginal Maximum Likelihood (MML). In essence, the secondary researcher merely reproduces the steps that the test-maker would have undertaken had the test-maker used the same conditioning variables as are included in the secondary researcher's model. This approach is described in seminal articles by Mislevy (1991, and 1992 with others). [17] Its only major drawbacks are that it requires the secondary researcher to code up the IRT model $P_{IRT}\left(R_i|\theta;\Psi\right)$ and that numerical integration over the θ distribution is computationally costly. But modern computation techniques (e.g., Markov Chain Monte Carlo) facilitate the latter, and test-makers could do much to facilitate the former. (For example, some editions of the NAEP technical documentation include code for the IRT model probabilities.)

Briggs (2008) uses an MML-like approach, which he refers to as "Explanatory Item Re-

[17] Indeed, the American Institutes of Research (AIR) developed software intended, in part, to estimate such models. See http://am.air.org/contact2.asp.

sponse Theory," to assess the extent of bias in models where ability is the dependent variable and only posterior means without conditioning variables are available. He examines racial & ethnic gaps in student achievement using data from 10th graders in 1999 who were administered the Partnership for the Assessment of Standards-based Science (PASS) test. He compares gaps estimated from posterior mean science achievement to those obtained via a MML procedure that includes the student race indicators in the conditioning model that is estimated jointly with the IRT-based measurement model.[18] Table 4 reproduces his estimates. In column 1 of Panel A, we show estimates obtained using posterior mean scores. These indicate that the black-white achievement gap is -0.61 scale points (on the IRT model's native scale). But per the discussion above, we expect this to be attenuated. Indeed, when the model is estimated via MML (column 2), the black-white gap increases (in absolute magnitude) to -0.77 scale points. Columns 3 and 4 report estimates for Z-scores, created by dividing the scale scores by the standard deviation of these scores (column 3) or by the estimated standard deviation of latent proficiency (column 4). Again, the two sets of estimates give notably different answers: A black-white gap of 0.87 standard deviation units when posterior means are used, or 0.95 when computed via MML.

Panel B shows comparable results for subdomains of science achievement. Traditional psychometric methods model student achievement as a random effect that may be correlated across subdomains. Posterior means therefore "borrow" information from one subdomain in predicting the student's score on other subdomains. This may mask true differences in regression coefficients across subdomains. We see this in the Table: The black-white gap is estimated at 0.42 (in scale score units) in life science and 0.58 in physical science when the posterior mean scores are used. But when the MML model is used to estimate racial gaps on each subdomain, without shrinking them toward a common dimension, Briggs finds that the physical science gap is more than double the life science gap. (Meanwhile, the Hispanic-white gap is much more similar across subjects.)

Schofield et al. (2015) propose an MML-like procedure that can be used when the research model uses ability as an independent variable, which they refer to as a the "Mixed Effects Structural

[18] In each case, he uses only a subset of the test items to compute the student's score.

Equations" (MESE) model. Here, there are three equations: The IRT model for item responses R_i as a function of θ_i, $p_{IRT}(R_i|\theta_i;\Psi)$; the research model that specifies the distribution of Y given X and θ, $p_Y(Y_i|X_i,\theta;\gamma,\delta)$; and a conditioning model that includes only X_i, $p_\theta(\theta_i|X_i;\pi)$. Together, these specify the joint distribution of the observed variables $\{X_i,Y_i,R_i\}$:

$$p(Y_i,R_i|X_i) = \int p_{IRT}(R_i|\theta_i;\Psi)\, p_Y(Y_i|X_i,\theta;\gamma,\delta)\, p_\theta(\theta_i|X_i;\pi)\, d\theta. \tag{20}$$

This is maximized over the parameters $\{\gamma,\delta;\Psi;\pi\}$ to obtain consistent estimates of the research model coefficients.

Junker et al. (2012) use the MESE approach to assess the bias that results from using either ML estimates or PVs (based on the test-maker's conditioning set) as independent variables. They use data from the National Adult Literacy Survey (NALS), a nationally representative sample of U.S. adults in 1992 that contains information on cognitive ability along with with survey information on a variety of demographic and socio-economic outcomes such as educational attainment and earnings. They focus on a sub-sample of 25-55-year-old men and women who work full time, answered at least one item on the literacy test, report a weekly wage and self-report as black or non-Hispanic white. They estimate a standard wage regression in which the outcome is log weekly wages, and the primary explanatory variables include an indicator for race/ethnicity (black), a quartic in potential experience, and indicators for urbanicity and census region.

Table 5 reproduces their results for their sample of 3,267 men. Column 1 shows results that do not include any control for individual literacy. These indicate that weekly wages for blacks are roughly 36.6 log points (or 30.6%) lower than for whites. Column 2 controls for a maximum likelihood estimate of individual literacy that the authors generate using a standard IRT model. The implied black-white wage gap in this model drops dramatically to 14.4 log points (13.4%). However, recall from above that the ML measure of θ_i suffers from classical measurement error, and we would thus expect the literacy coefficient to be attenuated and the race indicator to be biased correspondingly. Column 3 presents estimates from the MESE model. As expected, the literacy

coefficient increases and the implied black-white wage gap drops to 9.4 log points (9%). This suggests that properly controlling for latent ability accounts for 74 percent of the unconditional black-white log wage gap (= 1 - (-0.094/ -0.366)) whereas using the more typical control would only account for 61 percent of the gap.

Column 4 shows results based on using the plausible values (PVs) reported in NALS. The PVs in NALS are based on a conditioning model that includes the first approximately 100 principal components from several hundred main effects and interactions of background variables collected in the survey. Importantly, the conditioning set includes measures of individual wages (the outcome variable in the research model above) as well as other highly related measures such as family income and occupation, though the complex conditioning procedure makes it difficult to understand the functional form assigned to the relationship between ability and wages. As noted above, Schofield et al. (2015) demonstrate that use of PVs as an explanatory variable in such cases will result in bias if the functional form does not conform to that used in the research model. Indeed, the race coefficient in column 4 is -0.121 compared with the MESE estimate of -0.094 in column 3.

Unfortunately, it is unusual for analytic samples to include item-level responses. One approach to unbiased estimation in the independent variable case that does not require item-level data is to instrument for a noisy measure of ability, $\hat{\theta}_i$, with a second, independent, measure. This is of course only feasible if two independent measures are available, but in many circumstances this might be possible. For example, the NAEP test consists of two separate blocks of items; one could use the fraction correct from the first block as an instrument for the fraction correct on the second, perhaps rescaling each to correspond to the desired θ_i scale. Alternatively, many tests report subscores for particular subdomains (e.g., for reading comprehension and vocabulary within an ELA test). If one is willing to assume that none of the subdomain proficiency levels are related to Y_i conditional on the composite score, one subdomain score can be used as an instrument for another.[19]

[19] As discussed above, some testing systems "borrow" information from one domain in computing a student's score on another. In this case, the error in the subscores is not independent across

When the secondary researcher has access to two independent measures that are each unbiased estimates of θ_i (such as ML estimates of ability), this IV approach will produce unbiased estimates of δ and γ. The first stage from this regression has an intuitive interpretation: The $\hat{\theta}_i$ coefficient estimates $R_{\theta|X}$, the conditional reliability of the measured test score given X_i, and the fitted values from this first stage are posterior means conditional on X_i and the first θ_i measure. When the two measures are biased but independent estimates of θ_i (e.g., if the two are posterior means on different subdomains, with no cross-subdomain pooling), the IV approach can still be used to identify γ, though the IV estimate of δ will be biased due to the rescaling of θ. [20]

4.4 Subscore comparisons

A useful illustration of the practical importance of the measurement issues we consider here is the estimation of sub-score impacts. A common topic for investigation is whether a program or treatment has differential effects on different areas of competency. For example, Dee and Jacob (2011) estimate the effect of school accountability under the No Child Left Behind law on student performance across different math and reading subscales, looking for evidence that accountability led to changes in instructional emphasis. Unfortunately, the methods by which subscale scores are computed on most assessments are not well suited to support these kinds of analyses, and can be expected to bias them toward a conclusion of common effects across subscales.

First, when test-makers are selecting among candidate test items, they evaluate items in part based on their association with examinees' overall performance.[21] These evaluations treat achievement as unidimensional. Thus, items that successfully identify students who are strong on one dimension but weak on other dimensions will tend to be rejected.[22]

---

subdomains.

[20] For example, if the two measures are posterior means with uniform shrinkage factors λ, δ' will equal δ/λ. But γ' will be unbiased: Merely rescaling one explanatory variable does not affect the coefficients on other explanatory variables.

[21] In the notation of the 3PL model introduced above, they look for items with large discrimination parameters and item characteristic curves that have the logistic shapes predicted by the model, using a uni-dimensional model for θ.

[22] A related problem, which we do not explore in depth, is the consideration of the relative performance of different groups of students when selecting items. For example, items are commonly

Second, because the number of items relevant to any particular subscale is typically very small (often only 5-10), the measurement issues discussed in Section 3 become much more important. Estimates of student performance based solely on the items from a single subscale would be extremely unreliable. Skorupski (2008) examines subscale scores from four state assessment systems and finds that these scores have reliabilities in the 0.4 to 0.7 range. This is too low to be practically useful for individual assessments, particularly because the quantity of interest in subscale scoring is a student's *relative* strength on one subdomain relative to others. As individual proficiency is likely to be strongly positively correlated across subscales, the reliability of between-subscale differences in performance, computed only from subscale items, is extremely low. For example, with subscale reliabilities of 0.6 and a correlation of true proficiency across subscales of 0.8, the reliability of the estimated difference in a student's proficiency between two subscales is only 0.23.

As a consequence, subscale scores are rarely based solely on the items corresponding to the subscale. Exact procedures vary, but in one way or another many measurement systems "borrow" information about an examinee's performance from other sources, either the examinee's performance on other subscales or the performance of other examinees with similar observables (via the use of conditioning variables as discussed in Section 3.2.1). The unreliability of the raw subscale scores means that the amount of shrinkage is substantial.

The NAEP illustrates this. As discussed above, the NAEP relies on a conditioning model with a long list of student and school conditioning variables. In the math assessment, the questions are divided up into five subscales, with only a handful of questions from each, and the ability parameter θ_i in equation (7) is five dimensional with an unrestricted variance-covariance matrix Σ. In practice, the estimated correlations are very high – many around 0.99. As a consequence, while subscale proficiencies are allowed to have different relationships to the conditioning variables, a question on one subscale counts essentially as much toward the student's posterior mean on other

discarded if they show larger-than-expected black-white gaps. This is sensible in isolation. But since the black-white gap is to be estimated from the selected items, a practice of discarding items with unusual gaps is likely to bias it.

subscales as it does toward the posterior mean on the subscale to which the question is associated. This means that there is little if any scope to identify differential treatment effects of programs like NCLB school accountability on the different subscales.

Other assessments use different methods, but the implications are the same. Skorupski (2008) discusses several approaches to measuring subscale performance, and finds that all yield subscale scores that are correlated 0.97 or higher across subscales.

5 Conclusions

Modern psychometrics utilizes a variety of sophisticated models and techniques to develop cognitive assessments and produce individual ability scores. The applied researcher who does not possess at least a rudimentary understanding of these methods is liable to mis-use test scores in a way that can lead to serious biases. Perhaps most importantly, researchers need to pay close attention to how the individual score measures are generated. If one uses ability as a dependent variable, it is critical to know whether the assessment reports a raw score, a maximum likelihood of individual ability, or a Bayesian measure such as a posterior mean. When Bayesian ability measures are used as the outcome, model coefficients typically will be biased, though it may be possible to adjust the results to eliminate this bias. If ability is used as independent variable, each of the available measurement choices will create bias in coefficients, both that on the ability measure and those for other explanatory variables, but the biases will depend importantly on the type of test score measure employed. Here too information provided by the assessment – namely, the reliability of the ability measure – can in some cases be used in an errors-in-variables framework to generate consistent estimates, or, even better, the reported scores can be discarded in favor of analyses that draw directly on examinees' item responses.

Scaling also presents challenges to researchers who use cognitive ability measures. Practitioners and researchers routinely use test scores in a way that assumes they have interval properties. As discussed above, there is no compelling justification for this assumption. Given this inherent disconnect, how should the applied researcher proceed? Should one only use the ordinal infor-

mation contained in test scores, and forgo making any statements about the magnitude of effects, for example? While we understand this inclination, we are inclined to be less nihilistic. We are inclined toward the approach outlined by Nielsen (2015), who seeks to narrow the class of scale transformations that are considered reasonable. But at a minimum, we recommend that researchers make greater effort to test the robustness of their results to changes in the test score scale. In the context of a randomized program evaluation, for example, using only the ordinal nature of test scores one can calculate at what point in the control group distribution the median treatment student would fall. More generally, a P-P plot comparing the treatment and control groups would allow the researcher to fully characterize how the two distributions compare without relying on a particular scale. Where analyses will use the scale scores as interval measures, researchers might test their sensitivity to modest scale transformations such as the log of the reported scale score or its inverse transformation, $\exp(\theta)$.

The common practice of standardizing reported scores by dividing by the standard deviation, converting to normal curve equivalents, or constructing percentile scores, raises particular concerns. Each of these depends critically on the measurement properties of the score and on the sample used for the standardization. So, what practical guidance can be given to the applied researcher? A useful rule of thumb is that z-scores and effect sizes should always be computed from an estimate of σ_θ rather than from $\sigma_{\hat{\theta}}$. The former cannot be computed directly from the analysis sample (except when plausible value scores are reported) but can often be computed from information – such as, for example, the test-retest reliability of $\hat{\theta}^{ML}$, which estimates $\sigma^2_\theta/\sigma^2_{\hat{\theta}}$ – reported in the assessment's technical documentation. Moreover, secondary researchers should use a σ_θ that pertains to the broadest possible population, even if their study focuses on a more homogenous subpopulation, and comparisons of standardized effect sizes across studies should account for differences in the populations used to construct σ_θ.[23] A large effect size computed using a subpopulation estimate of σ_θ may actually correspond to a *smaller* effect than that of an intervention

[23] Both of these guidelines apply as well to the computation and analysis of percentiles or normal curve equivalents; in these cases, the distribution function of θ (or an approximation to it) is required, and is equally sensitive to the choice of population.

with a smaller reported effect size that is based on a more representative population's σ_θ.

The landscape of testing in U.S. schools is changing rapidly, driven by the widespread adoption of the Common Core State Standards (CCSS). The Common Core is an unprecedented effort, led by consortia of states with strong encouragement from the federal government, to develop and implement a set of common academic standards that will be used across state borders. The goal is twofold: to provide a consistent framework, in the absence of any national curriculum or testing system, like those that exist in many countries; and to emphasize the knowledge and skills that students will need in order to be "college and career ready" by the time they graduate high school. The standards articulate in some detail what students should know and be able to do in each grade and subject in elementary and secondary school. A theme running through them is reduced emphasis on memorization and rote computation, in favor of more problem-solving and higher-order thinking. As of August 2015, 42 states and the District of Columbia had adopted the CCSS in English Language Arts (ELA) and math.

The transition to the Common Core state standards (CCSS) has been accompanied by the introduction of new assessments in most states. States grouped into two consortia to develop common assessments designed to measure student mastery of the new standards: the Partnership for Assessment of Readiness for College and Careers (known as PARCC) and the Smarter Balance Assessment Consortium (known as "Smarter Balance" or SBAC). In Spring 2015, 11 states administered the PARCC assessments in Spring 2015 and 18 states administered the Smarter Balance tests.

The standards and associated assessments have become the source of considerable controversy, with many stakeholders resisting the diminished local control which would result. While several states have indicated that they will no longer administer the common assessments, in most cases these states are incorporating large segments of the new assessments under the "brand" of a state's own test. In any case, our view is that these assessments – or some large component of them – are likely to become the predominant student assessments used in elementary and secondary schools over the next decade, and it is thus important for researchers to understand them.

The new assessments have much in common (see Table 6 for a summary of the two tests). Both are administered on computers, incorporate performance-based tasks and open (constructed) response items along with standard multiple-choice items, and include mid-year interim assessments designed to provide teachers with information on students' strengths and weaknesses. Both assessments rely on sophisticated IRT models both to generate the exams and to calculate estimates of individual proficiency. The Smarter Balance assessment reports maximum likelihood scores, linearly transformed to range between 2,000 and 3,000 and to be vertically comparable across grades.[24] PARCC has reported much less detail about its scoring procedures, but appears to report ML score estimates as well. The Smarter Balance tests (though not the PARCC assessments) are computer adaptive, so that a student who does well on early items is routed to harder items later in the test. This can allow for more efficient estimation of student proficiency by ensuring that students are given many items that are appropriately difficult for them, but makes the resulting scores much more model dependent and sensitive to the IRT specification and measurement model. Both tests will report sub-scale scores as well as overall scores, and as on other tests students' relative proficiency on different subscales is likely to be quite imprecisely measured.

The measurement and scaling issues discussed here are not likely to go away anytime soon. Indeed, just the opposite. Student tests are continuing to see wider use in empirical economics research. Moreover, the push for accountability in higher education is leading some to advocate for the development of standardized assessments aimed at college students which will no doubt rest on the same psychometric foundations outlined in this article. And psychometric methods are are spreading beyond the realm of cognitive skill assessment. Common measures of "non-cognitive" traits such as persistence, self-esteem, and socio-emotional regulation, as well as of more cognitive traits such as working memory, rely on the same IRT measurement models discussed above, typically applied to batteries of very few survey items (Schofield, Forthcoming). Test score-like measures are also being used in health, as health care reform has encouraged increased empha-

[24] We have not discussed vertical equating in detail. As when tests claim to use interval scales, however, researchers should be wary of the assumptions behind a claim that a test uses a vertical scale.

sis on quantitative measurement. Finally, the rise of Empirical Bayes methods for measuring the contribution of teachers to students or firms to workers' wages has brought psychometric-like measures to data sets used in a wide range of recent empirical economics research. Across all of these domains, secondary researchers will need to account more carefully for measurement processes than has been typical in the past in order to draw appropriate conclusions.

Bibliography

References

Angrist, Joshua D., Atila Abdulkadiroglu, Susan M. Dynarski, Thomas J. Kane, and Parag A. Pathak, "Accountability and Flexibility in Public Schools: Evidence from Boston's Charters And Pilots," *The Quarterly Journal of Economics*, 2011, *126* (2), 699–748.

Barlevy, Gadi and Derek Neal, "Pay for Percentile," *American Economic Review*, August 2012, *102* (5), 1805–1821.

Blau, Francine D and Lawrence M Kahn, "Do cognitive test scores explain higher US wage inequality?," *Review of Economics and Statistics*, 2005, *87* (1), 184–193.

Bond, Timothy N. and Kevin Lang, "The Evolution of The Black-White Test Score Gap in Grades K-3: The Fragility of Results," *The Review of Economics and Statistics*, 2013, *95* (5), 1468–1479.

_ **and** _ , "The Black-White Education-Scaled Test-Score Gap in Grades K-7," October 2015.

Briggs, Derek C., "Using Explanatory Item Response Models to Analyze Group Differences in Science Achievement," *Applied Measurement in Education*, 2008, *21* (2), 89–118.

Carstens, R and D Hastedt, "The effect of not using plausible values when they should be: an illustration using TIMSS 2007 grade 8 mathematics data," in "4th IEA International Research Conference (IRC-2010) at the University of Gothenburg, Sweden" 2010.

Cascio, Elizabeth U. and Douglas O. Staiger, "Knowledge, Tests, and Fadeout in Educational Interventions (Working Paper No. 18038)," Technical Report, National Bureau of Economic Research 2012.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff, "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, September 2014, *104* (9), 2593–2632.

_ , _ , and _ , "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, September 2014, *104* (9), 2633–79.

Cunha, Flavio and James J. Heckman, "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation," *The Journal of Human Resources*, 2006, *43* (4), 738–782.

_ , _ , and Susanne M. Schennach, "Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Econometrica*, May 2010, *78* (3), 883–931.

Dee, Thomas S and Brian Jacob, "The impact of No Child Left Behind on student achievement," *Journal of Policy Analysis and Management*, 2011, *30* (3), 418–446.

Embretson, Susan E. and Steven Paul Reise, *Item Response Theory for Psychologists Multivariate, Applications Book Series*, Lawrence Erlbaum Associates, Inc., 2000.

Ferrer, Ana, David A Green, and W Craig Riddell, "The effect of literacy on immigrant earnings," *Journal of Human Resources*, 2006, *41* (2), 380–410.

Figlio, David N. and Cecilia Elena Rouse, "Do Accountability and Voucher Threats Improve Low-Performing Schools?," *Journal of Public Economics*, January 2006, *90* (1-2), 239–255.

Flynn, James R., "Massive IQ Gains in 14 Nations: What IQ Tests Really Measure," *Psychological Bulletin*, March 1987, *101* (2), 171–191.

_ , *What Is Intelligence?: Beyond the Flynn Effect*, paperback ed., Cambridge University Press, 2009.

Glas, Cees AW and Jonald L Pimentel, "Modeling nonignorable missing data in speeded tests," *Educational and Psychological Measurement*, 2008, *68* (6), 907–922.

Hart, Ray, Michael Casserly, Renata Uzzell, Moses Palacios, Amanda Corcoran, and Liz Spurgeon, "Student Testing in America's Great City Schools: An Inventory and Preliminary Analysis," Technical Report, Council of Great City Schools October 2015.

Heckman, James J., Seong Hyeok Moon, Rodrigo Pinto, Peter A. Savelyev, and Adam Yavitz, "Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program," *Quantitative Economics*, 2010, *1* (1), 1–46.

Ho, Andrew D. and Edward H. Haertel, "Metric-Free Measures of Test Score Trends and Gaps with Policy-Relevant Examples (CSE Report 665)," Technical Report, Graduate School of Education & Information Studies University Of California, Los Angeles 2006.

Ho, Andrew Dean, "A Nonparametric Framework for Comparing Trends and Gaps Across Tests," *Journal of Educational and Behavioral Statistics*, June 2009, *34* (2), 201–228.

Holman, Rebecca and Cees AW Glas, "Modelling non-ignorable missing-data mechanisms with item response theory models," *British Journal of Mathematical and Statistical Psychology*, 2005, *58* (1), 1–17.

Jackson, C. Kirabo, Jonah E. Rockoff, and Douglas O. Staiger, "Teacher Effects and Teacher-Related Policies," *Annual Review of Economics*, 2014, *6*, 801–825.

Junker, Brian, Lynne Steuerle Schofield, and Lowell J Taylor, "The use of cognitive ability measures as explanatory variables in regression analysis," *IZA Journal of Labor Economics*, 2012, *1* (4), 1–19.

Kleiner, Morris M. and Alan B. Krueger, "Analyzing the Extent and Influence of Occupational Licensing on the Labor Market," *Journal of Labor Economics*, April 2013, *31* (2), S173–S202.

Krueger, Alan B., "Experimental Estimates of Education Production Functions," *The Quarterly Journal of Economics*, May 1999, *114* (2), 497–532.

Lafortune, Julien, Jesse Rothstein, and Diane Whitmore Schanzenbach, "School Finance Reform and the Distribution of Student Achievement," 2015.

Mislevy, Robert J, "Randomization-based inference about latent variables from complex samples," *Psychometrika*, 1991, *56* (2), 177–196.

_ , Albert E Beaton, Bruce Kaplan, and Kathleen M Sheehan, "Estimating population characteristics from sparse matrix samples of item responses," *Journal of Educational Measurement*, 1992, *29* (2), 133–161.

Morris, Carl N., "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association*, 47-55 1983, *78* (381), 1983.

Neal, Derek A and William R Johnson, "The Role of Premarket Factors in Black-White Wage Differences," *The Journal of Political Economy*, 1996, *104* (5), 869–895.

Neal, Derek and Diane Whitmore Schanzenbach, "Left behind by design: Proficiency counts and test-based accountability," *The Review of Economics and Statistics*, 2010, *92* (2), 263–283.

Nielsen, Eric R., "Achievement Gap Estimates and Deviations from Cardinal Comparability," in "Finance and Economics Discussion Series 2015-040," Board of Governors of the Federal Reserve System, 2015.

Reardon, Sean, "Differential growth in the Black-White achievement gap during elementary school among initially high-and low-scoring students," *Institute for Research on Education Policy & Practice Working Paper*, 2008, *7*.

Rothstein, Jesse, "Teacher quality in educational production: Tracking, decay, and student achievement," *The Quarterly Journal of Economics*, 2010, *125* (1), 175–214.

_ , "Revisiting the Impact of Teachers," 2015.

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

Rubin, Donald B, "Multiple Imputation after 18+ years," *Journal of the American Statistical Association*, 1996, *91* (434), 473–489.

Schofield, Lynne Steuerle, "Correcting for Measurement Error in Latent Variables Used as Predictors," *Annals of Applied Statistics*, Forthcoming.

_ **, Brian Junker, Lowell J. Taylor, and Dan A. Black**, "Predictive Inference Using Latent Variables with Covariates," *Psychometrika*, September 2015, *80* (3), 727–747.

Shapiro, Gary, Pam Broene, Frank Jenkins, Philip Fletcher, Liz Quinn, Janet Friedman, Janet Ciarico, Monica Rohacek, Gina Adams, and Elizabeth Spier, "Head Start Impact Study Final Report," Technical Report, U.S. Department of Health and Human Services, Administration for Children and Families January 2010.

Skorupski, William P., "A Review and Empirical Comparison of Approaches for Improving the Reliability of Objective Level Scores," Technical Report, Council of Chief State School Officers August 2008.

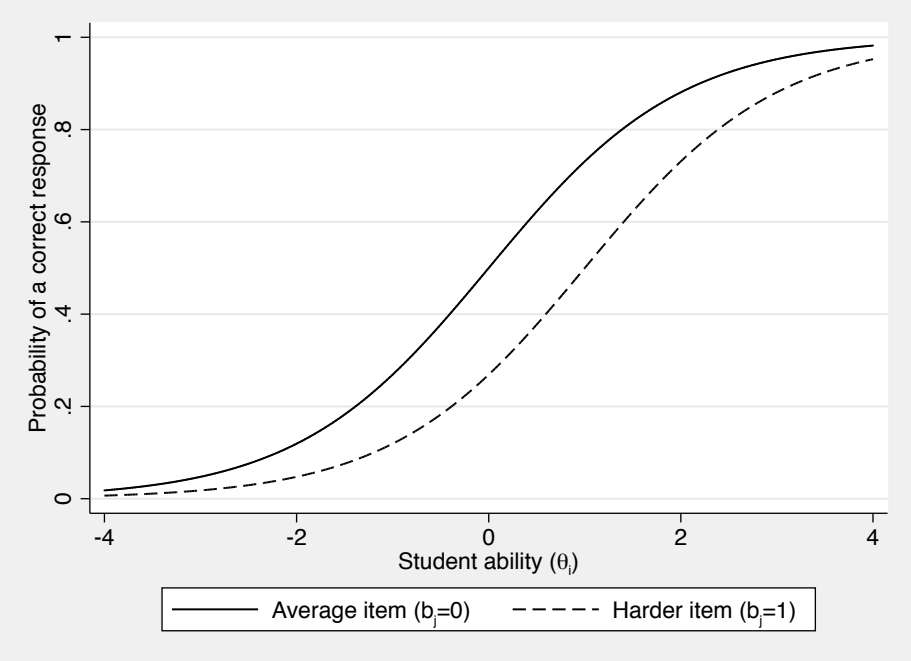
Stevens, S. S., "On the Theory of Scales of Measurement," *Science*, June 1946, *103* (2684), 677–680.

Thorndike, Robert L., "Intellectual Status and Intellectual Growth," *Journal of Educational Psychology*, June 1966, *57* (3), 121–127.

van der Linden, Wim J and Ronald K Hambleton, *Handbook of Modern Item Response Theory*, Springer, 1997.

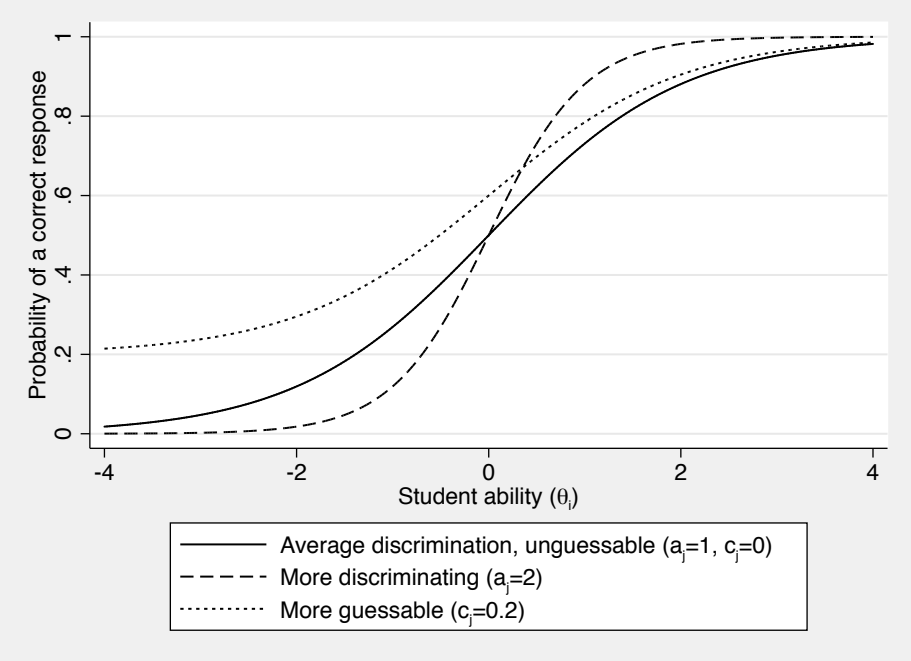
von Davier, Matthias, E Gonzalez, and R Mislevy, "What Are Plausible Values and Why Are They Useful?," Monograph, IERI 2009.

Figure 1. Item Characteristic Curves for 1-parameter logistic (Rasch) IRT items



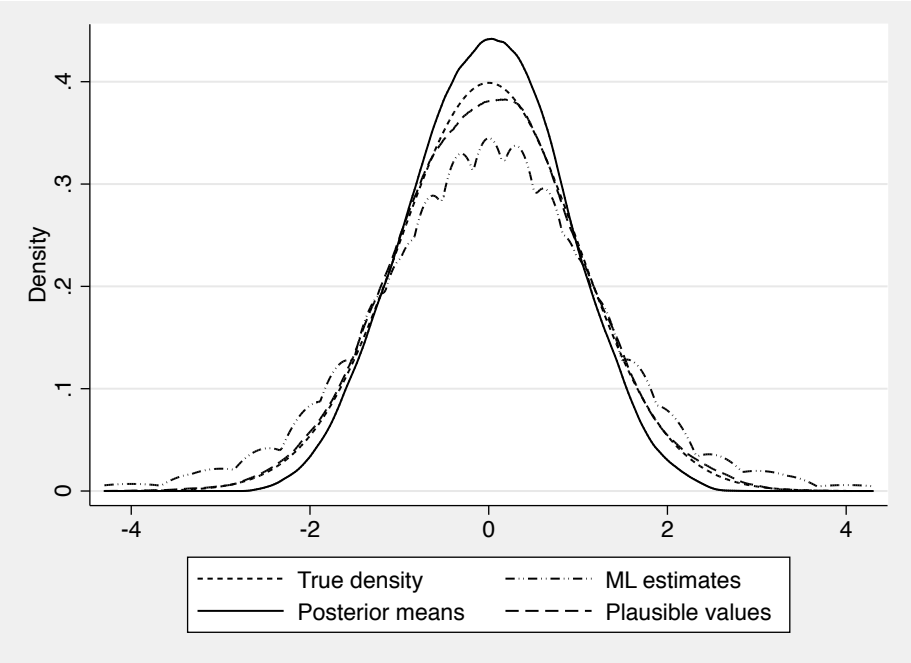
Notes: Series show the probability of a correct response as a function of the examinee's ability (θ) under the Rasch IRT model (equation 4).

Figure 2. Item Characteristic Curves for 3-parameter logistic IRT items



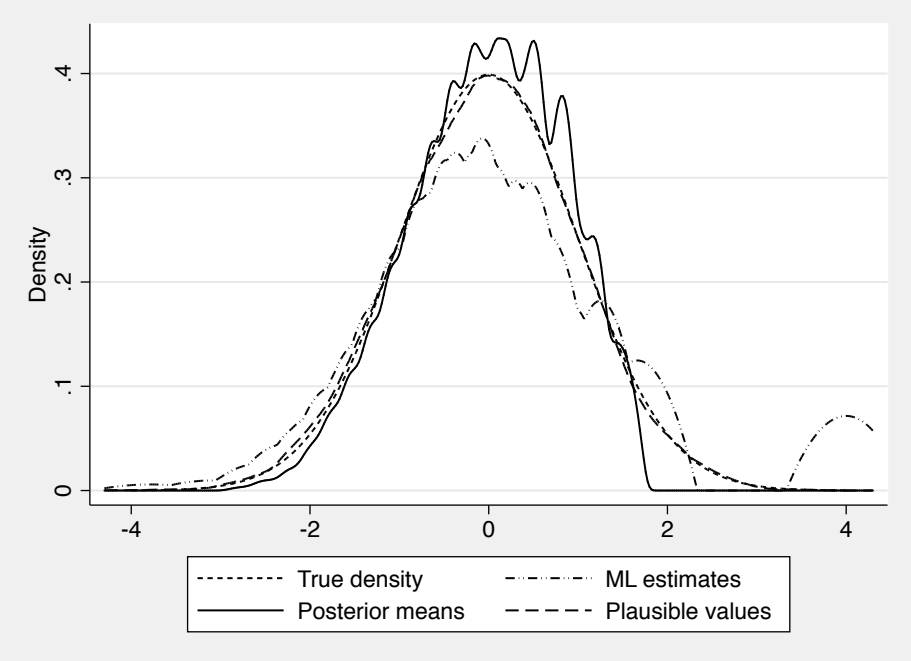
Notes: Series show the probability of a correct response as a function of the examinee's ability (θ) under the 3PL IRT model (equation 6). All items are assumed to have $b_j = 0$.

Figure 3. Distribution of reported scores under different measurement models



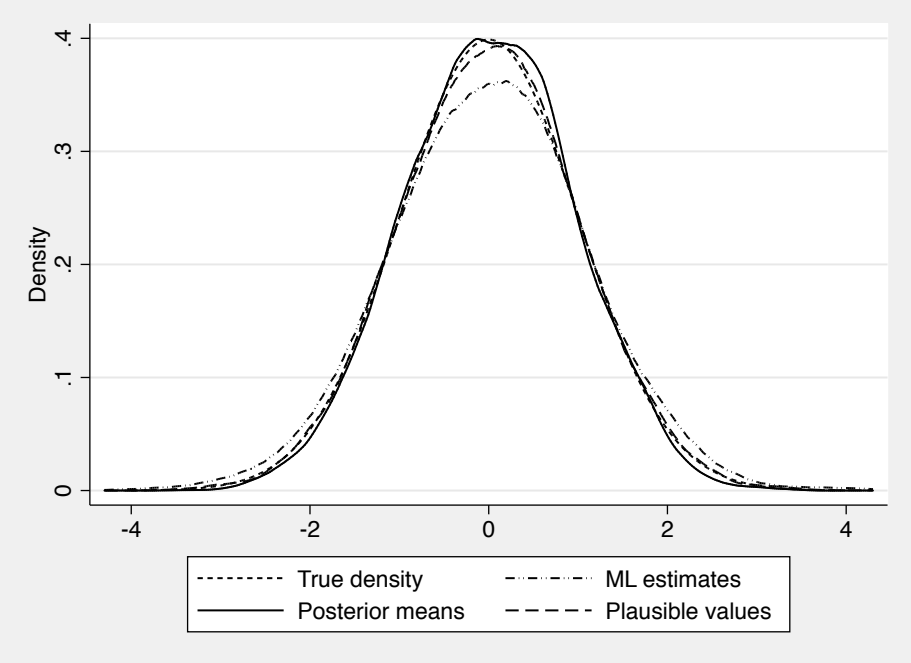
Notes: True θ distribution is assumed to be standard normal. Exams consist of 15 Rasch items with b_j drawn randomly from U[-2,2], independently for each examinee. ML estimates assign scores of -4 to students answering all items incorrectly, and +4 to students answering all items correctly. Kernel density estimates are plotted, using bandwidths of 0.3 (ML estimates) or 0.15 and Epanechnikov kernels.

Figure 4. Distribution of reported scores under different measurement models: Easy tests



Notes: True θ distribution is assumed to be standard normal. Exams consist of 15 Rasch items with b_j drawn randomly from $N(-1, 0.25^2)$, independently for each examinee. ML estimates assign scores of -4 to students answering all items incorrectly, and +4 to students answering all items correctly. Kernel density estimates are plotted, using bandwidths of 0.3 (ML estimates) or 0.15 and Epanechnikov kernels.

Figure 5. Distribution of reported scores under different measurement models: Long tests



Notes: True θ distribution is assumed to be standard normal. Exams consist of 50 Rasch items with b_j drawn randomly from U[-2,2], independently for each examinee. ML estimates assign scores of -4 to students answering all items incorrectly, and +4 to students answering all items correctly. Kernel density estimates are plotted, using bandwidths of 0.3 (ML estimates) or 0.15 and Epanechnikov kernels.

Table 1: Distributional assumptions used for simulations of secondary analyses of test scores

Primitive error terms {X,e,u,v}

	Expectation	Variance	SD
	0.00	0.50	0.71

Correlation

	X	e	u	v
X	1	0	0	0
e	0	1	0.50	0
u	0	0.50	1	0.50
v	0	0	0.50	1

Composites

	Expectation	Variance	SD
θ=X+e	0	1	1
W=X+u	0	1	1
Y=4+X+θ+v	4	3	1.73

Correlation

	X	θ	W	Y
X	1	0.71	0.71	0.82
θ	0.71	1	0.75	0.87
W	0.71	0.75	1	0.87
Y	0.82	0.87	0.87	1

Simulation

# examinees	10,000
# test items	20
Item specifications	Rasch, b=0

Table 2. Simulations of research models using test scores as dependent variables

Conditioning variables:	(1) None	(2) W	(3) X	(4) W,X
True theta				
X	1.008			
	(0.010)			
Maximum likelihood estimate				
X	1.016			
	(0.012)			
Posterior means				
X	0.816	0.931	1.021	1.021
	(0.010)	(0.009)	(0.008)	(0.008)
Plausible values				
X	0.817	0.923	1.022	1.012
	(0.012)	(0.011)	(0.010)	(0.010)

Notes: Each entry represents the X coefficient from a separate regression, with the indicated test score measure as the dependent variable. See Table 1 for description of simulation sample. N=10,000. ML estimates are computed using maximum and minimum obtainable scores of +2 and -2, respectively.

Table 3. Simulations of regressions with ability as an independent variable

Conditioning variable	(1) None	(2) X	(3) X,Y	(4) X,ln(Y)	(5) X,W	(6) X,W,Y
Panel A - True theta						
θ	0.989					
	(0.010)					
X	1.029					
	(0.014)					
Panel B - Maximum likelihood estimate						
θ	0.645					
	(0.009)					
X	1.371					
	(0.015)					
Panel C - Posterior means						
θ	0.809	0.972	1.301	1.180	1.113	1.301
	(0.012)	(0.014)	(0.010)	(0.011)	(0.012)	(0.010)
X	1.366	1.034	0.698	0.801	0.890	0.698
	(0.015)	(0.018)	(0.013)	(0.015)	(0.016)	(0.013)
Panel D - Plausible values						
θ	0.560	0.640	0.974	0.858	0.775	0.966
	(0.010)	(0.012)	(0.010)	(0.011)	(0.012)	(0.010)
X	1.569	1.372	1.043	1.132	1.242	1.030
	(0.015)	(0.018)	(0.014)	(0.016)	(0.017)	(0.014)

Notes: Each panel and column reports a separate regression with the same dependent variable, using the indicated test score measure as an independent variable and X as a control. See Table 1 for description of simulation sample. N=10,000. ML estimates are computed using maximum and minimum obtainable scores of +2 and -2, respectively.

Table 4 - Biases when using estimates of latent ability as a dependent variable (from Briggs 2008)

Panel A: Dependent Variable = Overall Science Score

	Scale scores		Z scores	
	(1)	(2)	(3)	(4)
	PM	MML	PM	MML
Intercept	0.9	0.96	1.29	1.19
Black	-0.61	-0.77	-0.87	-0.95
Hispanic	-0.52	-0.67	-0.75	-0.83
Asian	-0.1	-0.115	-0.14	-0.14
Other	-0.3	-0.373	-0.43	-0.46
SD	0.7	0.81	1	1

Panel B: Dependent Variable = Science Subscore

	Life science	Physical science	Life science	Physical science
	Scale scores		Z scores	
	(1)	(2)	(3)	(4)
	PM	MML	PM	MML
Intercept	0.43	0.91	0.5	1.01
Black	-0.42	-0.58	-0.38	-0.97
Hispanic	-0.36	-0.48	-0.61	-0.71
Asian	-0.07	-0.08	-0.25	0.04
Other	-0.22	-0.31	0.02	0.5
SD	0.481	0.661	0.58	0.75

Notes: Estimates reproduced from Briggs (2008). N = 433. Columns 1 and 3 report estimates when posterior means (without conditioning variables) are used as the dependent variable; Columns 2 and 4 report estimates obtained via the Marginal Maximum Likelihood method discussed in the text. Briggs does not report standard errors, but all Intercept, Black, and Hispanic coefficients are significantly different from zero at the 1% level, while none of the Asian or Other coefficients are reported to be significant at the 5% level.

Table 5 - Biases when using estimates of latent ability as an independent variable (from Junker et al. 2012)

	Dependent variable = log(weekly wage)			
	Estimate of literacy skill used in model			
	No skill control (1)	MLE of literacy score (2)	MESE (3)	PVs (4)
Black	-0.366	-0.144	-0.094	-0.121
	(0.033)	(0.033)	(0.033)	(0.041)
Literacy skill		0.151	0.191	0.221
		(0.008)	(0.010)	(0.015)
Effect of a one SD change in literacy skill		0.19	0.218	0.221

Notes: Estimates reproduced from Junker et al. (2012). N = 3,267. MESE = Mixed Effects Structural Equations. PV = Plausible Values. See text for detailed description of estimated models.

Table 6. Common Core Assessments

	PARCC	*Smarter Balanced*
States Administering Test in Spring 2015[1]	Arkansas*, Colorado, Illinois, Louisiana, Maryland, Massachusets^, Mississippi*, New Jersey, New Mexico, Ohio*, Rhode Island	California, Connecticut, Delaware, Hawaii, Idaho, Maine*, Michigan, Missouri*, Montana, Nevada, New Hampshire, North Dakota, Oregon, South Dakota, Vermont, Washington, West Virginia, Wisconsin*
Test duration (Math and ELA combined)	8 - 10 hours	7 - 8.5 hours
Format	Computer-based, non-adaptive	Computer adaptive
Item Types[2] [3]	**For ELA:** Evidence-based selected response (EBSR), Technology-enhanced constructed-response **For Math:** Tasks assessing concepts, skills and procedures (Type I); Tasks assessing expressing mathematical reasoning (Type II); Tasks assessing modeling / applications (Type III)	Selected Response, Constructed Response, Extended Response, Performance Tasks
Psychometric contractor	Educational Testing Service	Educational Testing Service (design) American Institutes for Research (implementation)
Scaled score range	650 - 850	2000 - 3000
Vertically equated scale?	Yes	Yes
Calculation of scaled score [4,5]	Unknown	Maximum likelihood estimation is used to calculate a theta score. The final vertical scale score is the linear transformation of the post-vertically scaled IRT ability estimate (theta score).

Notes:

*State has subsequently left this testing consoritium, though in many cases the state plans to use substantial portion of the consortium's assessment.

^Massachusetts allowed districts to choose between giving their own tests and PARCC, so not all districts administered the PARCC exam in spring 2015

Sources:

1. Ujifusa, A. (2015, November 16). Common Core's Big Test: Tracking 2014-15 Results - Education Week. Education Week.

2. Ford, L. A., Michaels, H. R., & Johnston-Fisher, J. L., (2015). Examination of Test Construction

3. "Sample Items and Performance Tasks." Sample Items and Performance Tasks. Smarter Balanced Assessment Consortium, n.d. Web. 18 Dec. 2015. <http://www.smarterbalanced.org/sample-items-and-performance-tasks/>.

4. PARCC: Technical Memorandum for Field Test Phase. Tech. N.p.: Educational Testing Service, 2014. Print. Materials across Multiple Assessment Programs. Washington,

5. *Smarter Balanced Scoring Specification 2014-2015 Administration*, American Institutes for Research.