The Measurement of Student Ability in Modern Assessment Systems

Brian Jacob and Jesse Rothstein*

June 2016

Economists often use test scores to measure a student's performance or an adult's human capital. In the research literature on the economics of education, student test scores are often used to estimate teacher effectiveness, or "value-added" (for example, Chetty et al. 2014a); to measure and attempt to explain the black-white achievement gap (for example, Fryer and Levitt 2004, 2006, 2013; Rothstein and Wozny 2013); or to measure the impacts of state- or district-level educational policy choices such as finance or accountability rules (for example, Dee and Jacob 2011; Lafortune et al. 2016). In the broader labor economics literature, test scores are often used as well as proxies for human capital, for example in examining the black-white wage gap conditional on cognitive ability as in Neal and Johnson (1996).

In our experience, many researchers think of an individual's score as a noisy but unbiased measure of true ability like, for example, the simple fraction of test items a

* Brian Jacob is Walter H. Annenberg Professor of Education Policy, Professor of Economics, and Professor of Education at the University of Michigan, Ann Arbor, Michigan. Jesse Rothstein is Professor of Public Policy and Economics and Director of the Institute for Research on Labor and Employment (IRLE) at the University of California, Berkeley, California. Both authors are Research Associates, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are bajacob@umich.edu and rothstein@berkeley.edu.

student answers correctly. Unfortunately, the student achievement measures provided in modern assessment systems are rarely – if ever – so straightforward. Assessments commonly have multiple forms and are often adaptive, meaning that the questions students receive are based on their performance on previous questions. As a result, students are frequently presented with different questions that may not be of comparable difficulty. Moreover, modern test-making practice disparages simple summaries like the fraction correct in favor of estimates from complex statistical models that attempt to extract more information from the pattern of correct and incorrect responses. The scores that these models produce are generally not unbiased measures of student ability, and may not be suitable for many secondary analyses that economists would like to perform.

Consider the well-known National Assessment of Educational Progress (NAEP, also known as "the Nation's Report Card"). A little-known fact is that the scores computed for students who take the NAEP depend not only on the examinees' responses to test items but also on their background characteristics, including race and gender. As a consequence, if a black and white student respond identically to questions on the NAEP assessment, the reported ability for the black student will be lower than for the white student—reflecting the lower average performance of black students on this assessment. This adjustment does not affect reported aggregate statistics such as the unconditional black-white test score gap, but, as we explain below, it can introduce important biases into many secondary analyses. Other testing systems do not incorporate students' background characteristics into their scores, but report posterior mean scores for students that are biased estimates of the students' ability, and therefore unsuitable for many of the secondary analyses that economists perform, which typically use the test scores as

dependent variables (for example, to estimate the effects of programs or even just the black-white test score gap).

Even in the relatively rare case that the underlying student ability measure comes from a simple statistic such as the fraction correct, assessments often present transformed "scale" scores for each individual. Research using these test scores virtually always assumes that the ability measure has an interval property – that is, a one unit change has the same meaning at every point on the scale (e.g., an increase from 400 to 450 on the SAT represents the same improvement in student knowledge as an increase from 700 to 750). However, as explained below, this assumption is entirely unwarranted. This fact, widely recognized in the testing community but often ignored, undermines many of the purposes to which test scores have been put.

And, finally, the fact that test scores are inherently "noisy" measures of student ability has important implications for analyses that use the scores as explanatory variables, such as in wage regressions. As we discuss in more detail below, a recent paper demonstrates that the failure to properly account for measurement error in individual ability, when this is used as a control in a standard wage regression, would lead an analyst to overstate the black-white wage gap conditional on ability by nearly 50 percent (Junker et al. 2012).

Our goal in this paper is to familiarize applied economists with the construction and properties of common cognitive score measures and with their potential implications for economics research using these measures. Information about how scores are constructed is often buried deep in technical manuals, if presented at all. While the literature in psychometrics (the field concerned with the theory and methodology of

psychological measurement) has explored many if not all of the issues that we discuss, economists and other applied researchers are generally unaware of them and frequently misuse test score measures, with potentially serious consequences for their analyses. These issues will become even more important in the coming years as new assessments, developed in conjunction with the new Common Core State Standards, are gradually rolled out in schools around the country.

We begin by discussing the domain covered by a test, and then the problem of assigning a quantitative scale to latent student ability.[1] We next turn to the statistical models used to convert examinees' responses to a series of test items into scores on the chosen scale. We then discuss the secondary analysis of test scores, when test scores are used as either dependent or explanatory variables, focusing in particular on how the test's measurement model can influence results. We attempt to provide both applied researchers and research consumers with practical guidance for evaluating the many research studies that use test-based cognitive ability measures.

What Does the Test Measure?

The first decision that must be made in designing a test concerns what is to be measured. Historically, psychometricians have distinguished between tests of aptitude and achievement. IQ tests (like the Wechsler Intelligence Scale for Children or Raven's Progressive Matrices) are designed to measure mental aptitude, conceptualized as a fixed

[1] Throughout this paper, we use "ability," "proficiency," "achievement," and "aptitude" interchangeably to refer to a latent trait that governs test performance. In many contexts these terms have distinct meanings—for example, some argue that IQ tests measure innate aptitude but not learned achievement—but such distinctions are not important for the purposes of this paper.

trait that is unaffected by educational interventions. Respondents recite long strings of digits from memory or recognize patterns in abstract figures. By contrast, achievement tests aim to capture an individual's stock of accumulated knowledge and not his or her innate ability.

The distinction between aptitude and achievement is not always clear, however. IQ scores are affected by educational interventions such as preschool attendance (Heckman et al. 2010) or by the amount of accumulated schooling (Cascio and Lewis 2006), though one might expect innate aptitude to be invariant to both. The Peabody Picture Vocabulary Test (PPVT), which is administered to children in the National Longitudinal Survey of Youth (NLSY), measures a child's "receptive vocabulary"—that is, number of words that the child recognizes and understands. The PPVT is sometimes described (and was designed) as an aptitude test. Yet a child's receptive vocabulary is surely affected as much by the quality of that child's educational experiences as by innate aptitude, particularly given evidence of substantial variation across socioeconomic groups in the number of words to which young children are regularly exposed (Hart and Risley 1995). In our view, all cognitive scores should be seen as measures of what a test-taker can accomplish on the day of the test, which is influenced by a combination of the subject's innate ability, the educational and non-educational inputs received in the past, and other factors extraneous to the testing process like testing conditions, health, and mood.

Two related distinctions, central to psychometrics but largely ignored by economists, concern the domain covered and the malleability of the trait being measured. It is common to have separate tests for each core academic subject, including math and

language arts. Within these broad subjects, many assessments have separate questions aimed at different sub-domains, like grammar vs. reading comprehension, or computation vs. geometric reasoning. Scores are sometimes reported for each sub-domain. Given a choice of domain, tests also differ in what is known as "instructional sensitivity." For example, a history test that focuses on facts which might have been covered in class is likely to be very sensitive to the quality and nature of the instruction that the student has received. By contrast, a test of historical reasoning, divorced from specific dates, names, and places, may better measure the student's accumulated skills across several academic subjects but be less sensitive to the specific curriculum or teaching methods of the most recent class. A related idea is that some tests may be more affected by the student's familiarity with the test form and scoring method — for example, students taking multiple choice tests must decide whether and how to guess at an item when the right answer is unknown. In many cases, it may be easier to improve scores by teaching test-taking strategies than by teaching the underlying material. Barlevy and Neal (2012) argue that avoiding this outcome should be a central consideration in the design of testing systems to be used for teacher accountability.

Scaling

Test scores are reported on different and arbitrary scales. The National Assessment of Educational Progress (NAEP), which bills itself as "The Nation's Report Card," reports scale scores, ranging from roughly 100 to 400 with standard deviations around 30, as well as discrete proficiency categories (basic, proficient, and advanced). The verbal and math sections of the SAT college entrance exam are scaled to have

approximately normal distributions with means around 500, standard deviations around 100, minimum scores of 200, and maximum scores of 800. The SAT's competitor, the ACT, uses integers between 1 and 36 for each of four subjects, with means around 21 and standard deviations around 6. These scales are arbitrary in their location (mean), range, and distribution. That is, there is no reason the College Board could not assign the lowest performing student on the SAT a score of 100, or have the highest score be 1000, or set the standard deviation to be 50 or 150 instead of 100, or even adopt a scale that makes scaled scores approximately uniformly distributed.

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. This assumption is implicit in any analysis based on score averages. However, 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 ordinal, not cardinal. This fact has important implications for virtually all empirical analyses of test scores.

Bond and Lang (2013) illustrate the importance of arbitrary scaling decisions in the calculation of a widely cited statistic in education research and policy: the black-white test score gap. Consider a test of three items, each testing a different skill, with the skills ranked cumulatively: A student must master skill 1 before mastering skill 2, and skill 2 before skill 3. Students can answer zero, one, two or all three test items correctly.

Suppose we have a sample of two black students who correctly answer 0 and 2 items, respectively, and two white students who answer 1 and 2 items correctly. The count of correct items is known as the "raw score." In this example, 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.

Now suppose that the three skills are the ability to recite the alphabet, to recognize letters, and to read fluently. In this case, one might consider the incremental knowledge represented by advancing from skill one (reciting the alphabet) to skill two (recognizing letters) to be smaller than the steps from zero (no measured pre-literacy) to one (reciting the alphabet) or from two (recognizing letters) to three (reading fluently). In this example, the difference between the two groups is driven by the black student who scored 0 and the white student who scored 1. If we assume the difference between these students' achievement is much larger than that between the two white students (who also differ in one skill), the black-white gap in average achievement approaches 1 full point. By contrast, if we assume the difference in knowledge between zero and one skill is arbitrarily small relative to that between one and two skills, the black-white test score gap approaches zero. More elaborate examples, where the distribution of one group does not stochastically dominate the other, could even produce reversals of the sign of the gap as the weight put on different skills varies.

This problem worsens if one considers changes over time. Assume that over the school year each student progresses one skill level, so the black students correctly answer 1 and 3 items correctly, and the white students answer 2 and 3 items correctly. The raw gap remains unchanged at 0.5 points. But if we assign more weight to the first skill

(reciting the alphabet) than the second (recognizing letters), we would conclude that the black-white gap had shrunk; if we reverse these weights, we would conclude it had grown. Empirically, estimates of the black-white gap in achievement growth across grades turn out to be extremely sensitive to transformations of the test score, in a way that varies across test and grade level. Depending on the transformation and assessment used, Bond and Lang (2013) find that the change in the black-white test score gap between kindergarten and third grade can be as small as zero or as large as 0.6 standard deviations.

As another example, consider value-added estimates of how teachers affect the achievement of students. Setting aside questions about the causal interpretation of these estimates (Rothstein 2010, 2015; Chetty et al. 2014a,b), any comparison of value-added across teachers whose students start with 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 10 points to the impact of a peer who raises the scores of higher-scoring students 15 points.[2]

The ordinality of test scores thus poses a serious challenge for those working with test score data, and has inspired several types of responses. A first approach, favored by many psychometricians and education researchers, is to develop a parametric model that defines an interval scale for the achievement parameter, and then treat the resulting scores as interval. (Some scholars interpret "item response theory" models, discussed below, in this way.) However, just as with similar uses of parametric utility functions, it is not clear

[2] Indeed, the equation of teachers' causal effects on their students with their effectiveness relies on a much stronger assumption about the test score production process: One needs to assume that a teacher would have the same impact, in scale score points, regardless of the students' initial achievement. Even with an interval scale, this assumption of homogeneity of treatment effects may not hold.

how one might evaluate the claim that a proposed scale of knowledge generated has an interval property.

A second approach, advocated by Bond and Lang (2013), is to accept the ordinality of test scores, limiting conclusions to those that are robust to arbitrary monotonic transformations of the scores. This approach drastically limits the statements that can be made. When scores are treated as ordinal, group achievement is only partially ordered; one group's achievement can only be said to exceed another's if the former's scale score distribution stochastically dominates the latter's. A related approach focuses on students' percentile scores. Reardon (2008) calculates the probability that a randomly chosen black student will have a test score higher than a randomly chosen white student. Ho (2009) and Ho and Haertel (2006) describe how this information can be converted to a standardized metric-free gap measure. These measures permit complete orderings and are invariant to test-makers' scaling decisions, but are nevertheless non-interval; they amount to re-scaling the original test, but do not avoid concerns about assigning importance weights to achievement gains at different points in the distribution.

Some value-added models—known as the "Colorado Growth Model" or the "student growth percentile model"—also rely on percentile scores to sidestep some scaling issues. In these models, each student is assigned a "growth percentile" corresponding to the student's percentile in the distribution of test scores among the sample of students who had the same test score in the prior year. A teacher's value-added is computed as the median growth percentile of students in that teacher's class. Again, this measure is insensitive to the particular test score scale chosen, but nevertheless provides a complete ordering. Barlevy and Neal (2012) propose building teacher

accountability and compensation systems around measures closely related to student growth percentiles. Interpretation of this ordering as reflecting teacher effectiveness depends on interval-like assumptions, however, as 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.

While a focus on the ordinal nature of test scores is clearly more defensible from a psychometric perspective, it does limit the questions that can be answered in research and policy evaluation. An approach that has received recent attention 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; Nielsen 2015b). For example, an attainment-scaled test score would be the average eventual educational attainment — looking ahead in time — of all students with a particular test score. Bond and Lang (2015) use this approach to measure the black-white gap at various grades. The attainment-scaled reading gap is roughly constant from Kindergarten through grade seven at around 0.7 years of predicted educational attainment, while the math gap is close to a full year.

This forward-linking approach yields an interpretable scale that is plausibly interval, but it also raises questions. First, how should one choose the specific outcome to which the test scores are linked? There is no assurance that the scale defined by educational attainment will correspond to that defined by another outcome (such as earnings), nor that either corresponds to a hypothetical scale representing units of knowledge at the time of testing. For example, it might require more inputs to move a student from 9 to 10 years of education than from 11 to 12 years or 15 to 16 years.

Second, scores on a forward-linked scale depend on both the inputs that the tested students received prior to the test and the inputs that earlier students received after the test. For example, the existence of an effective intervention program for low-scoring adolescents will raise the average educational attainment of children who scored poorly on kindergarten tests, and thus compress the left tail of forward-linked kindergarten scores relative to what would be seen from the same kindergarten test responses in a setting without such an adolescent intervention. 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 forward-linked scores may seem intuitive, they can sometimes produce odd results. For example, the black-white gap in attainment-scaled achievement at every grade is likely to be larger than the actual black-white gap in educational attainment, as black students tend to wind up with higher attainment than do white students with the same test scores. Overall, we regard this forward-linking approach as promising but underdeveloped, and not yet ready for broad application.

Finally, it might be possible to assume that raw test scores 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 50 times as large. The challenge then is to parameterize and define this notion in some defensible way. Nielsen (2015a) provides a first step in this direction. His empirical results, like those of Bond and Lang (2013), suggest that cross-sectional achievement gap estimates (for example, for black/white and high-/low-income

gaps) are robust to scale misspecification, but that changes in achievement gaps over time are considerably more sensitive to the choice of scale.

Standardized Scores

When analyzing tests that use well-known scales, such as the SAT, researchers often use unadjusted scale scores. When the scale is not familiar, economists frequently convert (or "standardize") scores to a known scale. There are three common methods: Z-scores are the difference between the examinee's scale score and the mean scale score, divided by the scale score standard deviation; percentile scores are the examinee's rank in the distribution; and normal curve equivalents (NCEs) are obtained by applying the standard normal inverse distribution function to the percentile score. These ad hoc transformations aim to be comparable across tests and samples, but they yield scales that are no more or less correct than the raw or scale scores.

Even when researchers are willing to set aside concerns about non-interval scales, there are several practical challenges to using these transformations. The challenges derive from the fact that each transformation is defined relative to some norming population, which in practice can be small and non-representative. Comparability across assessments depends on the use of norming populations with identical interval-scaled ability distributions, which is difficult to assess unless the two populations are given the same test. Consider, for example, a comparison between two states that administered different exams. Z-scores constructed from samples from the two states are comparable only if the mean and standard deviation of latent achievement, if measured on the same scale, would be identical in the two states; comparison of percentile or normal curve

equivalent scores requires even stronger assumptions about latent achievement distributions. The same problem arises when comparing across ages or cohorts.

Cascio and Staiger (2012) reconsider a 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. They ask whether this could be attributable to the standardization process, rather than an indication that achievement becomes less malleable as children age. Scores are typically standardized separately by age. Differences in the effects of interventions carried out upon students of different ages might therefore reflect either differences in the interventions' true effects or differences in the distribution of scores across students. If the standard deviation of achievement increases with age, a plausible hypothesis as older students have been exposed to more out-of-school influences whose effects may accumulate, this could explain the observed pattern of declining coefficients with age. Cascio and Staiger 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 transitory component that combines what they refer to as "short-term knowledge" with pure measurement error on the test. Based on this model, they conclude that while the variance of latent achievement does increase with age, this cannot fully explain the age pattern of estimated treatment effects.

Even when interval-scaled ability is similarly distributed across groups, measured ability may not be. The age-specific standard deviation combines true variability of ability among children with measurement error in the test. The measurement error component may vary with age even if true ability does not. Dividing by age-specific

standard deviations of measured scores will tend to make between-group differences (like, the black-white gap) in z-scores larger at ages where test measurement error is smaller.

Practical Guidance on Scaling

Both those who consume and those who carry out research routinely use test scores in a way that assumes they have interval properties, although this assumption has no compelling justification. Should one only use the ordinal information contained in test scores, and forgo making any statements about the magnitude of effects? For example, a percentile-percentile plot comparing treatment and control groups would allow the researcher to fully characterize how the two distributions compare without relying on a particular scale, though in many cases the groups will not be ordered without scaling restrictions. While we understand this inclination, we are inclined toward the approach outlined by Nielsen (2015a), who seeks to narrow the class of scale transformations that are considered reasonable. At a minimum, we recommend that researchers make greater effort to test the robustness of their results to changes in the test score scale. For example, researchers might test their sensitivity to modest scale transformations such as the log or exponential of the reported scale score.

The common practice of standardizing reported scores also raises concerns. Secondary researchers should standardize based on the broadest possible population, even if their study focuses on a subpopulation. Comparisons of standardized effect sizes across studies should account for differences in the norming populations. Moreover, in most cases the true (net of measurement error) standard deviation should be used for

standardization; in cases where this cannot be computed from measured scores, sometimes it can be backed out from information in the assessment's technical documentation, such as estimates of the test-retest reliability.

Measurement

Scaling involves the conversion of some initial ability measure into scores with a desired distribution. In this section, we discuss how test-makers obtain those initial ability estimates. The simplest estimate of ability is the fraction of items an individual answers correctly, often referred to as the "raw" score. But this approach has several limitations. First, a student's performance on a test, typically with relatively few items, measures student ability with error. Second, raw scores obtained from different tests or even from different test forms are not comparable. This is a particular issue for "adaptive" testing, where the student's performance on early items determines the difficulty of the items presented later. Third, even within the same form, items of moderate difficulty provide more information about a student's proficiency than do items that are very easy or very hard for that student; holding constant the difficulty of questions, some items may be better or worse at discriminating between more and less able individuals. For example, a test item about baseball statistics may measure knowledge of the sport better than it does statistical proficiency. These considerations motivate use of more complex performance measures, typically based on what is known as "item response theory."

Item Response Theory

An Item Response Theory (IRT) model specifies the probability that a student will answer each test item correctly as a function of a latent parameter representing the student's ability and of parameters relating to the item (van der Linden and Hambleton 1997). In one of the simplest specifications, the probability of a correct answer is a logit function of the difference between the student's ability and the item's difficulty. This is known as the "1 parameter logistic" (1PL) model. The implicit assumption here is straightforward: higher ability students are more likely to answer each item correctly than are lower ability students; all students are more likely to correctly answer simple than difficult items; and both relationships follow a simple, parametric form. (Some psychometricians argue that the implicit scale assigned to student ability by this model should be treated as interval, and refer to it as the "Rasch Model.")

Most item response theory models are more complex. The most common is "three parameter logistic" (3PL) model, which adds two item parameters to the 1PL. One parameter represents a test item's "discrimination" between high- and low-ability students. The more discriminating an item, the steeper the relationship between the student's ability and the probability of a correct answer, and the less overlap there is in ability between those who answer it correctly and those who do not. The second is "guessability" – the probability that even a very low ability student will guess the correct answer. There are also item response theory models for essay questions or multiple-correct-answer questions that are scored in ways other than simply right or wrong.[3]

[3] For a more complete discussion of item response theory models, see van der Linden and Hambleton (1997). Embretson and Reise (2000) provide a readable introduction to the field for non-psychometricians.

Measuring Student Ability in Test Scoring

After a particular item response theory specification—say, the three-parameter logistic version— is chosen, the next steps are to estimate the test item parameters (e.g., difficulty, discrimination, and guessability), and then to use a student's particular combination of right and wrong answers, along with the item parameters, to generate a measure of the student's latent ability. The item parameters are well-identified as the number of tested students gets large, and can typically be estimated with relatively little error. But the typical test has relatively few items, so that the ability of an individual student is not precisely identified.[4] Modern assessment systems vary in the way they handle this.

Some testing systems, including most state tests used for accountability purposes, treat student ability as a fixed effect to be estimated directly via maximum likelihood methods or some variant thereof, applied to the sequence of right and wrong answers. The resulting estimate is (approximately) unbiased in most cases, but can be very noisy. Moreover, when a student gets all questions incorrect or all correct, a maximum likelihood estimate does not exist. A former state-mandated test in Michigan (the Michigan Educational Assessment Program) assigned students who answered all items correctly a score 10 percent higher than what was otherwise possible, and conversely assigned students who answered all items incorrectly a score 10 percent lower than what was otherwise possible. Other tests simply set minimum and maximum scores, and assign students with perfect scores to the endpoints.

[4] The problem is similar to that which arises in many panel data models in econometrics, with the individual effect the object of interest rather than a nuisance parameter.

Other testing systems estimate student ability using random effects models, which generate posterior distributions for each student's ability (including for those who answer all items correctly or all incorrectly). To assign a single score to a student, some tests report the mean or mode of this posterior distribution. Posterior mean scores can be seen as Empirical Bayes estimates of students' latent ability (Morris 1983), which "shrink" the individual's own score (roughly, the maximum likelihood estimate) toward the population mean in proportion to the noisiness of the maximum likelihood score. In item response theory models, ability is estimated most precisely for individuals near the middle of the measured ability distribution. This is because test items are most "discriminating," in the sense that a right or wrong answer provides the most information about the student's ability, when the probability of a correct answer is close to 50 percent.[5] For this reason, the reported ability measure in this framework will be shrunk more for students who score extremely high or low on the exam.

Importantly, neither posterior means nor posterior modes are unbiased *estimates* of student ability. Recall, an unbiased estimate is one in which the estimation error (i.e. the difference between the estimate and the individual's true ability) is zero in expectation and is uncorrelated with the true ability. As noted above, a student's posterior mean is "shrunk" toward the population mean. So, we would expect that the individual's posterior mean is on average smaller (in absolute value) than his or her true ability – it is a biased estimate.

[5] 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 $p(1-p)/N$, and this is highest when p is close to 0.5. Intuitively, logistic item response theory models stretch out the tails of the ability scale relative to raw scores, even as test performance provides relatively little information to discriminate amongst students who do very well or very poorly on the test.

On the other hand, posterior means are unbiased *predictors* of true latent ability. In other words, the prediction error (i.e., the difference between a student's true latent ability and that student's posterior mean score) is mean zero in expectation, and is uncorrelated with the posterior mean score. This difference stems from the fact that when predicting how an individual will do in another context, it is optimal to adjust your prediction to account for the measurement error inherent in the student ability measure generated from a prior assessment. This insight has important consequences for secondary analysis of the scores, which we discuss below.

Most of the longitudinal databases created and distributed by National Center for Education Statistics, including the Early Childhood Longitudinal Study, the National Educational Longitudinal Study of 1988 (NELS:88), the Educational Longitudinal Study (ELS), and the High School Longitudinal Study (HSLS), report scores constructed from posterior means. The Armed Services Vocational Aptitude Battery (ASVAB) scores reported in the 1997 wave of the National Longitudinal Study of Youth (NLSY97) are posterior modes.

Several major assessments, including the National Assessment of Educational Progress, attempt to provide more information about the posterior distribution than a single mean or mode by reporting several "plausible values," which are random draws from the examinee's posterior distribution. Plausible values are closely related to multiple imputation for missing data, and derive from Rubin's (1987; 1996) work on the topic. For excellent summaries of plausible values, including guidance on how to properly use them in secondary analyses, useful starting points are von Davier et al. (2009) and Carstens and Hastedt (2010).

Plausible values are neither unbiased estimators (like maximum likelihood estiamtes) nor unbiased predictors (like posterior means) of individual ability. Their primary benefit is that the variance of plausible values across students equals (in a large sample) the variance of latent ability, which allows one to calculate population variances. In contrast, the variance of an unbiased estimator (like maximum likelihood) will overstate the population variance while the variance of an unbiased predictor (like posterior means) will understate it. On the other hand, as we discuss below, while maximum likelihood and posterior mean ability estimates can each support some secondary analyses without further adjustment, there is essentially no multivariate secondary analysis that would be of interest to economists for which plausible values will yield unbiased estimates.

Incorporating Conditioning Variables into the Generation of Latent Student Ability Measures

To minimize examinee burden, tests are often kept short. Tests with relatively few questions will not provide precise (posterior) estimates of individual ability. To increase precision, some assessments — including the premier US and international assessment systems, the NAEP and the Program on International Student Assessment (PISA), respectively — use priors that vary with student background characteristics. In this approach, a "conditioning model" relates performance on the exam to students' background characteristics (for example, race, gender, family income), and then the prior used for computation of each student's posterior ability distribution is centered at the predicted values from this conditioning model.

As with random effects approaches described above, the posterior distribution from a conditioning model can be summarized by its mean or by several plausible values. The posterior mean still can be viewed as an Empirical Bayes or shrinkage estimator, but instead of being shrunk toward the unconditional mean, a student's performance is shrunk toward the predicted performance of students with similar background characteristics.

While the conditioning approach permits more precise estimates of students' ability (that is, the posterior distributions are tighter), it means that a student's measured score depends on personal background characteristics, even conditional on that student's test responses. Suppose, for example, that race is one of the background variables (as it is in National Assessment of Educational Progress tests), and that on average black students perform less well on the assessment than white students. Now consider two students, one black and one white but otherwise identical in their background characteristics and in their test item responses. Our two students' performance, initially identical, 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 does not bias the average black-white test score gap. The average score of all black students remains the same because the scores of high-performing black students are pushed down just as the scores of low-performing black students are pushed up, and the same for white students (with each pushed toward a group-specific mean). However, individual scores are affected. Scores generated in this way are at odds with the expectations of many data users, and as we discuss below can create important biases in more complex secondary analyses.

Recent administrations of the National Assessment of Educational Progress use hundreds of student and school characteristics in the conditioning model, including student demographics (like race, gender and age), family background characteristics (like parental employment and parental education), school characteristics (including racial composition of the school and whether a school location in in an urban location), student self-reports of study habits and school performance (including overall grades, expected educational attainment, time spent on homework), and teacher reports of aspects of the curriculum and of school policies. The model contains few variables that are likely to be of interest for policy evaluations, however. For example, it 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. Moreover, none of these policy variables are likely to be well-proxied by the student-level characteristics that are included. As we discuss below, this may mean that program evaluations using NAEP scores as outcomes will understate programs' true effects.

Secondary Analysis with Latent Ability Measures

In this section, we discuss how the scaling and measurement issues described above can influence secondary analyses using test scores. For simplicity, we focus on ordinary least squares regressions and refer to the regression of interest as the "research model," distinguishing this from the "measurement model" (that is, the item response theory specification) and the "conditioning model" sometimes used to construct the test scores. For example, if one is interested in estimating the poverty achievement gap, the

research model might be a regression of student ability on a binary measure of being above or below the poverty line. To focus specifically on issues arising from scaling and measurement, we ignore both sampling variability (essentially assuming that the number of examinees is large) and omitted variable bias (essentially assuming that linear projections are of interest, perhaps because the research design supports their causal interpretation).

In many cases, simple estimation of the research model using the test performance measure provided by a test-maker will lead to biased estimates of the relationships of interest. It is important for secondary researchers and consumers of this research to be aware of these biases. Their existence and magnitude depend on the type of ability measure used — that is, whether it is a fixed effects approach to student ability based on a direct maximum likelihood estimate, a posterior mean, or a plausible value, and in the latter cases whether the prior distribution is unconditional or conditional on background characteristics — and also on whether the ability measure is a dependent or independent variable in the research model. 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, often but not always item-level test data that can be hard to acquire.

While the sign of the bias arising in various scenarios is clear, the magnitude of the bias is not. We present illustrative evidence from two studies that assess the magnitude of biases that arise, one regarding racial and ethnic test score gaps and the other examining a "Mincerian" wage regression—that is, a regression that uses schooling

and experience as explanatory variables--that also includes measures of ability derived from test scores as an explanatory variable.

Ability as the Dependent Variable

Measures of student ability based on test score data are common outcomes in both descriptive analyses (for example, of disparities across demographic groups) and evaluations of the causal effects of education programs or policies. A lesson of basic statistics is that classical measurement error in a regression dependent variable will not lead to biased coefficient estimates, though it may reduce the precision of such estimates. When the available test score is of the fixed effects (maximum likelihood) type, this result is likely to apply. But none of the random effects approaches discussed above yield test scores that can be approximated as true ability plus classical measurement error, and regressions that use them as dependent variables are likely to be biased. Consider, for example, estimation of the test score gap between poor and non-poor students. If poor children have lower ability on average, then the poor/non-poor gap in posterior mean scores (without conditioning) will understate the poverty achievement gap. The same is true when the available scores are plausible values, which are merely the sum of the posterior mean and a random component uncorrelated with student ability.

How potentially important is this bias? To measure this, we need to compare the biased estimates to unbiased results from the same test. Few databases of test scores report both random effects and fixed effects ability estimates. Some testing systems, however, report individual item responses. With these data, it is possible to obtain unbiased estimates by estimating a system of equations combining the item response

theory measurement model together with the research model. This system specifies the likelihood for the observed item responses in terms of the item parameters and the research model coefficients, in essence using the research model covariates as the conditioning set. This approach, known as Marginal Maximum Likelihood (MML), is described in seminal articles by Mislevy (Mislevy 1991; Mislevy et al. 1992).[6]

Briggs (2008) assesses the extent of bias in estimates of racial and ethnic gaps in student achievement that rely on posterior mean scores without conditioning variables. He uses a sample of 10th graders in 1999 who were administered the Partnership for the Assessment of Standards-based Science (PASS) test. Table 1 reproduces his estimates. Column 1 shows gaps in scaled posterior mean scores. These indicate that the black-white achievement gap is -0.61 scale points. Column 2 shows unbiased Marginal Maximum Likelihood (MML) estimates. These indicate a black-white gap of -0.77 scale points in the same sample. 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. Elsewhere, Briggs shows that the biases are even larger when considering sub-domains within the larger test.

[6] Implementing the model requires that the researcher invest some time in coding and in computational techniques (like Markov Chain Monte Carlo). The National Center for Education Statistics once contracted with the American Institutes of Research to develop software intended to estimate such models (at http://am.air.org/contact2.asp), although this software is now dated.

Another application where this issue has arisen is in the examination of teacher value-added, which is often computed via Empirical Bayes procedures. Chetty et al. (2014a, Appendix Table 2) assess inequities in access to good teachers by regressing teacher value-added on observable student characteristics. They estimate that the value-added scores are shrunken by 36 percent, on average, and attempt to undo this by multiplying their estimated coefficients by 1.56 = 1/1-0.36. But this is only an approximation. Because the Empirical Bayes estimates are shrunken differently for each student, the bias need not be uniform.

The above discussion applies to random effects estimates of ability without conditioning variables. When a conditioning model is used, the potential biases become more complicated. As described above, the inclusion of conditioning variables can be thought of as shrinking a student's individual performance toward the group-specific mean for those sharing the characteristics of that student. Thus, only the portion of achievement that is not predicted by the conditioning variables is shrunken. An implication is that the coefficients in the research model are unbiased if all of the explanatory variables in the research model were also included in the conditioning model (Mislevy 1991). However, this is unlikely to be the case in many applications. Recall that the conditioning model used in National Assessment of Educational Progress includes many student background characteristics but few, if any, variables that relate to education policies or programs. So if one regressed test scores on student background characteristics and, say, an indicator for whether the school had a high-stakes teacher evaluation system, the coefficient on the teacher evaluation system would be likely to be attenuated, and the student background coefficients might also be biased if the

background and policy measures were correlated. Mislevy (1991) reanalyzes data from the 1984 NAEP Long-Term Trend reading assessment. He finds that biases in coefficients on variables not included in the conditioning model can be substantial.

But modern assessment systems typically include hundreds of variables in the conditioning model, much more than were used in the 1984 National Assessment of Educational Progress. It is not clear how important this type of bias is in today's NAEP. We have investigated this as it applies to two specific examples: an evaluation of the federal school accountability policy No Child Left Behind (Dee and Jacob 2011) and an assessment of the effects of school finance reform on inequalities in spending and achievement across districts (Lafortune et al. 2016), both of which rely on difference-in-differences regression using a state-by-year panel. We found that cross-sectional regressions of NAEP performance on either the state's accountability rule or the district's funding were insensitive to the use of plausible values versus a Marginal Maximum Likelihood (MML) approach. However, other studies that examine different policies or programs may show greater bias. We view this as an important subject for future research.

Ability as an Independent Variable

Now consider a research model in which ability is an independent variable: for example, a regression of wages on education, family background, and an ability measure (for example, Neal and Johnson 1996). Again, economists are generally familiar with the idea that classical measurement error in an explanatory variable leads to an attenuated coefficient on that variable — in this case, the ability measure — and to biases of

predictable sign and magnitude in other coefficients. Once again, however, this result applies only to test scores generated by a fixed-effects method. By contrast, when test scores are posterior means or plausible values, measurement error in these scores is correlated (generally negatively) with the student's true ability. Intuitively, "shrinkage" estimators pull an examinee's reported score more toward the mean the further is that person's true score from the mean. Hence, classical measurement error results do not apply. In this setting, ordinary least squares coefficients are unbiased only in restrictive circumstances: for example when the test score is a posterior mean and the conditioning model includes the covariates from the research model but no other variables that are correlated with the research model outcome. Unlike in the dependent variable case, likely biases are quite different for posterior means than for plausible values, though as before they depend importantly on the presence and form of the conditioning model.

Schofield et al. (2015) model the likelihood of the outcome variable jointly with that for item responses. The resulting estimator, the "Mixed Effects Structural Equations" (MESE) model, is similar in spirit to the Marginal Maximum Likelihood (MML) approach discussed above, and permits unbiased estimation. As with MML, this requires both access to item responses and bespoke programming and computational methods.[7]

Junker et al. (2012) use this approach to assess the bias in a simple wage regression using data from the National Adult Literacy Survey, a nationally representative sample of US adults in 1992 that contains information on cognitive ability

[7] Another approach, not pursued in the literature to our knowledge, would be to instrument for a noisy measure of ability with a second, independent, measure, if available. For example, the National Assessment of Educational Progress 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.

along with survey information on a variety of demographic and socioeconomic 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. Their research model specifies log weekly wages as a linear function of race, a quartic in potential experience, indicators for urban status and census region, and the literacy test score. Table 2 reproduces their results for their sample of 3,267 men. Column 1 shows that the racial gap in wages is 36.6 log points (30.6 percent) without ability controls. Column 2 adds a maximum likelihood estimate of individual literacy, generated from a standard item response theory model. The implied black-white wage gap in this model drops dramatically to 14.4 log points (13.4 percent). However, recall from above that the literacy coefficient is attenuated due to classical measurement error in the maximum likelihood score, implying that the racial gap is overstated here. Column 3 presents unbiased Mixed Effects Structural Equations (MESE) estimates. As expected, the literacy coefficient increases and the implied black-white wage gap drops to 9.4 log points (9 percent). This finding suggests that latent ability accounts for 74 percent of the unconditional black-white log wage gap (= 1 - (-0.094/ -0.366)) when properly controlled, but that a naive estimator would indicate that it accounts for only 61 percent of the gap.

As it happens, the National Adult Literacy Survey data report ability as a set of plausible values, based on a conditioning model that includes 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 measures highly related to wages 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. Schofield et al. (2015) demonstrate that this sort of ability measure typically will result in bias. Indeed, the race coefficient when controlling for the plausible value scores (column 4) is -0.121, overstating the unbiased estimate by roughly 33 percent.

Again, this example makes clear the importance of sometimes obscure measurement choices in the construction of test scores to the substantive conclusions from secondary analysis of these scores. Regressions with test scores as dependent variables are plausibly unbiased when the score is constructed as a fixed effects estimate or as a random effects estimate with a sufficiently large conditioning set, but in nearly all other cases bias is likely. The most likely result is that the coefficients on key policy variables (which are unlikely to be included in conditioning models) will be attenuated, while those on demographic covariates will be overstated. When the test score is an independent variable, in the most common case using plausible values and conditioning on a wide range of predictors of individual ability (but not the dependent variable itself), we are aware of no general results on the sign or magnitude of bias. Information provided by the assessment — namely, the reliability of the ability measure — can in some cases be used in to generate consistent estimates. Or even better, the reported scores can be discarded in favor of analyses that draw directly on examinees' item responses. However, it remains unusual for analytic samples from test score data to include item-level responses; in any event, few secondary analysts are likely to be willing to invest in the appropriate analysis of these responses, which remains tedious.

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 misuse test scores in a way that can lead to serious biases. These biases have not been widely recognized in the literature to date, and may be important to our understanding of key issues in education and labor economics. In this concluding section, we discuss their implications for several of the running examples discussed throughout this article.

The black-white test score gap is a commonly cited statistic, used by educators and policymakers not only to judge specific schools or districts but also to evaluate the effectiveness of reform efforts. Recent studies using the nationally representative Early Childhood Longitudinal Survey (ECLS) have received considerable attention (for example, Fryer and Levitt, 2004, 2006, 2013). As Bond and Lang (2013) illustrate, this statistic is quite sensitive to arbitrary decisions about how to scale test scores, and changes in the gap are particularly unstable depending on such choices. A less recognized, but perhaps as important, concern is that the ECLS test scores are posterior means generated without any conditioning variables—that is, individual ability measures in ECLS are shrunk toward the population mean. This almost certainly implies that the black-white scale score gap in ECLS is attenuated in the cross section, although we are not aware of any research that seeks to assess the magnitude of this bias.

Value-added measures are becoming increasingly common in education, health care and other fields. Indeed, value-added measures of teacher effectiveness are currently used to evaluate teachers in many states, and value-added indicators of quality and cost effectiveness are used to reward hospitals as part of Medicare reforms in the recent Affordable Care Act. Choices about how to scale the outcome measure can have substantial impacts on the resulting statistic, and possibly important policy implications depending on exactly how such measures are used. As discussed above, we recommend that researchers assess the sensitivity of value-added measures by comparing the results of models that use scale scores with those that rely only on percentile ranks. We also caution researchers and policymakers to more carefully match the calculation of value-added (particularly the choice between fixed-effects-style estimators and random-effects-style predictors) to the use to which the scores will be put, as mismatches create biases of the forms discussed above.

Regressions that control for some measure of human capital are common in labor economics (for example, Neal and Johnson 1996). While measures of cognitive ability can be powerful controls in many models, estimated coefficients will be biased under typical conditions. If a maximum likelihood-based estimate of cognitive ability based on underlying test scores is used as a predictor, the coefficient on ability will likely be attenuated, and its relationship with other covariates will be under-controlled. In this case, if the test-maker reports the reliability of the test score, standard errors-in-variables results allow unbiased coefficients to be reverse-engineered. If the test score is instead derived from a random effects framework, either a posterior mean or a plausible value, the nature of the bias is much harder to determine as it depends on the other covariates in

the model and the correlation between these covariates and ability. There are no simple fixes, other than to be cautious in interpreting results.

Finally, policy evaluations that use aggregate panel data (at the state-by-year level, for example) on student outcomes may be biased due to inappropriate construction of the underlying test scores. While the few cases that we have explored (specifications like those used in Dee and Jacob 2011 and Lafortune et al. 2016) do not seem to suffer from important biases, there is no guarantee of the same result in other contexts. In such cases, researchers must first make sure that they understand the cognitive ability data they are using well enough to recognize what biases might be relevant. Also, we suggest that researchers test the sensitivity of their results as much as possible. For example, with surveys such as National Assessment of Educational Progress it is possible to obtain item-level data, with which one can either implement one of the more sophisticated approaches, such as the one suggested by Schofield et al. (2015), or a quick-and-dirty check such as testing robustness of results to using the fraction of items correct for each student as an alternative outcome.

The issues that arise in quantitative analysis of cognitive traits are only becoming more salient. The landscape of testing in US schools is changing rapidly, driven by the widespread adoption of the Common Core state standards for K-12 education.[8] In spring 2015, more than half the states introduced new assessments to match the Common Core

[8] The Common Core standards have been developed by a consortium of states, with strong encouragement from the federal government. They articulate in some detail what students should know and be able to do in each grade and subject in elementary and secondary school. A running theme is a reduced emphasis on memorization and rote computation, in favor of more problem-solving and higher-order thinking. Despite considerable controversy, as of August 2015, 42 states and the District of Columbia had adopted the Common Core standards in English/language arts and math.

standards. These assessments all rely on sophisticated item response theory models to generate the exams and to calculate estimates of individual proficiency. One of the two new assessments (Smarter Balance) is computer-adaptive, so that a student who does well on early items is routed to hard items later in the test. This method 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 more sensitive to the underlying item response theory specification and measurement model.

There is some discussion of developing of standardized assessments aimed at college students, too. Moreover, psychometric methods are are spreading beyond 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 item response theory-based measurement models discussed above, typically applied to batteries of very few survey items Schofield (2015). Test score-like measures are also being used in health, as health care reform has encouraged increased emphasis on quantitative measurement. Across all of these domains, secondary researchers will need to account more carefully for scaling and measurement issues.

References

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

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 18038, National Bureau of Economic Research 2012.

___ and Ethan G. Lewis, "Schooling and the Armed Forces Qualifying Test: Evidence from School-Entry Laws," *The Journal of Human Resources*, 2006, 41, 294—318.

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 P. Reise, *Item Response Theory for Psychologists* Multivariate Applications Series, Lawrence Erlbaum Associates, Inc., 2000.

Fryer, Roland G and Steven D Levitt, "Understanding the Black-White Test Score Gap in the First Two Years of School," *Review of Economics and Statistics*, 2004, 86 (2), 447—464.

___ and ___ , "The Black-White Test Score Gap Through Third Grade," *American Law and Economics Review*, 2006, 8 (2), 249—281.

___ and ___ , "Testing for Racial Differences in the Mental Ability of Young Children," *American Economic Review*, 2013, 103 (2), 981—1005.

Hart, Betty and Todd R. Risley, *Meaningful Differences in the Everyday Experience of Young American Children*, Paul H Brookes Publishing, 1995.

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.

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.

Lafortune, Julien, Jesse Rothstein, and Diane Whitmore Schanzenbach, "School Finance Reform and the Distribution of Student Achievement," Working paper 22011, National Bureau of Economic Research 2016.

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.

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

___ , "The Income-Achievement Gap and Adult Outcome Inequality," Finance and Economics Discussion Series Paper 2015-041, 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. Working paper.

___ and Nathan Wozny, "Permanent Income and the Black-White Test Score Gap," *Journal of Human Resources*, 2013, 48 (3), 510—544.

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

___ , "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*, 2015, 9 (4), 2133—2152.

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

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.

Table 1. Biases when using posterior mean test scores as a dependent variable (from Briggs 2008)

	Logit units		Z scores	
	(1)	(2)	(3)	(4)
	Posterior Mean	MML	Posterior Mean	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
N	420	433	420	433
SD of test score	0.7	0.81	1	1

Notes: Estimates reproduced from Briggs (2008), Tables 4 and 6. N = 433. MML = Marginal Maximum Likelihood. Data pertain to performance on a 10th grade science assessment. Columns 1 and 3 report estimates when posterior mean (without conditioning variables) are used as the dependent variable in an OLS regression; Columns 2 and 4 report estimates obtained via the Marginal Maximum Likelihood method discussed in the text. In Columns 1-2, scores use the scale of a logit index (so that the probability of a correct answer equals the logit function applied to the scaled score with an additive adjustment); in Columns 3-4 these are divided by their estimated standard deviation. 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 2. 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)	PV (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. MLE = Maximum Likelihood Estimates; MESE = Mixed Effects Structural Equations; PV = Plausible Values. In Columns 2 and 4, MLE and PV test scores (respectively) are entered as regressors in OLS regressions. Column 3 applies the MESE system-of-equations method. The research model in each column includes controls for a quartic in potential experience as well as indicators for urban status and census region.