Braun and von Davier *Large-scale Assess Educ (2017) 5:17*
DOI 10.1186/s40536-017-0050-x

Large-scale Assessments in Education

RESEARCH Open Access

# The use of test scores from large-scale assessment surveys: psychometric and statistical considerations

Henry Braun[1*] and Matthias von Davier[2*]

*Correspondence: hbraun@bc.edu; mvondavier@nbme.org
[1] Lynch School of Education, Campion Hall, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA
[2] National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA 19104, USA

## Abstract

**Background:** Economists are making increasing use of measures of student achievement obtained through large-scale survey assessments such as NAEP, TIMSS, and PISA. The construction of these measures, employing plausible value (PV) methodology, is quite different from that of the more familiar test scores associated with assessments such as the SAT or ACT. These differences have important implications both for utilization and interpretation. Although much has been written about PVs, it appears that there are still misconceptions about whether and how to employ them in secondary analyses.

**Methods:** We address a range of technical issues, including those raised in a recent article that was written to inform economists using these databases. First, an extensive review of the relevant literature was conducted, with particular attention to key publications that describe the derivation and psychometric characteristics of such achievement measures. Second, a simulation study was carried out to compare the statistical properties of estimates based on the use of PVs with those based on other, commonly used methods.

**Results:** It is shown, through both theoretical analysis and simulation, that under fairly general conditions appropriate use of PV yields approximately unbiased estimates of model parameters in regression analyses of large scale survey data. The superiority of the PV methodology is particularly evident when measures of student achievement are employed as explanatory variables.

**Conclusions:** The PV methodology used to report student test performance in large scale surveys remains the state-of-the-art for secondary analyses of these databases.

**Keywords:** Large-scale assessment, Imputation, Plausible values, Conditioning model, IRT, Unbiasedness

## Background

Recent publications have re-ignited interest in the approach to modeling used to generate achievement measures for large scale assessments such as NAEP and PISA. Even though the foundations and statistical methodology behind these models have been extensively covered for over 3 decades (Mislevy 1984, 1985; Mislevy and Sheehan 1987), there continue to be concerns about their ability to provide appropriate estimates of population statistics such as means and variances (e.g. Goldstein 2004; Cohen and Jiang 1999). In addition, the latent regression methodology used to estimate population characteristics which, in practice, is combined with the method of plausible values (a form of multiple imputation) to produce achievement measures for secondary analyses, continues to be scrutinized as to whether such measures are suitable inputs for econometric modeling (Jacob and Rothstein 2016). Although the latent regression modeling and the associated imputation methodology have been the focus of a large number of publications showing that these methods produce unbiased population estimates (e.g. Mislevy et al. 1992; von Davier 2007; von Davier and Mislevy 2009: Marsman et al. 2016), the recent article by Jacob and Rothstein (2016) [henceforth JR] questions the increasing use by economists of the test scores so generated as credible measures of human capital.

The goal of that article was to address important issues that arise when such measures of student ability are employed in statistical analyses. The article's broad coverage is, in our view, both welcome and somewhat problematic: The issues arising with conventionally designed standardized tests (e.g. end-of-course tests, college admissions tests) are different from those that arise in the analysis of data from large-scale assessment surveys (LSAS) such as the National Assessment of Educational Progress (NAEP), Programme for International Student Assessment (PISA), Trends in International Math and Science Studies (TIMSS), Progress in International Reading Literacy Study (PIRLS), and Programme in the International Assessment of Adult Competencies (PIAAC). Consequently, it is important to clearly distinguish between these two assessment categories.

In this article we address many of the questions and concerns related to the conduct of LSAS and the analysis of data that result from their administration. Although the original impetus was to respond to the JR article, our present goal is broader: To provide a clear but comprehensive description and evaluation of the present state of the technology for LSAS. In particular, drawing on an extensive psychometric and statistical literature, we argue that secondary analysts, following generally accepted procedures, can indeed draw valid and useful results from LSAS databases. Our principal methodological focus is on the use of so-called plausible values that are related to the estimation of an individual's cognitive proficiency (described below).

LSAS draw probability samples from the target population and administer to the sampled individuals one or more cognitive tests and an extensive background questionnaire (BQ). Crucially, LSAS are specifically designed—and only intended—to yield group-level statistics, and are not aimed at reporting results for individuals. However, LSAS do generate and employ individual level imputations that incorporate information from both test performance and the BQ. The imputations are random draws from a conditional distribution that represents an estimate of the individual's proficiency, as well as the uncertainty associated with that estimate. These imputations are called plausible values in the LSAS literature. Plausible values play a critical role in obtaining unbiased estimates of group-level descriptive statistics (averages, percentiles, etc.), as well as of regression coefficients in models of relationships between cognitive skills and background variables such as gender, educational attainment, and immigrant status (e.g. von Davier et al. 2009).

LSAS produce publicly available databases that can be used to conduct a large variety of statistical analyses. There are a number of software tools available through the organizations conducting LSAS that facilitate access to and analyses of these databases. For example, the National Center for Education Statistics (NCES) provides an online analysis tool called the International Data Explorer (available at https://nces.ed.gov/surveys/international/ide/) which allows access to databases from the PISA, TIMSS, PIRLS, and PIAAC.

The focus of LSAS on facilitating the study of (only) group-level statistical associations between test performance and background variables, or between test performance and other outcomes, offers test developers greater flexibility than usual. This has led to test designs that differ from the ones used in more typical settings. The designs used in LSAS have many advantages, among them an ability to incorporate a much larger pool of tasks than would be possible when administering the same test to all test takers, However, they also require the application of more sophisticated measurement models in order to generate the cognitive data that are used in primary (descriptive) and secondary (model-based) analyses (Mislevy 1991; von Davier et al. 2007). Unfortunately, these measurement models and the properties of their output have sometimes caused confusion resulting, on occasion, in inappropriate analyses and faulty interpretations.

JR set out to clear up this confusion and to offer cautions on the use of the cognitive data from LSAS in secondary analyses. As measurement scientists and test designers we applaud efforts to inform economists (and others) about the relevant issues and to encourage best practices. However, we feel that in this instance the effort falls somewhat short of the goal and, in some matters, obscures rather than clarifies the potential these databases offer for secondary analysis. Accordingly, we offer our perspective on these matters.

## Theories and models of measurement

JR correctly emphasize the importance of understanding what it is that the test measures; that is, the underlying construct that delineates the cognitive skills that are the target of inference. To say that the test measures first year high school math is insufficient as there are many ways to conceptualize this construct, leading to very different tests. For example, both PISA and TIMSS assess topics appearing in early secondary mathematics curricula. However, the assessment frameworks (OECD 2013; Mullis and Martin 2009) they have developed are quite different, as are the assessment instruments that are aligned to those frameworks. Consequently, it is not surprising that the relative performance of subpopulation groups and administrative jurisdictions can vary considerably across these two LSAS (e.g. Wu 2010). Related issues regarding careful definition of the underlying constructs and how they affect test development and test use are treated in Braun and Mislevy (2005) and Mislevy and Haertel (2006).

Traditionally, the statistical models employed in the analysis of test data are those associated with classical test theory (CTT) or item response theory (IRT) (Lord and Novick 1968). Briefly, CTT conceives of an observed test score as the sum of a “true score” and a random disturbance. Under reasonable assumptions, CTT leads to definitions and calculation formulas for such familiar quantities as test reliability. It is still an important tool in day-to-day test analysis, especially if a single test form with a simple (linear) test design is used.

By contrast, in its simplest incarnation, IRT begins with the notion of a latent trait that represents the construct and the assumption that each individual has some (unknown) value with respect to that trait. The goal of the test is to estimate as well as possible the unknown value for each individual. For each item in the test it is assumed that there is a "dose–response" curve that describes the probability of obtaining the correct answer (the response) as a function of the values of the latent trait (the dose).

JR describe different types of estimators of proficiency used in test analyses with IRT. However, their statements regarding these estimators are at odds with the literature. Specifically, JR state that "…the typical test has relatively few items…" and directly below that "…student ability [is] …estimated directly via maximum likelihood …[and] resulting estimate is (approximately) unbiased in most cases…" (p. 95) However, the maximum likelihood estimator (MLE) for the latent trait in IRT models is not unbiased (Kiefer and Wolfowitz 1956; Andersen 1972; Haberman 1977). The extent of the bias is directly related to the number of items, so that MLEs for tests with "relatively few items" will exhibit a more pronounced bias, as well as considerable noise. The bias of ML estimates can be examined formally, and even reduced, by the methods proposed by Warm (1989) for IRT and by Firth (1992, 1993) for a more general class of latent variable models. The weighted likelihood estimate (WLE), as the estimator proposed by Warm (1989) came to be known, eliminates the first order bias of the MLE (see also Firth 1993). It turns out that these bias corrections of the MLE are equivalent to Bayes modal estimators using the Jeffreys (1946) prior for a number of commonly used IRT models. This assertion holds for the Rasch model and the 2PL model (Warm 1989), and estimators of this type are available for several polytomous IRT models, including some polytomous Rasch models (von Davier and Rost 1995, von Davier 1996). Recently Magis (2015) verified this equivalency for 'divide by total' and 'difference' type polytomous IRT models.

The use of IRT in the construction of score scales is now considered best practice in many areas of test analysis (van der Linden 2016; Carlson and von Davier 2013; Embretson and Reise 2000). It has gained further support from research that shows IRT to be a special case of a much larger class of latent variable models that is commonly used in applied statistics (Takane and de Leeuw 1987; Moustaki and Knott 2000; Skrondal and Rabe-Hesketh 2004).

In IRT, the dose–response curve is referred to as the item response function and is usually modeled as a logistic function with one, two, or three parameters. A test of $K$ dichotomously scored items administered to N examinees yields a $NxK$ matrix of zeros and ones from which one can estimate the parameters of the $K$ items and the values of the latent trait for the $N$ examinees. The former is referred to as item calibration and the latter as ability estimation. Note that this estimation problem is an order of difficulty greater than that typically encountered in biostatistics where the doses are known.

Estimation can be done by some variant of maximum likelihood or by Bayesian techniques. Maximum likelihood estimation, either in the form of marginal maximum likelihood (MML) or conditional maximum likelihood (CML) estimation of item parameters in IRT, is the method of choice in LSAS (e.g. Adams et al. 2007; von Davier et al. 2007, 2013). Ability estimation is often done either using (weighted or bias-corrected) maximum likelihood or Bayesian approaches.

For use in LSAS and other more complex test designs, IRT models have been extended to allow for items that are not dichotomously scored and for multi-dimensional latent traits. In particular, LSAS test batteries represent the focal skill domains very comprehensively by employing a large number of test items. Consequently, inferences are based on an item pool that would typically require several hours of testing. Each respondent only takes a carefully selected, small subset of the item pool. However, through an application of an appropriate experimental design (balanced incomplete block designs), overall each item in the pool is administered to a random sample of respondents. Although the formulas and estimation techniques are necessarily more complicated for this extended IRT approach, the basic ideas remain the same.

JR correctly point out that there is an essential indeterminacy in IRT estimation: The quality of the fit of the model to the data is the same under monotone transformations of the underlying scale. Secondary analysis results can differ with the choice of scale (Ballou 2009). Consequently, appropriate cautions in interpretation are in order. However, the reporting scales are typically set in the first round of assessment and establish a mean, standard deviation, and range that are based on the score distribution of a well-defined initial set of populations.

For example, the PISA scale, with a mean of 500 and a standard deviation of 100, was set with respect to the group of OECD countries that participated in the first administration. One can call this an arbitrary or a pragmatic approach, but it is certainly not an approach that is uncommon, or not used elsewhere. Temperature scales differ in how they set reference points, and even the metric versus imperial system of measures show by their mere existence and simple exchangeability that neither inches nor centimeters are more rational or more arbitrary than the other.

JR also correctly point out that there is no evidence that these psychometric proficiency scales have the interval scale properties that are implicit in many secondary analyses (e.g. regression modeling). This is indeed a point of concern, as the properties of the scale are an assumption, as is also the case with wages (log-wage is often used), time (and age) measures, as well as in medical measures. (Is a change in heart rate from 80 to 140 of the same medical concern as a change from 140 to 200? Is the change in physical strength or vocabulary between ages 1 and 6 the same as between 24 and 29? Aging 5 years may indeed mean very different things at different initial ages.) There are many measures in the physical as well as behavioral sciences that are represented as real values or integers, employed as such in models and that, at the same time, may not have the desired 'interval' properties (or, better, the same interpretation of scale score differences) on related scales of interest. On this point, a remark by Tukey (1969, p.87) seems apropos:

> *Measuring the right thing on a communicable scale lets us stockpile information about amounts. Such information can be useful, whether or not the chosen scale is an interval scale. Before the second law of thermodynamics—and there were many decades of progress in physics and chemistry before it appeared—the scale of temperature was not, in any nontrivial sense, an interval scale. Yet these decades of progress would have been impossible had physicists and chemists refused either to record temperatures or to calculate with them.*

In point of fact we cannot verify the scale properties of many variables that we use in our analyses, but we can perform a variation of what one may call a scaling sensitivity analyses. As an example, educational attainment is another variable frequently seen in regressions used by labor economists, either measured in years of schooling (Do we start counting at kindergarten, or pre-school?) or in universally defined levels of education described by the International Standard Classification of Education codes (ISCED; UNESCO 2011). While the ISCED codes do not perfectly cross-classify all national education systems, they do yield an ordered set of educational attainment levels that attempts to represent equivalent types of education rather than just the number of years someone remained at school.

Although the choice of a particular numerical scale in an application of IRT is arbitrary, there are mathematical results on monotonicity properties that describe how increasing scale values are associated with increased expected outcomes on task performance and other variables (e.g. Junker and Sijtsma 2000). Moreover, similar to the ISCED levels, most test score scales are accompanied by a categorization of levels; for example, they can be based on typical tasks that are carried out correctly with high probability by individuals in these levels. Other examples are scales that are anchored (i.e. given meaning) by relevant variables linked to the scale. The types of variables that are used for anchoring can be job categories or ISCED levels (What is the average score of test takers whose parents have a high-school degree as their highest degree of education? What is the average score of test takers with one or two parents with masters or PhD level degrees?). Other variables used for anchoring can include, for example, the average scores achieved by developed countries, by developing countries, by student populations defined by school type, by native language, etc. In assessments such as PIAAC, scales can be described further in terms of expected scores on the scale for individuals in different job categories, at different levels of educational attainment, or at different income levels for that matter.

The reporting scales in LSAS are usually anchored by a process called proficiency scaling, which defines contiguous intervals on the scale associated with typical classes of problems or activities that individuals scoring at that level can master on a consistent basis. Finally, repeated use of a particular scale, together with the associated validity data and descriptions of what students at different ability levels can typically do, does achieve a certain interpretive familiarity over time.

JR refer to an example by Bond and Lang (2013) in which three different skills are measured in black and white student groups and the gap in test scores depends on how the skills are weighted. We could not agree more with the statement that the values of derived variables, such as group differences, will depend on how the scale is constructed through the weights assigned to the different component skill subscales. This is similar to what one sees in stock market indices: Different companies from different segments of the economy are included or excluded, and indices will be differentially sensitive to different events that may trigger market reactions.

What underlies the Bond and Lang (2013) example is an issue that occurs whenever one measures a variable that is not directly observable. Whether it is the literacy skills of students, or the health of a market segment, the degree to which a measure is sensitive to differences or changes depends on the choice of indicators (tasks in student assessments and stocks in indices).

To some extent, we cannot escape this scaling (and weighting and selection of indicators) problem in any science, whether it is the use of adjustments in C14 carbon dating of archeological artifacts, or the measurement of cosmic distances based on signals collected by radio telescopes, or stock indices or educational tests as measures of underlying constructs. More importantly, the different indicators will be differentially sensitive to underlying group differences in the construct, as each indicator was selected based on the goal of representing different aspects of the underlying phenomenon. Thus, 'varying gaps' are not an indicator of a deficiency but, rather, a consequence of the different ways reasonable indices (or tests) can be constructed. The phenomenon is often referred to as the reliability/validity dilemma: A test (or index) that maximizes reliability will contain only very similar components and will hence not be sensitive to the differences on outcome variables that are either caused by or cause, or are just correlated with, a much broader range of other measures.

Indeed, one could turn the question around and argue that there is absolutely no reason why the black–white gap should be the same across different indicators of literacy. Different aspects of literacy are by definition distinguishable attributes of a broader construct. Changes in pedagogy or policy may affect one attribute more than another. To that point, the expectation, for example, that the gap should be time-invariant when the measurement instrument changes to account for more complex literacy related activities in higher grade levels seems somewhat counter-intuitive, but would at least need a rather elaborate theoretical explanation why that should be the case.

Coming back to the scaling and scale level issue raised by JR, it is possible—and one could argue even necessary—to exploit this essential indeterminacy of the latent variable scale by utilizing scale transformation and linking methods to make proficiency scales comparable across cycles. This means that once a scale has been set (and anchored by proficiency level descriptors), then it can be used as the reference for future assessments. The methods that help to ensure the comparability of the scales of future assessments typically utilize common blocks of items over time. Measurement invariance models based on factor analysis, and their IRT equivalents, are then used to align the results of the current assessment cycle to the reference scale (Yamamoto and Mazzeo 1992; Bauer and Hussong 2009; Mazzeo and von Davier 2008, 2013). This is done on an ongoing basis for NAEP, as well as for international assessments such as PISA, TIMSS, PIRLS, and PIAAC. Of course the defensibility of these linking procedures depends on the validity of certain invariance assumptions with respect to how the tasks on the test are responded to across different populations and cohorts (Mazzeo et al. 2008, 2013).

In addition, as JR point out (p. 92), it has to be understood that any transformation, whether based on small or large samples, arbitrarily applied to test scores may not yield comparable test scores even if they are numerically transformed onto the same scale. The linking methods and comparability/measurement invariance approaches cited above do not apply such transformations; rather they utilize (and test) invariance assumptions in the form of parameter constraints that lead to linked test scores on the same scale, thereby allowing comparisons across countries and over time.

## Plausible values as a special case of multiple imputations

Plausible values (PVs) are what the literature on missing data calls multiple imputations (Rubin 1987; Little and Rubin 2002). They are drawn from a model that describes the posterior distribution of one or more cognitive skills assessed with the test(s), given the responses to the test(s), as well as observed test taker characteristics. Our intent here is to make more transparent the technology underlying the generation of the cognitive data—in the form of plausible values—and to offer guidelines for use that we believe are consistent with that technology. As do JR, we address two cases: Cognitive data used as a criterion in a regression or used as a predictor in a regression.

As noted above, there is a difference between simpler test designs for individual-level reporting and decision making and LSAS for group- and population-level reporting. To reiterate, LSAS do not provide point estimates of an individual's skills; rather, they provide conditional distributions that represent an estimate of the individual's proficiency together with an estimate of the uncertainty associated with that estimate. Although one may assume that an individual has a true value on the latent trait, any finite test and any amount of additional information on the individual's background cannot provide certainty about that true value: Even a test that has 100 items will not produce an error-free measure of proficiency. Another equally well-constructed test of 100 items will likely yield a slightly different score, even if the same test taker takes the two versions of the test on two consecutive occasions. These 100 items can be like attempts at picking a winning stock 100 times, kicking a ball into a goal (an example used in the piece on plausible values by von Davier et al. 2009), solving 100 math problems or playing 100 chess matches. Replications generate some level of variation in performance, and no amount of information can provide absolute certainty—in particular because the value of a latent trait is not directly observed, but can only be inferred by looking at manifest outcomes.

Conditional distributions of proficiency are utilized to generate plausible values (multiple imputations) that are a representation of our finite knowledge concerning an individual's value on the latent trait scale, given the individual's pattern of correct and incorrect responses, as well as information on her background characteristics. Mislevy (e.g. 1991) has shown that this approach ensures unbiased estimation of group differences for those characteristics that are part of the imputation model.

Groups are usually defined by some combination of factors such as gender, race/ethnicity, location, etc. See also von Davier et al. (2009) for a comparison of the estimates when using—or not using—the background data in the imputation-based approach. Test makers have taken advantage of the flexibility afforded them by LSAS by building very large item pools for administration in order to ensure broad representation of the skill domain. The pool size is driven by the number of facets of the focal construct to be assessed, by the need to have a range of item difficulties, as well as by cost considerations. The first is intended to satisfy the design criterion of construct representation and the second the criterion of reasonably accurate measurement all along the proficiency scale.

A typical item pool can consist of hundreds of items, far too many to administer to any one individual, especially considering that testing large number of students or adults presents a considerable burden on participating schools or households, as well as on survey organizations. In addition, considerable time is required to complete the BQ that elicits information on various domains including demographics, socio-economic-status, education and extracurricular skill-related activities. In adult assessments, the BQ also collects data on work history and income, as well as work and non-work activities that may be related to skill development or labor market success.

The solution is to divide the item pool into a collection of carefully designed, mutually exclusive blocks (as they are called in NAEP) or clusters (as they are called in PISA). Depending on the LSAS, each examinee is administered one or more of these blocks, along with the (common) BQ block. In NAEP, for example, the blocks are organized into booklets, each consisting of two cognitive blocks, according to a balanced incomplete block design. That is, each block appears in the same number of booklets, each time paired with a different block and balanced overall with respect to order. A booklet is randomly assigned to each examinee. For more details, see Mazzeo and von Davier (2013). As a result, employing IRT and using all the data generated by the administration, it is possible to construct a single proficiency scale on which test performance can be represented.

A problem would arise, however, if individuals were assigned a single score on the proficiency scale based on their item responses alone—as would be the case in a typical end-of-course test administration. The problem is that because each individual is exposed to a set of items that constitute a small fraction of the full item pool, the corresponding estimate of proficiency would be associated with a large error variance. Aggregating these imprecise individual scores to the group level typically yields biased estimates of the proficiency distribution of the group. Notably, this would be the case no matter which individual level estimator was used, maximum likelihood estimates, or bias-corrected versions such as the WLE (Warm 1989), or Bayesian estimates such as the expected-a posteriori estimates (EAP). Each would yield a particular type of biased estimate. An illustration of the result of using individual-level estimates is given by (von Davier et al. 2009).

## Explanatory IRT using latent regressions in LSAS

The solution is to introduce the plausible value (PV) machinery based on a model that involves a combination of IRT and a latent regression (Mislevy 1991; Adams et al. 1997; Andersen 2004; von Davier et al. 2007). It is important to recognize that this approach was adopted specifically to produce unbiased estimates of group-level statistics. As JR note, this is an adaptation of Rubin's (1987) missing data imputation model. The technical details can be found in Mislevy (1991). For a recent overview of current developments see von Davier and Sinharay (2013).

The basic idea is to treat the individual's location on the (latent) proficiency scale as missing data. The observed data consists of her responses to the cognitive items and to the BQ questions. The machinery comprises a multi-dimensional IRT component (for item calibration) and a normal theory, latent regression (LR) model that links the estimand to the background factors. The structural part of the LR model is characterized by a matrix of regression coefficients, $\Gamma$, and a variance–covariance matrix, $\sum$. In practice, because the number of variables associated with the BQ factors numbers in the thousands, the original set of predictor variables is replaced by a large, but manageable, number of principal components sufficient to account for at least 90% of the variance.

The estimation process yields maximum likelihood estimates for both $\Gamma$ and $\sum$. In practice, $\sum$ is held fixed at its MLE, denoted by $\sum_{MLE}$, while multiple versions of $\Gamma$ are obtained as independent draws from an estimate of its sampling distribution, denoted by $\mathrm{F_G}$. This is a multivariate normal distribution with mean $\Gamma_{MLE}$ and an estimate of the variance–covariance matrix of $\Gamma_{MLE}$. Thus, to obtain K PVs for an individual, one makes K independent random draws from $\mathrm{F_G}$: $\Gamma_1, \ldots, \Gamma_K$. Denote one such random draw by g. Combining g with the vector of cognitive responses and the set of principal components for the individual yields a trial mean vector. The PV is then generated as a random draw from a multivariate normal distribution with that mean vector and variance–covariance matrix $\sum_{\mathrm{MLE}}$. The process is repeated K times. Details regarding this sampling process can be found in von Davier et al. (2007, 2009, 2013).

## The use of plausible values as dependent variables

In order to obtain estimates of desired quantities in secondary analysis such as group differences or, more generally, the parameters of a linear regression model, calculations are carried out K times, once for each set of PVs and the results averaged. Mislevy (1991) proves that this process yields unbiased estimates of mean proficiencies for groups defined by the factors incorporated in the latent regression model. This result is consistent with the broader literature on estimation with multiple imputations (Little and Rubin 2002). Further, if PVs are used as the criterion in a linear regression, then the corresponding regression coefficient estimates, obtained by combining the K estimates generated by the K sets of PVs, are unbiased (or approximately so because of the use of principal components rather than the original variables)—as long as the latent regression that generated the PVs is 'larger than' the secondary analyst's model (e.g. Mislevy 1991; von Davier et al. 2009; Junker et al. 2012, p. 736).

Standard results for the calculation of variance estimates and corrected degrees of freedom using multiple imputations (Little and Rubin 2002) apply directly to analyses using PVs. Parameter estimates are obtained by averaging the results from the K replications. The variance component that estimates measurement uncertainty is calculated following Little and Rubin (1987, 2002). The equations are also given in (von Davier et al. 2009). Note that the secondary analysis model is typically a subset of the latent regression model used to generate the PVs. However, if variables beyond those in the latent regression are used in a secondary analysis, then biased estimates may result (Mislevy 1991; Meng 1994). On the other hand, since the PVs generating model typically includes as many factors as are available ("kitchen-sink approach": Graham 2012), even these additional variables may be effectively included by proxy, to the extent that they are correlated with the variables incorporated in the latent regression.

Although not central for the argument made below with regard to PVs, it is appropriate to note that JR (p.100) features a table taken from Briggs (2008) that does *not* include estimates based on the PV machinery. Rather, it compares group mean estimates obtained with either ML or EAP. As expected, the EAP estimates differ from those obtained using maximum likelihood in the well-known way: Both sets of results are biased and, in fact, are likely biased in opposite directions.

PVs are neither maximum likelihood estimates nor EAPs, so that any conclusions drawn from that table are not germane to the points made by JR later on. Recall that PVs are random draws from an individual-specific family of posterior distributions based on a comprehensive imputation model that contains both background data and test performance indicators. Therefore, the *K* PVs associated with an individual are (unlike the values compared in the table) not test scores in the usual sense and do not at all correspond to the true score model of CTT. They should not be confused with the EAP estimators, as suggested in JR. Instead, PVs are intermediate values in the calculation of group level statistics such as group means, regression coefficients, or correlations.

JR also offer examples of situations where certain school-level characteristics are of interest but were not included in the conditioning model. In actual practice, this may not be a problem. Such characteristics are either drawn directly from items incorporated in the school questionnaire and are part of the conditioning, or indirectly, through inclusion of a dummy coded school identifier. If particular characteristics that become subsequently available are of interest, then supplementary latent regression models can be run to generate new PVs so as to ensure unbiased estimation. Software for conducting these latent regression model analyses is available upon request from organizations such as ETS (PC Windows version DGROUP, Rogers and Blew 2012) and ACER (Conquest; Adams et al. 1997).

## The use of plausible values as independent variables

Some of the issues that arise when PVs are used in regression models as independent variables are similar to those that arise with fallible predictors. (These are sometimes referred to as "errors in variable" models.) Others are specific to PVs. First, it is well-known (Fuller 2006) that in the presence of fallible predictors the corresponding regression coefficients may be deflated (i.e. biased toward zero). Potentially, the application of method of moments or more recent generalized approaches that take heteroscedasticity into account (Lockwood and McCaffrey 2014) can yield corrected estimates.

The concerns raised with the use of PVs appear to be a particular instance of a general problem treated by Meng (1994). Meng addresses the validity of results based on an analysis incorporating PVs. He defines the concept of congeniality between the model generating the multiple imputations and the model of the secondary analyst. When congeniality holds, the results are valid but when it fails to hold, bias is likely to arise. The specific case of PVs is treated in Junker et al. (2012), Schofield et al. (2015), and Schofield (2015). In effect, they argue that in many situations congeniality fails to hold.

Lack of congeniality can occur in many ways. Suppose, for example, that it is suspected that the outcome of interest, say wages, and the imputed proficiency values are statistically associated in non-linear ways, and so the secondary analyst's model contains non-linear transformation of the outcome variable or one or more of the independent variables; however, the latent regression does not include such terms. In this setting, a custom-made latent regression model, or a mixed effects model of the type suggested by Schofield et al. (2015) might prove of value. On the other hand, if the imputation model is a latent regression model that contains all cognitive response data and an extensive collection of background data in the form of a contrast-coded set of predictors, as is the case in LSAS, then the family of secondary analysis models that are congenial with the imputation model will be very large. In particular, the latent regression model used in LSAS is specified using predictors as ordinal variables in dummy-coded form, so that non-linear relationships between variables such as earnings, years of schooling, or home resources can be captured because they are incorporated as effects at different levels of these variables. Finally, LSAS do differ in the number and types of two-factor interactions among predictors that are formally incorporated into the pool of variables used for the latent regression. NAEP, for example, employs a comprehensive set of two-factor interactions.

JR (p. 102) display a real data example of the use of PVs that is taken from Junker et al. (2012). The data are an extract from the National Adult Literacy Survey (1992). They model log(weekly wages) as a function of race (Black, non-Hispanic White), literacy skill, and other variables. Comparing estimates using the Mixed Effects Structural Equation (MESE) approach and one using PVs, there are small but non-trivial differences in the estimated regression coefficients for the Black–White gap and for cognitive skills. With MESE, cognitive skills account for 74% of the Black–White log(weekly wages) gap. However, using PVs only 61% of the gap is explained. Thus, in this example, the extent to which the race gap in wages is accounted for by skill differences is somewhat smaller with PVs. JR prefer the MESE estimates, because the reduction is greater, and the model for generating the PVs is not compatible with the wage equation model.

Although it is neither explicit in JR nor in Junker et al. (2012), one reason may be that log(weekly wages) was used as the criterion but it is unlikely such a transformation was used in the NALS conditioning model. However, the extent to which a substantial bias occurs cannot be completely determined using real data, as the true ability variable is unobserved, and each estimate that is used in its place relies on certain assumptions and approximations. Also note that Junker et al. (2012) state that "If the form of the secondary analyst's research model is the same as the …institutional conditioning model …its estimate using institutional PVs … will be an unbiased…"

We now present a case in which the imputation model used to generate the plausible values (called the "institutional conditioning model" by Junker et al.) is compatible with the analyst's model, based on simulated data. A disadvantage of real data is that the true effects of the contributing variables are unknown. The advantage of simulation is that the "true" ability is known and is used to generate the data. It is then possible to compare the estimated regression parameters from different strategies for estimating ability with those obtained when true ability is used.

**Table 1 Estimated regression coefficients for a model predicting wages (raw) based on skill variable theta (either true value, or PVs, or WLE, MLE, EAP estimates), gender and education**

| Estimator for θ | Intercept | COG | SEX | EDU |
|---|---|---|---|---|
| True ability | 9.967 | *1.012* | 0.002 | *0.024* |
| PVs | 9.951 | *0.998* | 0.009 | *0.057* |
| EAP | 9.718 | ***0.810*** | 0.009 | ***0.539*** |
| MLE | 9.567 | ***0.548*** | 0.017 | ***0.871*** |
| WLE | 9.697 | ***0.668*** | 0.010 | ***0.592*** |

Simulation setup similar to the one found in von Davier et al. (2009)

The quantities printed in bolditalics and italics are those regression coefficient estimates that are adversely affected by the use of EAP, MLE as well as WLE estimates in the regression. These quantities are biased, while the estimates obtained using PVs are very close to the estimates calculated using the true person parameters

We model wages as a function of cognitive skills (COG in the table), gender (SEX in the table), education level (EDU in the table) with three ordered levels (low = 1, medium = 2, high = 3), and other variables. The generating model had the following characteristics: (i) Wages were highly correlated with COG but not with SEX or EDU; (ii) COG had a positive correlation with EDU but not with SEX. The PVs were generated using all item response data in a balanced incomplete block design given in (von Davier et al. 2009), as well as the covariates wage, EDU and SEX. The three estimators of COG (WLE, MLE, EAP) were all based on item responses alone.

Table 1 displays the results for this example and provides evidence of how estimates based on PVs and the true (generating) ability are quite similar, while estimates based on using EAP, MLE, or WLE are biased. The mean of the individual PVs is the EAP and, hence, a regression that uses the average of the 5 PVs rather than each of the 5 PVs in separate regressions will produce biased results that are similar to those obtained using the EAP. Note that these results are expected given the characteristics of the estimators. Also, similar effects were shown and explained by (von Davier et al. 2009) as well as OECD (2009), chapter 6.

It can be seen that the results obtained with true ability (the one used to generate the simulated data) and with the PVs agree more closely than those obtained with EAP, MLE, and WLE. The tabled values in boldface demonstrate that EAP, MLE and WLE do not fully control for ability in this regression, resulting in inflated estimates of the effect of EDU (educational level). The estimated effect of EDU is much reduced when the true (generating) theta or the PVs are employed, as both can fully control for this effect.

Table 2 shows the same regression estimated with log(wages) as the dependent variable. Apart from the scale change, the results appear to be quite similar. The estimated regression parameters when using true ability or the PVs agree well, while the parameter estimates from EAP, WLE and MLE are inflated for EDU and deflated for ability. What makes Table 2 particularly interesting is that one could argue that the conditioning model and the secondary analyst's model are not (particularly) congenial, as the log(wages) variable was not included in conditioning for the PVs.

Recent years have seen the problem of congeniality of the imputation model and substantive model earning greater attention. The main concern is not that all variables have to be in the same configuration in both models, but rather that the imputation model contains all variables (in the same transformed or untransformed form and, if needed, interacted with other variables). as the substantive model (Daniels et al. 2014; Bartlett

**Table 2 Estimated regression coefficients for a model predicting log(wages) based on skill variable theta (either true value, or PVs, or WLE, MLE, EAP estimates), gender and education**

| Estimator for θ | Intercept | θ | Gender | EDU |
|---|---|---|---|---|
| True ability | 2.2697 | *0.1074* | 0.0014 | *0.0038* |
| PVs | 2.2688 | *0.1067* | 0.0017 | *0.0067* |
| EAP | 2.2432 | ***0.0860*** | 0.0021 | ***0.0584*** |
| MLE | 2.2272 | ***0.0581*** | 0.0030 | ***0.0937*** |
| WLE | 2.2411 | ***0.0709*** | 0.0022 | ***0.0640*** |

Simulation setup similar to the one found in von Davier et al. (2009)

et al. 2014; von Hippel 2009; Quartagno and Carpenter 2016). However, log(wages) and wages are strictly monotone increasing functions of each other, so that a linear approximation of one by the other in a restricted interval is probably quite serviceable.

Although a single example is not definitive, it does suggest that the use of PVs as an independent variable can be a reasonable strategy when estimating a model that includes variables (or their strict monotone transforms) that were part of the imputation model for generating the PVs. This approach, performed separately for each set of PVs and then combined using the rules for calculations with multiple imputations proposed by Little and Rubin (2002) will allow researchers to evaluate the utility of the PVs as predictors in regressions.

## Discussion

As JR assert, the use of measures of cognitive skills in labor economic studies is becoming more common. Many econometricians, and other researchers as well, are unaware of the complex processes that generate cognitive data and their implications for analysis. In this article, we have focused on the issues that arise in the analysis of data from LSAS, which are quite different from those found in more traditional settings. The crucial distinction is that in the former case interest centers on estimating proficiency distributions at the group-level or population-level. Individual-level estimates are not of interest and are not produced.

We argue that the relevant psychometric literature on PVs, as well as the more general statistical literature on multiple imputations, gives reason for optimism. This is certainly the case when PVs are used as criterion variables and may generally be the case even when they are used as predictor variables. The latter question certainly deserves further attention, using both real and simulated data. Of course, following good statistical practice and exercising due caution in interpreting results is always recommended.

### Authors' contributions

HB and MvD developed the idea and the initial plan for the paper and analyses. MvD and HB developed an outline for the paper. HB drafted the first version and MvD reviewed this version. MvD revised the manuscript and HB reviewed the second version. Both authors contributed equally to the research. Both authors read and approved the final manuscript.

### Acknowledgements

We would like to Thank to Eugenio Gonzalez for running the reported analyses based on simulated data that emulate a simple wage equation.

### Competing interests

The authors declare that they have no competing interests.

### Availability of data and materials

Data from the small-scale simulation can be made available upon request.

### Consent for publication

We provide our consent to publish this manuscript upon acceptance for publication in the Springer open journal 'Large Scale Assessments in Education'.

### Ethics approval and consent to participate

No data collected on humans was used, the article contains mathematical arguments and simulated data only.

### Funding

Not applicable, no outside or grant funds were used.

## Publisher's Note

Received: 28 June 2017 Accepted: 20 October 2017
Published online: 07 November 2017

## References

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response modelling: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22*(1), 47–76.

Adams, R. & Wu, M. (2007). The mixed-coefficients multinomial logit model: a generalized form of the Rasch model. In M. von Davier & Carstensen, C. H. (Eds.), *Multivariate and mixture distribution Rasch models: extensions and applications* (pp. 57–76). New York: Springer.

Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society: Series B, 34*(1), 42–54.

Andersen, E. B. (2004). Latent regression analysis based on the rating scale model. *Psychology Science, 46*(2), 209–226.

Ballou, D. (2009). Test scaling and value-added measurement. *Education Finance and Policy, 4*(4), 351–383.

Bartlett, J., Seaman, S., White, I., & Carpenter, J. (2014). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research, 24*(4), 462–487.

Bauer, D. J., & Hussong, A. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods, 14*(2), 101–125. https://doi.org/10.1037/a0015583. **[PubMed: 19485624]**.

Bond, T. N., & Lang, K. (2013). The Evolution of the black–white test score gap in grades K-3: The fragility of results. *Review of Economics and Statistics, 95*(5), 1468–1479.

Braun, H. I., & Mislevy, R. M. (2005). Intuitive test theory. *Phi Delta Kappan, 86*(7), 489–497.

Briggs, D. C. (2008). Using explanatory item response models to analyze group differences in science achievement. *Applied Measurement in Education, 21*(2), 89–118.

Carlson, J. E., & von Davier, M. 2013. Item response theory. R&D Scientific and Policy Contributions Series SPC-13-05; Research Report 13–28, Educational testing service: Princeton. http://dx.doi.org/10.1002/j.2333-8504.2013.tb02335.x.

Cohen, J. D., & Jiang, T. (1999). Comparison of partially measured latent traits across normal populations. *Journal of the American Statistical Association, 94*(448), 1035–1044.

Daniels, M. J., Wang, C., & Marcus, B. H. (2014). Fully Bayesian inference under ignorable missingness in the presence of auxiliary covariates. *Biometrics, 70*(1), 62–72. https://doi.org/10.1111/biom.12121.

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah: Lawrence Erlbaum Associates Inc.

Firth, D. (1992). Generalized linear models and Jeffreys priors: An iterative generalized least-squares approach. In Y. Dodge & J. Whittaker (Eds.), *Computational statistics*. Heidelberg: Physica-Verlag.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika, 80*(1), 27–38.

Fuller, W. A. (2006). *Measurement error models*. Hoboken: Wiley.

Goldstein, H. (2004). International comparisons of student attainment: some issues arising from the PISA study. *Assessment in Education*. https://doi.org/10.1080/0969594042000304618.

Graham, J. W. (2012). *Missing data: Analysis and design*. New York: Springer.

Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *The Annals of Statistics, 5*(5), 815–841.

Jacob, B., & Rothstein, J. (2016). The measurement of student ability in modern assessment systems. *The Journal of Economic Perspectives, 30*(3), 85–107.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London, 186*(1007), 453–461.

Junker, B., Schofield, L. S., & Taylor, L. J. (2012). The use of cognitive ability measures as explanatory variables in regression analysis. *IZA Journal of Labor Economics, 1,* 4.

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement, 24*(1), 65–81.

Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics, 27*(4), 887–906.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. Hoboken: Wiley.

Little, R. J. A., & Rubin., D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken: Wiley.

Lockwood, J. R., & McCaffrey, D. (2014). Correcting for test score measurement error in ANCOVA models for estimating treatment effects. *Journal of Educational and Behavioral Statistics, 39*(1), 22–52. https://doi.org/10.3102/1076998613509405.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

Magis, D. (2015). A note on weighted likelihood and bayes modal estimation for polytomous IRT models. *Psychometrika, 80*(1), 200–204. https://doi.org/10.1007/S11336-013-9378-5.

Marsman, M., Maris, G. K. J., Bechger, T. M., & Glas, C. A. W. (2016). What can we learn from plausible values? *Psychometrika, 81*(2), 274–289.

Mazzeo, J., & von Davier, M. 2008. Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results. *doc.ref. EDU/PISA/GB(2008)28.* https://www.researchgate.net/publication/257822388_Review_of_the_Programme_for_International_Student_Assessment_PISA_test_design_Recommendations_for_fostering_stability_in_assessment_results.

Mazzeo, J., & von Davier, M. (2013). Linking scales in international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook international large-scale assessment: Background, technical issues, and methods of data analysis*. Boca Raton: Chapman and Hall/CRC.

Meng, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science, 9*(4), 538–558.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49,* 359–381.

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80,* 993–997.

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

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29,* 133–161. https://doi.org/10.1111/j.1745-3984.1992.tb00371.x.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practices, 25*(4), 6–20.

Mislevy, R. J., & Sheehan, K. M. (1987). Marginal estimation procedures. In A. E. Beaton (Ed.), *The NAEP 1983/84 technical report (NAEP Report 15-TR-20* (pp. 293–360). Princeton: Educational Testing Service.

Moustaki, I., & Knott, M. (2000). Generalized latent trait models. *Psychometrika, 65,* 391–411.

Mullis, I. V.S., Martin, M., Ruddock, G., O'Sullivan, C., & Preuschoff, C. 2009. *TIMSS 2011 assessment frameworks*. TIMSS & PIRLS International Study Center: Boston College. http://timss.bc.edu/timss2011/downloads/TIMSS2011_Frameworks.pdf.

OECD (2009). PISA data analysis manual: second edition—ISBN 978-92-64-05624-4.

OECD (2013). PISA 2012 Assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy. OECD Publishing. http://dx.doi.org/10.1787/9789264190511-en.

Quartagno, M., & Carpenter, J. R. (2016). Multiple imputation for IPD meta-analysis: Allowing for heterogeneity and studies with missing covariates. *Statistics in Medicine, 35*(17), 2938–2954.

Rogers, A., & Blew, T. (2012). *DGROUP—manual for the ETS software*. Princeton: Educational Testing Service.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Hoboken: Wiley.

Schofield, L. S. (2015). Correcting for measurement error in latent variables used as predictors. *Annals of Applied Statistics, 9*(4), 2133–2152.

Schofield, L. S., Junker, B., Taylor, L. J., & Black, D. A. (2015). Predictive inference using latent variables with covariates. *Psychometrika, 80*(3), 727–747.

Skrondal, A., & Rabe-Hesketh, Sophia. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton: Chapman & Hall/CRC.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52,* 393–408.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work. *American Psychologist, 24*(2), 83–91.

UNESCO. 2011. *International standard classification of education*. UNESCO Institute for Statistics, Montreal, Quebec. http://www.uis.unesco.org/Education/Documents/isced-2011-en.pdf.

van der Linden, W. (2016). *Handbook of item response theory 1*. Boca Raton: Chapman and Hall/CRC.

von Davier, M. (1996). Wnmira 1.74. A program for estimating dichtomous and polytomous rasch models, mixture distribution rasch models, and latent class models. software manual. Institute for Science Education: Kiel.

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful? In M. von Davier & D. Hastedt (Eds.), *IERI monograph series: Issues and methodologies in large scale assessments 2*. Princeton: IERInstitute.

von Davier, M., & Rost, J. (1995). Polytomous Mixed Rasch Models. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models—Foundations, Recent Developments and Applications* (pp. 371–379). New York: Springer.

von Davier, M., & Sinharay, S. (2013). Analytics in international large-scale assessments: Item response theory and population models. In L. Rutkowski, M. von Davier, & David Rutkowski (Eds.), *Handbook international large-scale assessment: Background, technical issues, and methods of data analysis*. Boca Raton: Chapman and Hall/CRC.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2007). The statistical procedures used in National Assessment of Educational Progress: Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics 26* (pp. 1039–1055). Amsterdam: North Holland-Elsevier.

von Hippel, P. T. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology, 39*(1), 265–291.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427–450. https://doi.org/10.1007/BF02294627.

Wu, M. (2010). Comparing the similarities and differences of PISA 2003 and TIMSS. OECD Education Working Papers no. 32, OECD Publishing. http://dx.doi.org/10.1787/5km4psnm13nx-en.

Yamamoto, K., & Mazzeo, J. (1992). Item response theory scale linking in NAEP. *Journal of Educational and Behavioral Statistics, 17*(2), 155–173. https://doi.org/10.3102/10769986017002155.