Large-scale Assessments in Education

REVIEW

Open Access

Considerations for the use of plausible values in large-scale assessments

Paul A. Jewsbury[1*], Yue Jia[1] and Eugenio J. Gonzalez[2]

*Correspondence:
pjewsbury@ets.org

[1] ETS, 660 Rosedale Road, Princeton, NJ 08541, USA
[2] Boston College, 140 Commonwealth Ave, Chestnut Hill, MA 02467, USA

Abstract

Large-scale assessments are rich sources of data that can inform a diverse range of research questions related to educational policy and practice. For this reason, datasets from large-scale assessments are available to enable secondary analysts to replicate and extend published reports of assessment results. These datasets include multiple imputed values for proficiency, known as *plausible values*. Plausible values enable the analysis of achievement in large-scale assessment data with complete-case statistical methods such as *t*-tests implemented in readily-available statistical software. However, researchers are often challenged by the complex and unfamiliar nature of plausible values, large-scale assessments, and their datasets. Misunderstandings and misuses of plausible values may therefore arise. The aims of this paper are to explain what plausible values are, why plausible values are used in large-scale assessments, and how plausible values should be used in secondary analysis of the data. Also provided are answers to secondary researchers' frequently asked questions about the use of plausible values in analysis gathered by the authors during their experience advising secondary users of these databases.

Keywords: Plausible values, Large-scale assessments, Group-score assessments, Multiple imputation

Introduction

Large-scale assessments are administered to representative probability samples from a population with the intention to report group-level statistics, and not results for the individual participants.[1] The National Assessment of Educational Progress (NAEP) is a large-scale assessment that reports on what students in the United States know and can do in a variety of subject areas. The Trends in International Mathematics and Science Study (TIMSS), the Programme for International Student Assessment (PISA), and the Progress in International Reading Literacy Study (PIRLS) are international large-scale assessments that report on the knowledge and skills of students in countries around the world. Another international large-scale assessment, but for adults, is the Programme for the International Assessment of Adult Competencies (PIAAC), which provides

[1] Throughout this paper we refer to the individuals participating in a large-scale assessment—the people who take the assessment—as *participants*.

international comparisons of adults' skills and competencies. Large-scale assessments are also classified as either school-based or household-based, depending on the population and subpopulations of interest and how individuals from that population will be selected. School-based assessments are those that select participants from the within-school population (e.g., NAEP, PISA, TIMSS), and household-based assessments are those that select participants directly from households (e.g., PIAAC, Early Childhood Longitudinal Study [ECLS]).

In addition to assessing what participants know and can do in domains such as reading, mathematics, science, literacy, and problem solving, large-scale assessments also collect extensive demographic information about the participants and administer contextual questionnaires to the participants and persons associated with them (e.g., teachers, principals, or parents). The purpose of collecting these data is two-fold: to classify participants into reporting groups (e.g., based on gender, socioeconomic status, language, migration status, participation in school programs, or employment status), and to study the distribution of achievement within these reporting groups. As an example, in the case of school-based assessments, questionnaire responses from teachers and school leaders become rich sources of information about what students know and can do and the contexts in which learning takes place.

Detailed reports describing the assessment results are published periodically by the organizations that administer the assessments, and together with these reports, supporting databases are also published. The databases allow researchers and secondary users not only to replicate the results as a measure of transparency and quality control of the assessment program, but also to allow users to conduct additional analyses of the data in support of their own research questions. Thus, with the availability of databases for secondary research, policymakers and practitioners can use large-scale assessment data to inform a wide range of issues of interest.

The published databases contain one record per participant and include their responses to the contextual questionnaires, demographic and other group classification information, sampling weights, and multiple imputed proficiency estimates, also known as *plausible values*. Secondary researchers can use these plausible values, along with the provided sampling weights, to obtain appropriate achievement estimates and associated uncertainties for reporting groups in the population.

Analyzing multiply imputed data such as plausible values is not straightforward. And despite a relatively large collection of technical documentation and guidance publicly available to support secondary analysts (e.g., Beaton et al., 2011; Rutkowski et al., 2010; von Davier et al., 2009), there are lingering doubts and questions about the need for plausible values and how to use them properly during analysis of large-scale assessment data.

With secondary analysts as our intended audience, this paper has two main goals. First, we will explain why and how plausible values are useful. Second, we will present guidelines and recommendations for using plausible values during secondary analysis. This paper attempts to supplement and reinforce previous efforts to clarify misunderstandings about using plausible values in secondary research.

We have organized the body of this paper in several sections. In the following section on *Goals and Designs of Large-scale Assessments*, we describe and characterize large-scale assessments and their goals. This is followed by the section on *Score Estimation for*

Large-scale Assessments, in which we describe the methodology to generate scores that was designed to best suit the goals and design of large-scale assessments. The description is rather technical, but as there are technical reasons supporting the use of plausible values, this is unavoidable. In the final section on *Using Plausible Values in Secondary Research*, we present guidelines and recommendations for working with plausible values in secondary analysis. We conclude with a set of recommendations and list of resources available for secondary users of the data and respond to an extensive list of frequently asked questions in the Appendix.

Goals and designs of large-scale assessments

Broadly stated, the purpose of a large-scale assessment is to measure and report on groups of participants' knowledge and skills, or what they know and can do, in one or more domains. For example, the purpose of NAEP is to report on what students in the United States know and can do in a variety of subject areas (e.g., reading, mathematics, and science).

The assessments most familiar to the public are college admissions tests (e.g., university entrance exams such as the SAT and ACT in the United States, the Gaokao in China, or the A-levels in the United Kingdom), K-12 standardized tests (e.g., national curriculum assessments), and exit examinations such as end-of-course or graduation assessments. These assessments share a common goal: to report accurate and comparable scores for *individuals*. Students taking those assessments receive individual score reports. Scores from these assessments may be used to make potentially high-stakes decisions affecting an individual, such as college admissions, graduation, or assignment to a course. These scores are also meant to be directly comparable for the individuals within and between administrations. These types of assessments are known as *individual-score assessments*, even if large in scale.

Large-scale assessments, on the other hand, aim to accurately report on what *groups* of participants know and can do. Because of this, large-scale assessments are at times referred to as *group-score assessments* (Beaton & Barone, 2017; Mazzeo et al., 2006) to explicitly distinguish them from *individual-score assessments*. Participants in a large-scale assessment do not receive individual score reports, and no action that specifically affects them are expected to be taken based on their performance on the assessment. Their responses to the assessment items contribute to the estimation of the proficiency distribution for the groups to which they belong and represent (e.g., eighth graders nationwide). The main goal of large-scale assessments is to report on the distribution of knowledge and skills between and within groups of interest (e.g., groups defined by categories of gender, race and ethnicity, socioeconomic status, immigration status or country).

An assessment design is a comprehensive plan that guides all aspects of the development and administration of an assessment. The ideal assessment design varies based on the specific reporting goal one aims to optimize. The assessment design contains specifications for the content to be measured, types and difficulty range of the test questions, length and timing of the test, administration procedures, the analysis methodology used to estimate test scores, and how results will be reported. This careful tailoring of all aspects of the assessment in pursuit of the intended reporting goals explains why large-scale assessment designs differ, often markedly, compared to designs for individual-score assessments.

There are three key aspects that distinguish the assessment designs of large-scale assessments from those of individual-score assessments: sampling of individuals, sampling of items, and collection of contextual information.

Sampling of individuals

Typically, assessing all members of a target group to report on what they, in the aggregate, know and can do is impractical and unnecessary, much in the same way that tasting an entire pot of stew to assess its quality or drawing all the blood from a person for a platelet count would also be impractical and unnecessary. Therefore, large-scale assessments are designed to be administered to representative samples of participants. Selecting a representative sample requires a carefully defined sampling design that specifies, among many other details, an unambiguous definition of the population and the demographic characteristics or population subgroups that must be represented in the sample. Again, as in the case of tasting a stew, we want to make sure all the major ingredients are tasted before forming a judgment about its quality.

Depending on the characteristics of the population of interest, sampling designs for large-scale assessments can be complex, often making use of stratified and multistage procedures to select the sample. The goal of sampling of individuals is to ensure proper representation of the different groups that make up the population, as well as the internal composition of the groups on which scores will be reported. Comprehensive accounts of sampling designs used in large-scale assessments are provided by Rust and Johnson (1992) and Rust (2013).

Sampling of items

To adequately cover the broad range of contents and topics in a subject domain, and adequately cover subdomains for which results will be reported, a large pool of test questions, also referred to as assessment items, is needed. For instance, the grade 8 NAEP mathematics assessment, which measures and reports on students' knowledge and skills across five mathematics content strands,[2] typically comprises about 150 items. Administering all 150 items to each student to be able to report on the overall domain and each of the five content strands would require about five hours of testing time. Such a lengthy test would be far too burdensome for participating students and schools and, hence, untenable for a large-scale assessment. And as is the case for sampling individuals, administering all items to any one individual is unnecessary when the goal of the assessment is to report aggregate results for groups of participants, and not to individuals.

Therefore, a significant concern for large-scale assessments is covering the entire assessment domain across the different reporting groups without significantly increasing the burden on the individuals participating in the assessment. Consequently, large-scale assessment designs reflect the need to minimize the burden on participating schools and

[2] The five mathematics content strands assessed at grade 8 are number properties and operations; measurement; geometry; data analysis, statistics, and probability; and algebra. Results are reported separately for these five content subscales and for an overall mathematics scale.

individuals as a means of maximizing participation when there is no direct personal benefit to their participation. Additional benefits accrue from a shorter testing time, such as the potential for participants to remain more engaged and experience less fatigue while responding to the assessment.

Given that reducing burden on participants is integral to keeping participation rates high, as there are no individual consequences for the participants, and because the goal is to report results aggregated at the group level and not for individuals, large-scale assessments are not required to have all participants attempt to respond to all items, or even items in all the reported domains. Instead, each person assessed is asked to respond to only a relatively small portion—a sample—of the total item pool, which requires much less of the participant's time.[3] Selecting and administering subsets of items from an overall item pool in this manner is known as a *matrix item sampling* design.

Determining which items are administered to any participant can be done a priori, by assigning a fixed set of items either in the form of a printed booklet, a predetermined sequence of items pre-programmed on a computer delivery system, or by influencing the selection of items based on information about the individual obtained prior to or during the assessment period. In any of these contexts, participants are tasked with responding to a subset of the items in the assessment. Applications of matrix sampling designs in large-scale assessments are discussed in detail by Gonzalez and Rutkowski (2010) and Frey et al. (2009).

Collection of contextual data

Large-scale assessments collect contextual data that is then used to group participants, and report on the proficiency distribution within these groups. Some of the contextual data may be obtained from external sources, such as demographic information about the participants (e.g., gender) and where they are located (e.g., school location and region of the country), whereas some is collected via contextual questionnaires administered to participants, their school administrators, teachers, or parents.

Score estimation for large-scale assessments

The designs of large-scale assessments, while optimized to efficiently collect and report information at the group level, introduces several technical challenges in the analysis of the data. As individuals only receive a subset of the entire assessment under the matrix item sampling design, the set of items each participant receives may differ in terms of properties and content (e.g., content domain distribution, item difficulty, and reliability), and we have relatively little cognitive information on the reported domains from individuals compared to individual-score assessments. These differences in the psychometric properties of the assessments, and the sparsity of the information we have from each individual, precludes the applicability of conventional methods of aggregating the response data from individuals. An analysis methodology that is capable of accounting

[3] The exact length of the assessment administration varies across large-scale assessments but is usually less than two hours. NAEP administrations typically require less than 90 min of students' time, spending 60 min to answer the assessment questions and another 15 min to complete survey questionnaires.

for both of these challenges is required to ensure data from large-scale assessments can yield comparable and reliable assessment results.

Item response theory

In education assessment and research, the most common methods to estimate scores while accounting for differences in the assessment design are *latent trait* models. Latent trait models assume that test takers receive an assessment with items that relate to a common construct, but account for differences in the psychometric properties of the assessments, such as differences in difficulty and reliability of the assessments.

Large-scale assessments typically use latent trait models within the framework of item response theory (IRT), which are suited for the item types used in these assessments. IRT models typically assume that the construct within a given domain can be characterized by a single dominant dimension. The items can be ordered along this dimension going from easy to difficult, depending on how much proficiency is required to be likely to answer the item correctly. Participants can also be ordered along this dimension going from less proficient to more proficient, depending on how likely they are to answer the items correctly.

Within IRT, different item types may be modelled with different item response functions. For example, the two-parameter logistic (2PL) model is commonly used for dichotomously-scored items where the correct answer cannot be guessed without the knowledge or skills being measured. Under the assumptions of this model, the response to an item is a function of how much of the latent trait (e.g., mathematics or reading proficiency) the participant possesses, how difficult the item is, and how differences in performance on the item reflects differences in proficiency of the test taker. Under the 2PL model, the probability of a correct response is defined as:

$$P\left(Y_j = 1|\theta_i, \boldsymbol{\beta_j}\right) = \text{logistic}\left(a_j\left(\theta_i - b_j\right)\right)$$

where Y_j is the response variable for item *j* scored as 1 (correct) or 0 (incorrect), θ_i is the latent proficiency for participant *i*, $\boldsymbol{\beta_j} = \{a_j,b_j\}$ is the item parameters for item *j*, a_j is the discrimination for item *j*, b_j is the difficulty of item *j*, $P(\cdot|\cdot)$ is a conditional probability and logistic() is the standard logistic function.[4]

For other item types, other models such as the three-parameter logistic (3PL) model and the generalized partial credit model (GPCM; Muraki, 1992) can be used.

To model the full pattern of a participant's item responses on the assessment, the item responses are assumed to be *locally independent*, that is,

$$P\left(\boldsymbol{Y} = \boldsymbol{y}_i|\theta_i, \boldsymbol{\beta}\right) = \prod_{j=1}^{J} P(Y_j = y_{ij}|\theta_i, \boldsymbol{\beta}_j)$$

where $\boldsymbol{Y} = \left\{Y_j\right\}_{j=1}^{J}$ is the vector of response variables for all items, $\boldsymbol{y}_i = \left\{y_{ij}\right\}_{j=1}^{J}$ is the vector of responses for participant *i* on all items, y_{ij} is the response for participant *i* on

[4] The standard logistic function is defined as $\text{logistic}(k) = \frac{1}{1+e^{-k}}$. In many large-scale assessments, there are multiple latent proficiencies associated with different subscales or domains. Here, one latent proficiency is assumed to simplify the presentation.

item j, $\boldsymbol{\beta} = \left\{\boldsymbol{\beta}_j\right\}_{j=1}^{J}$, and J is the total number of items. That is, the probability of a response to an item is a function of the underlying proficiency of the participant (θ_i) and is otherwise independent of responses on all other items in the assessment.

When IRT is used to summarize or aggregate large-scale assessment response data, the matrix item sampling design becomes a missing data problem: for each individual, responses to most items are missing, as those items are not presented to the individual. The assignment of items is usually randomized to ensure this missing data for the item responses can be classified as Missing Completely at Random (MCAR; Rubin, 1976), but that is not always the case.

More recently, assignment of items that are adaptively targeting at students' performance estimated on their responses to items administered and scored on-the-fly during the assessment have been implemented by some large-scale assessment programs. The adaptive assignment of items can be classified as Missing at Random (MAR), provided that an appropriate IRT model is used for the design (e.g., Jewsbury & van Rijn, 2020, in press).

To estimate the parameters of the IRT model in the presence of missing data from matrix item sampling, the following likelihood is maximized:

$$L\left(\boldsymbol{\beta},\mu,\sigma^2|\boldsymbol{y}\right) = \prod_{i=1}^{N} w_i \int \prod\nolimits_{j\in s_i} P\left(Y_j = y_{ij}|\theta_i,\boldsymbol{\beta}_j\right) f(\theta_i|\mu,\sigma^2)d\theta_i$$

where $L(\cdot|\cdot)$ is a likelihood, μ is the mean of the latent variable, σ^2 is the variance of the latent variable, $\boldsymbol{y} = \left\{\boldsymbol{y}_i\right\}_{i=1}^{N}$, N is the total number of participants, w_i is the sampling weight for participant i, s_i is the set of items that participant i received, and $f(\cdot|\cdot)$ is a density function.

Finally, the metric of the latent proficiency is set to ensure the comparability of scores between assessments administered in different years or with different administration modes with *linking methods*. These linking methods can usually be classified as either common-item or common-population linking (for more detail, see Jewsbury, 2019; Jewsbury et al., 2020; Jewsbury, Jia & Xi, in press; Yamamoto & Mazzeo, 1992).

IRT-latent regression

While IRT enables the scoring of large-scale assessment data and reporting results on a common scale, another challenge is the optimal estimation of the characteristics of the proficiency distribution for a population and each subpopulation. A general problem in statistics is that the optimal estimator for one level of inference may not be optimal for another level (e.g., Efron & Morris, 1977; Stein, 1956). In the context of IRT, this trade-off manifests itself in the observation that aggregating optimal individual-level scores may not result in optimal estimation of the mean or standard deviation of the proficiency distribution for a subpopulation (Mislevy et al., 1992), where optimality is defined as minimum bias or mean squared error. For example, when the maximum likelihood estimate (MLE) is calculated for each participant's proficiency, taking the mean or variance of these MLEs for a group of participants does not produce the maximum likelihood estimate of the mean or variance of the group (Mislevy et al., 1992; von Davier et al., 2009; Wu, 2005). Instead, this estimated mean from MLEs of the participants tends to be outwardly biased towards more extreme values, and the variance tends to be overestimated.

Because large-scale assessments are focused on reporting group estimates, such as the mean and variance of proficiency within a population or subpopulation, methods that produce optimal group-level inference are preferred over methods that produce optimal individual-level inference. In the context of IRT, IRT-latent regression models (IRT-LRM; Mislevy, 1984, 1985) can be used to obtain optimal group-level estimates.

In large-scale assessments, IRT-latent regression models are estimated by maximizing the following marginal likelihood,

$$L\left(\boldsymbol{\Gamma}, \sigma^2|\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\beta}\right) = \prod_{i=1}^{N} w_i \int \prod\nolimits_{j\in s_i} P\left(Y_j = y_{ij}|\theta_i, \boldsymbol{\beta}_j\right) f(\theta_i|\boldsymbol{\Gamma}, \sigma^2, \boldsymbol{x}_i) d\theta_i$$

where $\boldsymbol{x} = \{\boldsymbol{x}_i\}_{i=1}^{N}$, $\boldsymbol{x}_i$ is the vector of contextual information for participant i, $\boldsymbol{\Gamma}$ is the vector of regression coefficients, and σ^2 is the residual variance.[5]

Based on the IRT-latent regression, estimators for group-level parameters of interest may be defined for any group g, where a dummy-coded variable indicating membership in group g is included in $\boldsymbol{x}$. For example, the IRT-latent regression estimator for the mean of proficiency in group g is

$$\widehat{\mu}_g = \widehat{\boldsymbol{\Gamma}}' \overline{\boldsymbol{x}}_g$$

where $\widehat{\boldsymbol{\Gamma}}$ is the maximum likelihood estimator for $\boldsymbol{\Gamma}$, and $\overline{\boldsymbol{x}}_g$ is the mean of $\boldsymbol{x}$ within group g.

The IRT-latent regression estimator for the variance of proficiency within group g is,

$$\widehat{\sigma}_g^2 = \widehat{\boldsymbol{\Gamma}}' S_{\boldsymbol{x}g} \widehat{\boldsymbol{\Gamma}} + \widehat{\sigma}^2$$

where $S_{\boldsymbol{x}g}$ is the variance–covariance matrix for $\boldsymbol{x}$ within group g.

Multiple imputations

The statistical inappropriateness of using estimators designed for optimal individual-level inferences presents a challenge for secondary users of large-scale assessment data to conduct research and to replicate official results. While IRT-latent regression models yields estimators for group-level parameters of interest, not all researchers have the sophistication or computational resources available to directly estimate and interpret IRT-latent regression models, which presents a challenge for secondary users to replicate official results or conduct their own analyses with the data.

To aid in secondary user analysis and allow for official results to be easily replicated by secondary users, large-scale assessments use imputation methods to produce imputed values for latent proficiency. The intention underlying the use of imputed values is to allow secondary users to calculate estimates for group-level parameters that

[5] Usually, θ_i conditional on $\boldsymbol{x}_i$ is assumed to be normally distributed, with a mean of $\boldsymbol{\Gamma}\prime\boldsymbol{x}_i$ and a variance of σ^2. Note that in practical and operational applications the item parameters are usually estimated prior to the IRT-latent regression analysis, and the estimates of the item parameters are treated as the known and fixed values of $\boldsymbol{\beta}$ when estimating the latent regression parameters $\{\boldsymbol{\Gamma}, \sigma^2\}$ (see Jewsbury, 2023).

are consistent with the IRT-latent regression estimators described above, but with more familiar statistical methods.

While methods to produce a single imputed value for latent proficiency for each participant are available, methods involving multiply imputed values are generally recommended when there is substantial uncertainty in the imputation (Sinharay et al., 2001). Specifically, the method of *multiple imputation* (Rubin, 1987) is commonly recommended as a statistically principled approach for dealing with missing data (e.g., Schafer & Graham, 2002), enabling consistent estimation of parameters in data with missingness under the assumption that the missing data is missing at random (MAR; Rubin, 1976).

Multiple imputation has two benefits over many other imputation methods. First, multiple imputation is justified by statistical theory, in contrast to some common methods that are ad hoc and generally produce invalid inferences, such as mean substitution (Rässler et al., 2013). Relatedly, these common methods produce estimates that are inconsistent with the IRT-latent regression estimators. Second, the use of multiple imputations, instead of a single imputation, both reduces the uncertainty in estimates associated with the imputation process, as well as enables the uncertainty due to imputation to be accounted for in the standard errors (Rubin, 1976; Sinharay et al., 2001). As a result, the institution sponsoring the large-scale assessment program creates the imputed data for both official and secondary user analysis and reporting (e.g., Martin et al., 2020; NCES, 2024; OECD, 2019a, 2019b, 2023; von Davier et al., 2023). The institution packages the imputations in data products for secondary users to replicate published reports and create new ones of their own. This practice provides transparency to the officially reported results and allows for external quality control and verification of the reports.

A key concept in multiple imputation is *congeniality*, where the analysis models applied to the imputed data must correspond to the imputation model used to generate the imputed values (Meng, 1994). For example, consider a regression analysis where some values on the dependent variable are MAR, as is the case for the proficiency scores from the assessment. Multiple imputation may be used to impute values for these missing values, allowing an institution to provide a completed dataset to secondary users. While the imputed values are generally not the results that would have been observed had all responses been obtained, nevertheless, the secondary user may conduct the regression analysis with the imputed values and obtain valid regression parameter estimates as-if the data matrix were complete. In this context, congeniality would be guaranteed if the independent variables used by the secondary user were used in a regression-based imputation model (Meng, 1994).

Plausible values

While multiple imputation is usually applied to impute values for missing responses on observed variables (Rubin, 1987), the multiple imputation method was generalized for use in latent trait models and IRT by Mislevy (1991). Mislevy's (1991) insight was that the latent variable in IRT representing proficiency may also be considered a variable missing for all participants. As missingness on the latent variable in IRT is guaranteed to be MAR, the assumption underlying multiple imputation, values may be imputed on the latent variable for all participants. Here, we defined *plausible values* as multiply imputed values for latent variables.

With values imputed for the latent variable, secondary users can then use standard statistical routines and programs that require complete data to calculate estimates that are consistent with the IRT-latent regression estimators. However, some additional steps are required to analyze the plausible values, as described in the section Using Plausible Values in Secondary Research, below. The analysis challenges associated with the latency of proficiency and the massive amount of missingness in the item responses are thus eliminated for the secondary user.

Specifically, plausible values are generated as random draws from an approximation to each participant's conditional proficiency distribution,

$$f\left(\theta_i|\boldsymbol{x}_i,\boldsymbol{y}_i,\boldsymbol{\beta},\boldsymbol{\Gamma},\sigma^2\right)\propto\prod\nolimits_{j\in s_i}P\left(Y_j=y_{ij}|\theta_i,\boldsymbol{\beta}_j\right)f(\theta_i|\boldsymbol{\Gamma},\sigma^2,\boldsymbol{x}_i)$$

In the case of large-scale assessments, multiple plausible values are drawn for each participant to account for uncertainty in the estimate of each participant's proficiency. The specific procedure to approximate the conditional proficiency distribution and draw plausible values can differ between large-scale assessment programs. Specific details are provided in the technical documentation for each study (e.g., Martin et al., 2020; NCES, 2024; OECD, 2019a, 2019b, 2023; von Davier et al., 2023) or in technical publications (e.g., Jewsbury, 2023; von Davier & Sinharay, 2013; von Davier et al., 2007).

In the context of large-scale assessments, the requirement of congeniality implies that, in order to support a potential secondary user's analysis involving estimating the mean proficiency for a given group of participants, a variable indicating membership in the group should be accounted for in the IRT-latent regression model (Mislevy, 1991). To account for the group membership variable, it may be included as a predictor in the regression (within $\boldsymbol{x}$). If the variable is independent of proficiency conditional on the predictors included in the regression, the variable does not need be included (Meng, 1994; Mislevy, 1991). Note that the amount of bias that may be introduced by a non-congenial analysis is not only dependent on the strength of the relationship between proficiency and that variable unaccounted for by the other variables included in the regression model, but also the reliability of the assessment (Marsman et al., 2016).

Anticipating a large number of potential secondary user analyses, and to avoid inadvertently excluding important variables, a large number of predictors are typically included in the IRT-latent regression model (Mislevy, 1991). This follows generally recommended practices in the multiple imputation literature that support being as inclusive as possible in constructing the imputation model (Collins et al., 2001).

To reduce model overfitting while maximally accounting for variance in the contextual variables, a principal components analysis is typically applied to the full set of contrast-coded[6] contextual variables as a data reduction technique. Using contrast-coded data allows also for easy inclusion of non-linear and interaction effects in the model, and

[6] Contrast-coding in this context refers to the replacement of a *K*-category discrete variable with *K*-1 dichotomous dummy variables. For example, a three-category variable would be replaced with two dummy variables, where the first category is coded as (0,0), the second category is coded as (0,1), and the third category is coded as (1,1) on the dummy variables.

using principal components addresses issues of collinearity among the predictor variables. The predictors used in the regression are typically the smallest subset of principal components that account for a large proportion of the variance of the contextual variables, usually between 80 and 90% of the variance (von Davier et al., 2006). In some large-scale assessments, when samples are relatively small, the maximum number of principal components may be restricted as a function of the sample size (OECD, 2019a, 2019b).

While large-scale assessments may have been the earliest use of multiple imputation for latent variables (see Rubin, 1996), these methods have been applied more recently and more widely in other latent variable contexts (e.g., Boeschoten et al., 2018; Carlin, 1992; Dicke et al., 2015; Gorter et al., 2015; Joinson et al., 2012; Rhee et al., 2013). The popular structural equation modelling software *Mplus*, for example, enables researchers to use plausible values for a wide range of latent variable models (see Asparouhov & Muthén, 2010). Many of these applications employ plausible values for the purpose of missing data treatment in primary user analysis, as is recommended by some authors (Bray et al., 2015; Graham, 2003), similar to how multiple imputation is often used as a general-purpose missing data treatment with observed variable missingness (Schafer & Graham, 2002; Sinharay et al., 2001).

Summary

In summary, plausible values are imputations of individuals' abilities based on their cognitive and contextual data, where the relationships between proficiency with cognitive response and contextual data has been estimated via IRT-latent regression. By imputing proficiency values and adding them to the rectangular matrix containing the contextual data, users of the data are then able to analyze the database with readily available statistical software, eliminating the need for complex or specialized programming and computing skills. As an added advantage, different users of the database can replicate the results obtained by other users and those of the sponsoring organization.

A simple step-by-step description of this analysis process for the generation of plausible values follows:

- Step 0: Estimate the dependency of the item responses on latent proficiency with the use of the IRT model.
- Step 1: Estimate the IRT-latent regression parameters (regression coefficients and residual variance–covariance matrix) that characterize the relationships between the contextual data and proficiency.
- Step 2: Using the estimates from steps 0 and 1, calculate the distribution of proficiency for each participant, based on their item responses and contextual data.
- Step 3: Generate imputed proficiency data for participants, or *plausible values*, from the calculated individual proficiency distributions.
- Step 4: Analyze the plausible values. For example, compute group-level summary statistics with the imputed data.

This process is sometimes misunderstood as circular, as if Step 4 informs or affects Step 1. But the agreement between the summary statistics in Step 4 and the regression coefficients in Step 1 used to impute the data simply confirms that steps 2 and 3, the imputation process, were successful and that the results reflect the estimates used for imputing the data.

One clarification to this process worth mentioning here is that in Step 1, each regression parameter estimate has an associated standard error. This error reflects the uncertainty due to sampling of individuals from the population of interest. In Step 2, rather than using the point estimate for each of the regression parameters, or the same set of values for all plausible values, the set of regression parameter estimates used for the imputations are drawn from the distribution of regression parameter estimates, thus accounting for the uncertainty associated with them. Consequently, the plausible values simultaneously reflect the combined measurement and sampling variance associated with the IRT-latent regression model.

This paper emphasizes the history of plausible values as a generalization of Rubin's multiple imputation methodology for missing data to situate plausible values within the wider research literature. However, plausible values may also be understood within a Bayesian framework without reference to missing data. In the latter context, plausible values are random draws from each participant's posterior distribution of proficiency, conditional on their item responses and contextual information. The IRT model defines the likelihood, and latent regression defines the prior.

Within the Bayesian framework, the distinction between plausible values in large-scale educational assessments and conventional scores used in individual-score assessments is clarified:

- Plausible values are random draws from each participant's proficiency distribution, which may be associated with greater bias and random error in individual-level inference than the mean or mode (expected a priori, EAP, or maximum a posteriori, MAP, respectively; Mislevy et al., 1992).
- Plausible values in large-scale educational assessments are drawn from proficiency distributions conditional on both item responses and contextual information, while scores in individual-score assessments usually only directly depend on item responses. As the contextual information is not part of the domain being measured by the assessment, individual-level inference with plausible values in large-scale assessments may be inappropriate and unfair.

Consequently, while plausible values may be alternatively understood within a multiple imputation or a Bayesian interpretation, both interpretations motivate the same guidelines in the analysis of plausible values. Specifically, plausible values should be used to estimate group-level parameters following the guidelines outlined in the next section and are not appropriate for making inferences about any individual participant.

Using plausible values in secondary research

Large-scale assessments are rich sources of data that are used to inform many policy-relevant topics. Large-scale assessment data are intended be used to describe different groups within the population of interest based on their performance on an assessment of skills and knowledge. These descriptions can entail ranking the different groups against each other and establishing an interpretable distance between the means of these groups to determine whether the observed group mean differences are statistically significant, at one point in time and over time. The data can also be used to test the hypothesized relationship between performance on the assessment and a wide range of demographic and contextual data about groups of participants.

In addition to the sponsoring institution, stakeholders, policymakers, and applied researchers may also be interested in analyzing the large-scale assessment data to address a wide range of research questions. This type of data use is commonly referred to as secondary analysis or secondary research.

Users of large-scale assessment data analyze published databases with various statistical software and programming languages, including SPSS, SAS, STATA, and R. Most statistical tools, commercially or freely available, work with what are called *rectangular data matrices*. These rectangular data matrices are characterized by having rows for each participant in the assessment and columns for each variable collected in the assessment, as well as many other combinations of the original response variables. Plausible values are created and included in large-scale assessment databases so that the data can be analyzed with readily available statistical programs that expect data in rectangular form. However, working with plausible values requires special treatment and consideration. In this section we will cover these special considerations and how they affect secondary analysis with plausible values.

Calculating point estimates

Point estimation involves the use of sample data to calculate a single value for a statistic, which is to then serve as a "best estimate" of an unknown population parameter. A point estimate could be a mean, a standard deviation, a regression coefficient, a group difference, or the percentage of students reaching a benchmark or proficiency level. Calculating point estimates with imputed data, such as plausible values, requires the calculation of the statistic, separately, with each set of plausible values, and then reporting the average for the statistic as the point estimate. This means that the analysis needs to be repeated as many times as there are plausible values in the database.

A user might be tempted to work with the average of the plausible values instead, and perform the calculations once, thus saving processing time. While this approach will yield the correct point estimate for means and mean differences, it does not work with other statistics that would be influenced by the dispersion of the data, such as percentiles, correlations and the percentage reaching achievement benchmarks. And even if the correct point estimate could be obtained, a second disadvantage of working with the average of the plausible values is that the proper interval estimate, also known as the standard error for the statistic, could not be calculated. Without the interval estimate for the statistic, significance could not be properly tested, as we demonstrate in the next section.

The point estimate, or $\overline{\varepsilon}$, is then calculated using the following formula, where ε_p is any point estimate calculated from each of the P plausible values.

$$\bar{\varepsilon} = \frac{1}{P}\sum_{p=1}^{P}\left(\varepsilon_p\right)$$

Note that although the point estimate will be valid regardless of how many plausible values are used; using fewer plausible values than available in the data will likely yield a different and nosier point estimate than would be estimated by using all plausible values.

Calculating interval estimates

Interval estimation is the use of sample data to estimate an interval of possible values of a parameter of interest. The most prevalent form of interval estimation is a confidence interval, which is calculated using the standard error of the statistic. When working with large-scale assessment data, and more specifically with plausible values, there are two components to the estimation of the standard error of the statistic. These are the within- and the between-imputation variance components. Note that in much of large-scale assessment literature, these components are traditionally referred to as "sampling" and "measurement" variance, respectively. However, as noted earlier, the plausible values are generated using random draws of the latent regression coefficients themselves, thus reflecting uncertainty from the sample selection of individuals, as well. Because of this, we prefer to follow the conventions in multiple imputation literature that apply more generally (e.g., Rubin, 1976), referring to the variance components used to calculate the interval estimates as within- and between-imputation variances.

The within-imputation variance component is computed in a manner analogous to calculating the variance of a point estimate. This typically involves estimating the variance separately for each set of plausible values and, subsequently, averaging the results. To achieve this, conventional methods for standard error variance calculation are employed for each set of plausible values, treating the plausible values as if they were observed scores. The resulting average is the within-imputation variance component.

Large-scale assessments typically have a complex multistage sampling design, so some common conventional procedures to calculate standard errors that assume a simple random sampling design do not apply and would generally underestimate standard errors (Rust, 2013). Instead, there are two alternative procedures commonly used in large-scale assessments: the jackknifing procedure and balanced repeated replicates procedure, or BRR. Under the jackknifing procedure, members of the sample are systematically deleted from the sample and their contribution effectively replaced by that of others within the same sample. Under the BRR procedure the relative contribution of some elements in the sample is systematically increased while that of others is decreased by the same factor. In the end, by altering the composition of the original sample, we obtain multiple samples that are referred to as replicate samples. The sampling variance is then calculated by accumulating the squared differences between the point estimates calculated using each of the replicate samples, and the point estimate calculated using the original sample. This sum is then multiplied by a factor (*f*), reflecting the specific replication approach used to rearrange the composition of the sample. Details of the specific implementation of the procedures in international large-scale assessments can be found in the corresponding technical report (e.g., Martin et al., 2020; NCES, 2024; OECD, 2019a, 2019b, 2023; von Davier et al., 2023).

The within-imputation variance WI_ε is then calculated using the following formula:

$$WI_\varepsilon = \frac{1}{P}\sum_{p=1}^{P}\left(f\sum_{r=1}^{R}\left(\varepsilon_{p,r} - \varepsilon_p\right)^2\right)$$

where $\varepsilon_{p,r}$ are the point estimates calculated from each of the replicate samples using plausible value p. The value of the factor f depends on the replication method (e.g., jackknife, BRR, or bootstrap). Details on the value of this factor can be found in the technical documentation for the relevant large-scale assessment study (e.g., Martin et al., 2020; NCES, 2024; OECD, 2019a, 2019b, 2023; von Davier et al., 2023). Some large-scale assessments, such as NAEP, approximate the within-imputation variance term using only the sum of squared differences calculated from the first plausible value, rather than the mean of all P calculations (NCES, 2024). Note that while the most well-known large-scale assessments currently use a replicate-samples method conforming to this paragraph, there are alternative valid methods for complex samples, such as Taylor series linearization and sandwich variance estimators that could also be used (Binder, 1983; Kish & Frankel, 1974).

While the within-imputation variance can be calculated by taking the average of the estimates from each set of plausible values, this estimate does not capture all uncertainties associated with the point estimate. The uncertainty between each set of plausible values, or the between-imputation variance component, must be calculated as well.

The between-imputation variance BI_e is related to the variability of the estimate calculated from each of the plausible values. The following formula is used:

$$BI_\varepsilon = \frac{\sum_{p=1}^{P}\left(\varepsilon_p - \overline{\varepsilon}\right)^2}{P-1}$$

It is also important to note here that using fewer plausible values than available in the data will likely yield a different interval estimate than would be estimated using all plausible values available in the data. See the Appendix for discussions and recommendations on how many plausible values to use in secondary analysis.

Once these two variance components have been calculated, they are combined to obtain the standard error SE_e for the point estimate. The following formula is then used:

$$SE_\varepsilon = \sqrt{WI_\varepsilon + \left(1 + \frac{1}{P}\right)BI_\varepsilon}$$

These calculations are generalizable to the calculation of the standard error for any point estimate that involves plausible values, whether they are means, differences between means, percentages, correlations, or regression coefficients.

For example, if we want to calculate the standard error for the difference between two groups, where $d = \overline{a} - \overline{b}$, and $\overline{a}$ and $\overline{b}$ involve the use of plausible values, the calculation will be as follows:

$$SE_d = \sqrt{\left[\frac{1}{P}\sum\nolimits_{p=1}^{P}\left(f\sum\nolimits_{r=\ 1}^{R}(d_{p,r}-d_p)^2\right)\right]+\left[\left(1+\frac{1}{P}\right)\frac{\sum_{p=1}^{P}\left(d_p-\overline{d}\right)^2}{P-1}\right]}.$$

Combining plausible values from different domains in the same analysis

It is often the case that a user wants to use plausible values from different domains in the same analysis. For example, using the PISA data, calculate the correlation between mathematics, science, and reading literacy. Or using the PIAAC data, calculate the likelihood of an adult being employed based on their literacy and problem-solving skills.

To understand how to use plausible values from different domains in the same analysis, it is helpful to briefly review how plausible values are created. Recall that the first step is estimation of a latent regression, and in the second step, the latent regression coefficients are used to create the imputed scores, or plausible values. When multiple domains are involved, the latent regression and subsequent imputation is carried out assuming the multiple domains are part of a multidimensional space, and therefore, the latent regression and subsequent imputation for the different domains is done simultaneously across all domains. As a result, the first imputation is for all domains simultaneously, then the second one, and so on. Because of this, the correlation between the first plausible values of two domains provides an unattenuated estimate of the relationship between the domains, whereas the correlation between the first plausible value for a domain and any other plausible value will be attenuated.

Therefore, when using plausible values from different domains in the same analysis, the recommended procedure is to do the calculations using the first plausible value from across the domains in the analysis, then the again using the second plausible value across all domains, and so on. The average of the statistics from across all the analysis will be the statistic to report, and the standard error is calculated as described earlier.

Combining plausible values with non-plausible values in the same analysis

In the interest of making different analysis scenarios explicit, we present an example. When combining plausible values with non-plausible values in the same analysis, the principles and recommendations described earlier apply. That is, conduct the analysis separately with each plausible value, treating the non-plausible value variable as constant for each case across the different analyses. Then, calculate the point estimates and interval estimates as described earlier.

Conducting analysis with model based selection of variables and cases

The analysis using plausible values can involve the selection of variables, such as when conducting multiple linear or logistic regression, or the selection of cases, such as discriminant analysis of assigning individuals to proficiency groups.

It can be somewhat confusing for the secondary user to run a linear regression using an automated variable selection procedure (i.e., backwards, forward, or stepwise selection) and obtain seemingly contradictory or unreconcilable results. Following the prescribed procedure for calculating statistics using each set of plausible values separately and then using the average of those statistics, the user could end up with a different selection of variables depending on which plausible value was used. In this case, the user should keep in mind that many of these selection procedures are not always directly applicable to data collected with complex sampling designs or imputations.

Users might be tempted to rely on a single plausible value for this analysis, which would provide a consistent estimate, but no information about the between-imputation error variance component. Furthermore, the estimate from a single plausible value would be less precise than the mean across the estimate for each set of plausible values. Users might also be tempted to use the average of the plausible values, but this would lead to underestimating variances and the relationship between the variables, not to mention not being able to calculate the between-imputation variance.

Instead, the recommended procedure is to run the regression analysis separately with each of the plausible values and report the average of the coefficients. The standard errors for the coefficients used in the selection of variables will come from combining the standard errors related to the sample selection, calculated using the replication method described earlier, and the standard error related to the imputation by using the variance of the coefficients computed with each of the plausible values.

The decision to include or exclude a variable in a model could defensibly be based on a more thought-out and theory-based approach to the inclusion of variables in an equation. The fact that a variable is included in a model with some plausible values and not included when other plausible values are used is not an indication of a problem with the method, but rather an indication that there is uncertainty about whether the variables should be included in the model, and the imputed scores reflect that.

On the flip side, when selecting cases into groups, the same principle applies of running the analysis separately with each plausible value and summarizing the resulting coefficients and calculating the errors of the coefficients using the prescribed formula. When the classification of cases is used in a subsequent analysis, the classification should also be treated as an imputed classification, and the classification should be done separately with each plausible value; then the analysis with the classified data would follow the same rules described earlier, that is, run the analysis with each classification and summarize the results accordingly. As with the selection of variables, when the classification of the cases varies depending on which plausible value is used, this reflects the uncertainty about where individuals should be classified or categorized.

Automated procedures for variable and case selection with multiply imputed datasets, such as plausible values, is an active and current area of research. Interested readers are directed towards recent methodology reviews and comparison studies (Gunn et al., 2023; Thao & Geskus, 2019; Zhao & Long, 2017). However, more research is needed, particularly in the context of large-scale assessments, to identify the optimal approach.

Considerations for analysis using standard statistical analysis tools (e.g., SPSS, SAS)

Analyzing data from large-scale assessments requires special treatment of the data during the analysis. The special treatment relates to obtaining consistent point estimates and their corresponding standard errors. Most standard analysis tools, such as SPSS, SAS, and STATA, are not readily capable of performing these operations automatically. These analysis tools require additional programming to properly treat the data and compute the correct statistics and standard errors according to the design.

However, there are several tools currently available that allow users to work with plausible values while using these statistical programs. For users of R, SPSS, or SAS, there is the IEA's IDB Analyzer, which was developed by the International Association for the Evaluation of Educational Achievement (IEA). The IDB Analyzer (IEA, 2021) is an application that can be used to combine and analyze data from most major large-scale assessments. Originally designed for IEA's international large-scale assessments, the IDB Analyzer has also been configured to work with national assessment databases, such as those from NAEP, PISA and PIAAC. The IEA's IDB Analyzer consists of two modules—the Merge Module and the Analysis Module—that are integrated into a single application. The Merge Module creates R, SPSS, or SAS syntax that can be used to combine files from large-scale assessment databases. The Analysis Module can be used to perform statistical analyses with these databases. Available analysis options include calculations of percentages and means, linear and logistic regression, Pearson and Spearman correlations, percentiles, and working with benchmarks of achievement. The IDB Analyzer works by creating R, SPSS or SAS syntax that incorporates information from the sampling design in the computation of sampling variance and handles the plausible values. The code generated by the IDB Analyzer enables the user to compute descriptive statistics and conduct statistical hypothesis testing among groups in the population without having to write any programming code. The most recent version of the IDB Analyzer also interfaces with SAS on Demand for Academics, a free online tool for statistical analysis available on the web.

For users of STATA, there is repest (Avvisati & Keslair, 2014). This application estimates statistics using replicate weights, thus accounting for complex sampling designs. It is specially designed to be used with the PISA, PIAAC, and TALIS datasets produced by the OECD. It also allows for analyses with multiply imputed values or plausible values.

For users of R, in addition to the IEA's IDB Analyzer, there is EdSurvey and the R-Analyzer for Large-Scale Assessments (RALSA). EdSurvey is an R statistical package designed for the analysis of national and international large-scale assessment data from the National Center for Education Statistics (NCES) in the U.S. The key functions of the current version of EdSurvey include downloading available databases and converting data to R; subsetting and merging of data; calculating summary statistics; calculating linear and logistics regression; multilevel modeling; direct estimation; gap analysis; percentiles; analysis of achievement levels and benchmarks, correlations, multivariate regression; quantile regression (Bailey et al., 2020). RALSA (Mirazchiyski, 2021) is also an R package for preparation and analysis of data from large-scale assessments and surveys that use complex sampling and assessment designs. The software can handle the design issues and apply the appropriate analytical methods automatically for every type of study, out of the box with little user intervention.

In addition to the tools described above, NCES has also made Data Explorers available. The Data Explorers are online analysis tools that are freely available for users to analyze data from most large-scale assessments, national and international, in which NCES participates, such as NAEP, PISA, PIAAC, TIMSS, and PIRLS. The Data Explorers are available from the NCES website and are designed to give users relatively quick and easy access to the data from large-scale assessments. Using the Data Explorers, a user can generate report tables, conduct linear regression analysis, conduct significance testing, and create diverse graphics without needing to have access to the databases or a statistical analysis program on the user's local machine. The user-friendly graphical interfaces of the Data Explorers facilitate analysis and report generation through simple point-and-click interactions. The user also has access to historical data from the assessment and, in most cases, to restricted-use data.

Summary

Throughout this paper we have described, in conceptual, technical, and practical terms, what plausible values are, why they appear in databases for large-scale assessments, and how researchers should use them in secondary analyses. The motivation for plausible values in large-scale assessments ultimately reflects the fact that these assessments serve a purpose different from individual-score assessments. Large-scale assessments aim to describe the knowledge and skills of populations and population subgroups rather than individual test takers, and this is reflected in the methods and procedures used to summarize the results from the assessment and, consequently, how to analyze and report them. Plausible values were designed to allow users to analyze large-scale assessment data with common statistical methods, enabling valid group-level inferences, despite the challenges associated with the complex assessment data structure.

Secondary analysts are often challenged by the complex and unfamiliar nature of plausible values, large-scale assessments, and their datasets. Users are encouraged to use the plausible values provided in the secondary-use databases and follow recommended procedures for using plausible values to compute statistics of interest and corresponding errors.

While this paper provides a broad overview of the meaning and use of plausible values, more detailed information on specific aspects of large-scale assessments is available in recently edited books (e.g., Lietz et al., 2017; Maehler & Rammstedt, 2020; Nilsen et al., 2022; Rutkowski et al., 2013), special journal issues on large-scale assessments (e.g., Cai, 2019; Stadler et al., 2016, 2017), technical reports (e.g., Martin et al., 2020; NCES, 2024; OECD, 2019a, 2019b, 2023; von Davier et al., 2023), and the specialized journal *Large-Scale Assessments in Education*.

Common misunderstandings and misuses of plausible values arise when users incorrectly assume that plausible values have the same meaning and statistical properties as test scores that are reported for participants from individual-score assessments. Throughout this paper, we intended to clarify and explain the meaning of plausible values, as well as to detail how best to analyze and interpret plausible values. We conclude this paper with an appendix list of frequently asked questions on the use of plausible values. These questions have been compiled by the authors based on common misconceptions and misuse of plausible values found in the literature.

Appendix: Frequently asked questions

The authors have extensive experience lecturing on the topic of large-scale assessments and training secondary users of large-scale assessment databases. What follows are some of the questions that most frequently come up during these lectures and trainings.

Why are results for large-scale assessments reported using plausible values?

Plausible values are used in reporting large-scale assessment results so that secondary users of the data can replicate the reported results using readily available statistical software and use the data to conduct their own research. Prior to generating the plausible values, the relationships between the contextual variables and proficiency are estimated via a latent regression. Using the model parameter estimates from this latent regression, combined with the data for each individual, the plausible values are generated. Plausible values may then be analyzed with standard statistical programs and methods.

Why are there multiple plausible values used in large-scale assessments?

Multiple plausible values are used following best practices in the imputation of missing data (Rubin, 1972). By repeating the analysis with each set of plausible values and averaging across the results, the final results are obtained with more precision than if one set of plausible values is used. Furthermore, calculating the variance of the results for each set of plausible values is part of the recommended standard error calculation. Depending on the large-scale assessment program, currently, anywhere from five to twenty plausible values are used.

How many plausible values should I use for my analysis?

We recommend using as many plausible values as are provided in the data file, as that is the number of plausible values used by the large-scale assessment program for reporting results. Doing so will facilitate your ability to replicate reported results and those of others. For exploratory purposes, you could use fewer plausible values, and when plotting graphics, you could use only one. But we recommend whenever possible using as many as are published in the available databases.

Why do different large-scale assessments use different numbers of plausible values?

Plausible values are a random sample of values from the estimated distribution of proficiency for each individual. As with any sample, more selections are better in terms of representing the estimated proficiency distribution, but at some point, more selections provide little additional information. As computer storage and processing power have become less expensive over the years, large-scale assessments have been able to incorporate more selections (i.e., more plausible values) for the estimation of proficiency and associated uncertainty. Therefore, the numbers of plausible values used by large-scale assessments vary, with some assessments using 5 plausible values, but others using up to 20 plausible values.

If my statistical analysis program can use only one plausible value, which one should I use?

On occasion, the program or procedure you are using might not allow you to use more than one plausible value (e.g., plotting a histogram). In this case we recommend using one set; however, once you have selected a set from the sequence of available plausible values, you should use the same set of plausible values for all cases. For example, you could select the 3rd set of plausible values to use for all cases. You should not select a different set of plausible values for each case, even if the selection is at random. Note that when using only one plausible value, you will not be able to compute the variance due to measurement imprecision.

In the data file, why does one person have different plausible values for the same subject?

The plausible values are random selections from the individual's estimated proficiency distribution. Since they are random selections from within a range of possible estimates, they are expected to vary. The more uncertainty we have about a person's proficiency, the more the plausible values are expected to vary. The less uncertainty we have, the less they will vary. This is one of the reasons why plausible values should not be reported to individuals as their "score on the test."

What is the proficiency distribution for an individual?

For any person responding to a large-scale assessment, we cannot estimate their proficiency precisely. All we have are the person's responses to the relatively small number of assessment items assigned to them, their responses to the background questionnaire administered to them, and any other contextual information we have about them. Using this information, and in combination with probabilistic models, we can estimate the likelihood of having the observed response pattern if the participant had a certain proficiency. Using this information, we can approximate a probability distribution of their proficiency. It is from this estimated proficiency distribution that plausible values are randomly selected.

In the data file, why do different individuals with the same performance on the assessment have different plausible values?

This is related to the question above. The plausible values are random selections from an individual's estimated proficiency distribution. Since they are random selections from within a range of possible estimates, they are expected to vary within and between individuals, even if they performed the same. The more uncertainty we have about the performance, the more these will vary. This is one of the reasons why plausible values should not be reported to individuals as their "score on the test." They could also vary because the individuals are from different groups in the population, and according to the latent regression these groups vary in the estimated proficiency. This is another reason why plausible values should not be reported to the individuals as their "score on the test."

What is wrong about using the average of the plausible values in my analysis instead of each of them separately?

There are two problems with doing this. First, using the average of the plausible values for the individuals will underestimate the variance of the proficiency estimates, because the variance of these averages will be an underestimate of the actual variance in the population. The second problem with using the average is that you will not be able to calculate the uncertainty due to measurement.

What does it mean to use variables for "conditioning"?

The process of estimating achievement in large-scale assessments consists of estimating a latent regression using variables believed to be potentially related to proficiency as predictors in the regression. The resulting coefficients from this latent regression tell us about the relationships between these independent variables and proficiency. We can then use these regression parameters to calculate predicted proficiency distributions. These predicted proficiency distributions play a role in the generation of the plausible values, to accommodate analyses involving regressing the plausible values on those contextual variables. When a variable is used as an independent variable in this latent regression, it is said that we are conditioning on this variable.

I hear that by conditioning on some variables, some groups of participants are given scores worse that they deserve based on their performance, is this true?

To the contrary, by conditioning on a variable we estimate with the latent regression the relationship between that variable and achievement, and therefore are able to use that variable in generating the plausible values that incorporates the estimated effect of that variable on proficiency. Depending on the relationships between proficiency and the independent variable, excluding a variable from the latent regression may lead to an underestimation of the effect of that variable when the institution generates the plausible values.

How do I know if the variables I am using in my analysis were used as conditioning variables?

Traditionally, all contextual variables are summarized by means of a principal component analysis, and the principal components accounting for 80 – 90% of the variance corresponding to the contextual variables are used as independent variables in the latent regression. While this is no guarantee that the contextual variables are represented in their entirety in the principal components used in the regression, most of the variance in the contextual variables are accounted for. You should consult the technical documentation of the corresponding assessment to find out how the contextual variables were selected and used in the model.

How are the background variables used as independent variables in the latent regression equation?

This depends on how the variable is coded. There is some pre-processing required. A categorical variable is converted to a set of contrast-coded variables. For example, a variable identifying public and private schools is converted to a dichotomous variable with a value of 1 assigned to public schools, and a 0 otherwise. If there is missing data for this variable, a second variable is created with a value of 1 for those cases missing this variable, and 0 otherwise. A continuous variable could be categorized into several groups, and then converted to a set of contrasted coded variables, or it could be used as-is, with the missing values replaced with a 'reasonable' value and creating an additional variable identifying the cases for which this was done. These converted variables can then be entered as-is, but what is more commonly done is that principal components are calculated from these variables and a selection of these are entered in the equation. This has the benefit of requiring fewer variables in the equation than the total, while accounting for a large proportion of the variance.

What should I do if a background variable I am interested in was not used as a conditioning variable in the analysis?

The standard practice is to use all contextual variables in the latent regression model, either directly or indirectly. Directly by using the unaltered variable in the model, or indirectly as part of the principal components. If the variable of interest was not used in the conditioning model at all, the effect for this variable on the estimated proficiency may be underestimated to the extent that this effect has not already been accounted by the other variables already included in the model.

How often are the effects from the variables in the regression equation calculated? If one group does relatively poor/well in an assessment, will they ever be able to show improvement/decline in subsequent assessments?

The latent regression coefficients are calculated for each assessment year, grade, and cycle. The fact that a group did poorly in one year has no effect on the regression model that is calculated for later years. Along the same lines, if a group does poorly in one assessment, this does not mean that they will do poorly in all assessments. For example, three latent regressions with different subject domain data could point to girls outperforming boys in reading, boys outperforming girls in math, and boys and girls doing about the same in science. A latent regression for Country or State A could point to girls outperforming boys in mathematics, but to boys outperforming girls in Country or State B in the same assessment year. The effects for the contextual variables are re-estimated for every assessment, and all participant-level contextual variables are typically used in the analysis.

Can I use plausible values with my statistics program?

Yes, plausible are created precisely so that they can be used with readily available statistics programs. However, when using them we recommend you follow the prescribed procedures for using them by running separate analysis using each plausible value and then reporting the average of the estimated statistics.

If I run the analysis separately with each plausible value, which results do I then report?

When using plausible values for an analysis, you will carry out the analysis separately using each of the plausible values and would then report the mean of the statistics computed. For example, when calculating a mean, you would report the average of the means obtained from each of the plausible values. When calculating a regression coefficient, you would report the average of the regression coefficients.

What does it mean that "plausible values are not individual scores"?

This is meant to be a warning to users of the data that plausible values are not meant to be used to report the results to individuals or otherwise make inferences about individuals. However, for the purpose of using readily available statistic programs to compute group-level statistics, the plausible values are there to be treated as individual scores for the cases in the data matrix.

How do I classify cases into achievement groups when each of them has several plausible values?

The classification of cases needs to be done separately with each plausible value, potentially resulting in different classifications depending on the plausible value used. The variability of this classification across members of the same group will give an estimate about the classification of the group across the different categories.

What should I do if I need to plot achievement results using the plausible values?

This is one of the few instances where you could pick one set of plausible values, and plot these.

I am not very good using statistical programs, is there a tool I could use to get answers from the assessment that uses plausible values?

Nowadays there are several tools available that incorporate the use of plausible values. If you use SAS or SPSS, we recommend the IEA's IDB Analyzer. If you are proficient with R, we recommend the package EdSurvey for the U.S. national and international assessments and the packages RALSA or IntSvy for the international large-scale assessments. Of these last three R packages, RALSA has a graphical user interface reducing the need for coding in R. If you have little experience with statistical programs, we recommend using the data explorers sponsored and supported by

NCES. The data explorers give you quick access to historical data from all the major national and international large-scale assessments, and a powerful data analysis tool.

Are the plausible values different manifestations of the same variable?

As per Mislevy (1993), plausible values bear a superficial resemblance to classical multiple indicators of a latent variable, as analyzed with structural equation modeling or hierarchical modeling. But analyzing them as such does not generally yield the correct results. Instead, they should be analyzed as imputed data by conducting separate analysis with each set of plausible values.

How do I work simultaneously with plausible values from two or more different domains? For example, correlate the plausible values from mathematics, science and reading literacy in PISA?

Before working with them simultaneously, you want to refer to the assessment's published technical documentation to confirm that the plausible values were estimated using a joint multidimensional model. If that is the case, you would analyze them in accordance to their construction by successively calculating the correlation between the first plausible values across all the domains, then the second one, and so on. You would then report the average of these correlation coefficients. You would not mix-and-match across different sets of plausible values.

Does the use of contextual data in the generation of the plausible values bias results for groups based on their demographics?

A common concern is that the use of contextual data may result in biased comparison of groups with different demographic data. This is an understandable concern given that the contextual data is used in the generation of the plausible values. However, the purpose of using the contextual data is to ensure that analysis with plausible values recovers the proficiency distributions that were estimated in the data with the IRT-latent regression. The use of contextual information does not bias the results towards previously collected data or non-data driven assumptions about the relative performance of demographic groups.

Abbreviations

2PL	Two-parameter logistic
3PL	Three-parameter logistic
EAP	Expected a priori
ECLS	Early childhood longitudinal study
GPCM	Generalized partial credit model
IEA	International Association for the Evaluation of Educational Achievement
IRT	Item response theory
MAP	Maximum a posteriori
MAR	Missing at random
MCAR	Missing completely at random
MLE	Maximum likelihood estimate
NAEP	National Assessment of Educational Progress
NCES	National Center for Education Statistics
OECD	Organisation for Economic Co-operation and Development
PIAAC	Programme for the International Assessment of Adult Competencies
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
RALSA	R-Analyzer for Large-Scale Assessments
TIMSS	Trends in International Mathematics and Science Study

Acknowledgements

The authors would like to thank Rebecca Moran, Jay Campbell, Katherine Castellano, Tim Davey, Robert Finnegan, Matthew Johnson and Daniel McCaffrey for helpful comments and suggestions. We also thank Emily Kerzabi for editorial assistance.

Author contributions

All authors were major contributors in developing and writing the manuscript. All authors read and approved the final manuscript.

Funding

The manuscript was funded by ETS.

Availability of data and materials

Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 14 December 2023 Accepted: 31 July 2024
Published online: 08 August 2024

References

Asparouhov, T., & Muthén, B. O. (2010). *Plausible values for latent variables using Mplus. Mplus Technical Appendix*. Muthén and Muthén.

Avvisati, F. & Keslair, F. (2014). *REPEST: Stata module to run estimations with weighted replicate samples and plausible values*, Statistical Software Components S457918, Boston College Department of Economics, revised 06 Jan 2020. https://ideas.repec.org/c/boc/bocode/s457918.html

Bailey, P., Lee, M., Nguyen, T., & Zhang, T. (2020). Using EdSurvey to Analyse PIAAC Data. In D. Maehler & B. Rammstedt (Eds.), *Large-scale cognitive assessment*. Cham: Springer. https://doi.org/10.1007/978-3-030-47515-4_9

Beaton, A. E., & Barone, J. L. (2017). Large-scale group-score assessment. In R. E. Bennett & M. von Davier (Eds.), *Advancing human assessment: The methodological, psychological, and policy contributions of ETS* (pp. 233–284). Springer. https://doi.org/10.1007/978-3-319-58689-2_8

Beaton, A. E., Rogers, A. M., Gonzalez, E., Hanly, M. B., Kolstad, A., Rust, K. F., Sikali, E., Stokes, L., & Jia, Y. (2011). *The NAEP Primer* (NCES 2011–463). U.S. Department of Education, National Center for Education Statistics. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2011463

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review/Revue Internationale de Statistique*, *279*–292.

Boeschoten, L., Oberski, D. L., De Waal, T., & Vermunt, J. K. (2018). Updating latent class imputations with external auxiliary variables. *Structural Equation Modeling: A Multidisciplinary Journal, 25*, 750–761.

Bray, B. C., Lanza, S. T., & Tan, X. (2015). Eliminating bias in classify-analyze approaches for latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 22*, 1–11.

Cai, L. (2019). Introduction to the Special Issue on Research and Development on Large-Scale Educational Assessment Programs. *Journal of Educational and Behavioral Statistics, 44*(6), 647–647.

Carlin, J. B. (1992). Meta-analysis for 2 × 2 tables: A Bayesian approach. *Statistics in Medicine, 11*, 141–158.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*, 330–351.

Dicke, T., Parker, P. D., Holzberger, D., Kunina-Habenicht, O., Kunter, M., & Leutner, D. (2015). Beginning teachers' efficacy and emotional exhaustion: Latent changes, reciprocity, and the influence of professional knowledge. *Contemporary Educational Psychology, 41*, 62–72.

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American, 236*(5), 119–127.

Frey, A., Hartig, J., & Rupp, A. (2009). An NCME instructional module on booklet designs in large scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28*(3), 39–53.

Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. In M. von Davier & D. Hastedt (Eds.), *IERI Monograph Series: Issues and Methodologies in Large Scale Assessments*, *3*, 125–156.

Gorter, R., Fox, J. P., & Twisk, J. W. (2015). Why item response theory should be used for longitudinal questionnaire data analysis in medical research. *BMC Medical Research Methodology, 15*, 55.

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 10*, 80–100.

Gunn, H. J., Hayati Rezvan, P., Fernández, M. I., & Comulada, W. S. (2023). How to apply variable selection machine learning algorithms with multiply imputed data: A missing discussion. *Psychological Methods, 28*(2), 452.

International Association for the Evaluation of Educational Achievement (2021). *IEA International Database Analyzer (IDB Analyzer)*. IEA. https://www.iea.nl/data-tools/tools#section-308

Jewsbury, P. A. (2019). *Error variance in common population linking bridge studies* (Research Report No. RR-19-42). Educational Testing Service.

Jewsbury, P. A. (2023). Educational surveys: Methodological foundations. In R. Tierney, F. Rizvi, & K. Ercikan (Eds.), *International encyclopedia of education: quantitative research/educational measurement*. Elsevier.

Jewsbury, P. A., Finnegan, R., Xi, N., Jia, Y., Rust, K., & Burg, S. (2020). *2017 NAEP Transition to Digitally Based Assessments in Mathematics and Reading at Grades 4 and 8: Mode Evaluation Study*. White paper published by the National Center for Education Statistics. Retrieved from https://nces.ed.gov/nationsreportcard/subject/publications/main2020/pdf/transitional_whitepaper.pdf

Jewsbury, P. A., Jia, Y., & Xi, N. (in press). Effects of mode transition on instruments and subpopulation performance in NAEP. In L. Khorramdel, M. von Davier, & K. Yamamoto (Eds.), *Innovative Computer-based International Large-Scale Assessments—Foundations, Methodologies and Quality Assurance Procedures*. Springer.

Jewsbury, P. A., & van Rijn, P. W. (2020). IRT and MIRT models for item parameter estimation with multidimensional multistage tests. *Journal of Educational and Behavioral Statistics, 45*(4), 383–402.

Jewsbury, P. A., & van Rijn, P. W. (in press). Item calibration in multistage tests. In D. Yan, D. J. Weiss, & A. A. von Davier (Eds.), *Research for Practical Issues and Solutions in Computerized Multistage Testing*. Taylor & Francis.

Joinson, C., Heron, J., Araya, R., Paus, T., Croudace, T., Rubin, C., Marcus, M., & Lewis, G. (2012). Association between pubertal development and depressive symptoms in girls from a UK cohort. *Psychological Medicine, 42*, 2579–2589.

Kish, L., & Frankel, M. (1974). Inference from complex samples. *Journal of the Royal Statistical Society: Series B, 36*, 2–37.

Lietz, P., Cresswell, J., Rust, K. F., & Adams, R. J. (2017). *Implementation of large-scale education assessments*. Wiley.

Maehler, D. B., & Rammstedt, B. (Eds.). (2020). *Large-Scale Cognitive Assessment: Analyzing PIAAC Data*. Springer International Publishing.

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

Martin, M. O., von Davier, M., & Mullis, I. V. S. (Eds.). (2020). *Methods and procedures: TIMSS 2019 Technical Report*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/methods

Mazzeo, J., Lazer, S., & Zieky, M. J. (2006). Monitoring educational progress with group-score assessments. In R. L. Brennan (Ed.), *Educational measurement* (pp. 681–699). Praeger.

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

Mirazchiyski, P. V. (2021). RALSA: The R analyzer for large-scale assessments. *Large-Scale Assessments in Education, 9*(21), 1–24. https://doi.org/10.1186/s40536-021-00114-4

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*, 177–181.

Mislevy, R. J. (1993). Should "multiple imputations" be treated as "multiple indicators"? *Psychometrika, 58*(1), 79–85.

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*(2), 133–161.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–176.

NCES. (2024). NAEP Technical Documentation on the Web. Retrieved from https://nces.ed.gov/nationsreportcard/tdw

Nilsen, T., Stancel-Piątak, A., & Gustafsson, J. E. (Eds.). (2022). *International Handbook of Comparative Large-Scale Studies in Education: Perspectives, Methods and Findings*. Springer Nature.

OECD (2019a). Technical Report of the Survey of Adult Skills (PIAAC) (3rd edition). Paris, OECD. Retrieved from https://www.oecd.org/skills/piaac/publications/PIAAC_Technical_Report_2019.pdf

OECD (2019b). *The use of test scores in secondary analysis: A dialogue between data users and data producers*. http://www.oecd.org/skills/piaac/The_use_of_test_scores_in_secondary_analysis_14_June_2019_Concept_Note.pdf

OECD. (2023). PISA 2022 technical report. Paris, OECD. Retrieved from https://www.oecd.org/pisa/data/pisa2022technicalreport/

Rässler, S., Rubin, D. B., & Zell, E. R. (2013). Imputation. *Wiley Interdisciplinary Reviews: Computational Statistics, 5*(1), 20–29.

Rhee, S. H., Friedman, N. P., Boeldt, D. L., Corley, R. P., Hewitt, J. K., Knafo, A., Lahey, B. B., Robinson, J. A., Van Hulle, C. A., Waldman, I. D., Young, S. E., & Zahn-Waxler, C. (2013). Early concern and disregard for others as predictors of antisocial behavior. *Journal of Child Psychology and Psychiatry, 54*, 157–166.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592.

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

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

Rust, K. (2013). Sampling, weighting, and variance estimation in international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 117–153). CRC Press. https://doi.org/10.1201/b16061

Rust, K. F., & Johnson, E. G. (1992). Sampling and weighting in the national assessment. *Journal of Educational and Behavioral Statistics, 17*(2), 111–129.

Rutkowksi, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher, 39*(2), 142–151. https://doi.org/10.3102/0013189X10363170

Rutkowski, L., von Davier, M., & Rutkowski, D. (Eds.). (2013). *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. Chapman & Hall/CRC Press.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177.

Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods, 6*(4), 317.

Stadler, M., Greiff, S., & Krolak-Schwerdt, S. (2016). Current methodological issues in educational large-scale assessments. Guest editorial. *Psychological Test and Assessment Modeling, 58*, 593–595.

Stadler, M., Greiff, S., & Krolak-Schwerdt, S. (2017). Editorial to the special issue current methodological issues in educational large-scale assessments. Part 2. *Psychological Test and Assessment Modeling, 59*, 31–33.

Stein, C. (1956). Inadmissibility of the usual estimator of the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium of Mathematical Statistics and Probability, Vol. 1* (pp. 197–206). University of California Press.

Thao, L. T. P., & Geskus, R. (2019). A comparison of model selection methods for prediction in the presence of multiply imputed data. *Biometrical Journal, 61*(2), 343–356.

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*, 9–36.

von Davier, M., Mullis, I. V. S., Fishbein, B., & Foy, P. (Eds.). (2023). *Methods and Procedures: PIRLS 2021 Technical Report*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://pirls2021.org/methods

von Davier, M., & Sinharay, S. (2013). Analytics in international large-scale assessments: Item response theory and population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook international large-scale assessment: Background, technical issues, and methods of data analysis*. 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* (pp. 1039–1055). North Holland-Elsevier.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*(2–3), 114–128.

Yamamoto, K., & Mazzeo, J. (1992). Chapter 4: Item response theory scale linking in NAEP. *Journal of Educational Statistics, 17*(2), 155–173.

Zhao, Y., & Long, Q. (2017). Variable selection in the presence of missing data: Imputation-based methods. *Wiley Interdisciplinary Reviews: Computational Statistics, 9*(5), e1402.

Publisher's Note