Article

Journal of Educational and Behavioral Statistics
2016, Vol. 41, No. 1, pp. 57–80
DOI: 10.3102/1076998615622221

On Matrix Sampling and Imputation of Context Questionnaires With Implications for the Generation of Plausible Values in Large-Scale Assessments

David Kaplan
University of Wisconsin–Madison

Dan Su
University of Wisconsin–Madison

This article presents findings on the consequences of matrix sampling of context questionnaires for the generation of plausible values in large-scale assessments. Three studies are conducted. Study 1 uses data from PISA 2012 to examine several different forms of missing data imputation within the chained equations framework: predictive mean matching, Bayesian linear regression, and proportional odds logistic regression. We find that predictive mean matching accurately reproduces the marginal distributions of the missing context questionnaire data due to matrix sampling. Study 2 uses data from PISA 2006 to examine the consequences of imputing context questionnaire data in terms of the generation of plausible values. We find that imputing context questionnaire data with predictive mean matching and using the imputed data to produce the plausible values yields very close approximation of the original marginal distributions but leads to underestimation of the correlation structure of the context questionnaire items. Study 3 examines imputation and plausible values generation within a partially balanced incomplete block design. We find that imputation within this design accurately reproduces the original marginal distributions and retains the correlation structure of the data. Implications for context questionnaire development are discussed.

Keywords: *matrix sampling; large-scale assessment; imputation*

A common concern facing most national and international large-scale assessments is the desire to present as much content as possible without overburdening the participants in the survey. For large-scale assessments that test the so-called "cognitive" outcomes, such as the National Assessment of Educational Progress (NAEP), the Program for International Student Assessment (PISA), the Program for the International Assessment of Adult Competencies (PIAAC), the Trends in International Mathematics and Science Study (TIMSS), and the Progress in Reading and Literacy Study (PIRLS), the method for increasing cognitive content is through the implementation of *multiple matrix sampling designs*. A classic study of multiple matrix sampling designs can be found in Shoemaker (1973) who provided procedural guidelines and computational formulas for a variety of matrix sampling designs. More recently, Frey, Hartig, and Rupp (2009) provided a didactic discussion of matrix sampling designs carefully outlining theoretical and practical implications for a variety of different designs. Gonzalez and Rutkowski (2010) also outlined a variety of matrix sampling designs and showed the impact of these designs on item and person parameter recovery in a simulation study of a large-scale assessment.

In addition to the cognitive assessments, policy makers and researchers alike have begun to focus attention on the context questionnaires (CQs) of large-scale assessments. CQs provide important exogenous and mediating variables for models predicting cognitive outcomes, and these variables have become important outcomes in their own right—often referred to as "noncognitive outcomes." Given that policy priorities still focus largely on the cognitive side of the assessment, there are only three ways to expand the content of the CQ: (a) increase the assessment time for the CQ, (b) provide "menus" of optional questionnaires and let countries decide which questionnaires they wish to administer, or (c) matrix sample the CQ in a manner similar to the design of the cognitive instrument. At present, option (a) is simply not feasible. Option (b) is still possible, but it would make cross-country comparisons of noncognitive outcomes difficult. The technical consequences of option (c) form the focus of this article.

The purpose of this article is to examine the implications of matrix sampling of the CQ with respect to the generation of the plausible values (PVs) of the cognitive assessment. Our article consists of three interrelated studies. Study 1 presents a comparison of three imputation methods using data from PISA 2012 (Organization for Economic Cooperation and Development [OECD], 2013). We chose PISA 2012 because this was the first cycle of PISA and the only large-scale assessment to date that implemented a matrix sampling of the CQ. The goal of Study 1 is to provide empirical evidence of the differences among imputation methodologies and to choose an imputation methodology for the remaining studies.

In Study 2, we use data from PISA 2006 (OECD, 2006) to recreate the matrix sampling design used in PISA 2012 and to study the impact of matrix sampling and imputation of the CQ with respect to the generation of the PVs. Study 2 provides a partial replication of a recent study by Adams, Lietz, and Berezner (2013) described in more detail below.

In Study 3 using complete data from PISA 2006, we present an alternative matrix sampling design to the one used in PISA 2012 to examine how it might preserve the correlation structure among the CQ items and the PVs.

The organization of this article is as follows. In the next section, we describe the imputation methodology to be used in this article. This is followed by a description of how the imputation methodologies will be validated. After this, we describe the methods and results of Studies 1, 2, and 3, respectively. This article closes with a discussion of additional research that needs to be conducted as well as operational considerations that need to be considered when deciding to implement CQ rotation.

Imputation Methodology

For this article, we will concentrate on issues related to the imputation of the CQ. In practice, this would then be followed by the use of the conditioning model for estimating latent proficiency distributions based on the fully imputed data (see Mislevy, 1991; Mislevy, Beaton, Kaplan, & Sheehan, 1992; von Davier, 2013). For imputation of the rotated CQ, a choice must be made regarding the imputation algorithm that would be used to fill in the missing data, and it is important to note that not all imputation algorithms provide the same results. As a consequence, validity criteria must be established.

We will describe three approaches within the fully conditional specification (chained equations) framework of missing data theory (van Buuren, 2012). These approaches follow a general Bayesian framework for imputation based on the fundamental work of Rubin (1987). Specifically, we will example *predictive mean matching* for all items and then *Bayesian linear regression* under the normal model for continuous items and *proportional odds logistic regression* for categorical items.

Chained Equations

The chained equations approach uses a univariate regression model consistent with the scale of the variable with missing data to provide predicted values of the missing data given the observed data. Once a variable of interest is filled in, that variable, along with the items for which there is complete data, is used in a sequence to fill in another variable. Once the sequence is completed for all items with missing data, the posterior distribution of the regression parameters is obtained via Gibbs sampling, and the process is started again. The algorithm can run these sequences simultaneously *m* number of times obtaining *m* imputed data sets. This is the method used in the R program "mice" (van Buuren & Groothuis-Oudshoorn, 2010), which we will use for our analyses below.

Predictive Mean Matching

Let X_{obs} be the predictors with observed data and let X_{miss} be the predictors with missing data on the target variable y.

1. Obtain $\hat{\beta}$ based on X_{obs} and let $\tilde{\sigma}^2$ be a draw based on the deviations $(y_{\text{obs}} - X_{\text{obs}}\hat{\beta})'(y_{\text{obs}} - X_{\text{obs}}\hat{\beta})/\tilde{g}$, where $\tilde{g}$ is a draw from a χ^2 distribution.
2. Draw $\tilde{\beta} = \hat{\beta} + \tilde{\sigma}\tilde{z}_1 V^{1/2}$, where $V^{1/2}$ is the square root of the Cholesky decomposition of the cross-product matrix $S = X'_{\text{obs}}X_{\text{obs}}$, and z_1 is p-dimensional vector of $N(0, 1)$ random variates.
3. Calculate $\tilde{\eta}(i,j) = |X_{\text{obs},[i]}\hat{\beta} - X_{\text{miss},[j]}\tilde{\beta}|$, where $i = 1, 2, \ldots, n_1$ and $j = 1, 2, \ldots, n_0$.

4. Construct n_0 sets W_j containing d candidate donors from y_{obs} such that $\sum_d \tilde{\eta}(i,j)$ is minimum. Break ties randomly.
5. Randomly draw one donor i_j from W_j for $j = 1, 2, \ldots, n_0$.
6. Impute $\tilde{y}_j = y_{i_j}$, for $j = 1, 2, \ldots, n_0$.

Bayesian Regression Imputation

Bayesian imputation under the normal model proceeds much like predictive mean matching.

1. Obtain $\hat{\beta}$ based on X_{obs} and let $\tilde{\sigma}^2$ be a draw based on the deviations $(y_{\text{obs}} - X_{\text{obs}}\hat{\beta})'(y_{\text{obs}} - X_{\text{obs}}\hat{\beta})/\tilde{g}$, where $\tilde{g}$ is a draw from a χ^2 distribution.
2. Draw $\tilde{\beta} = \hat{\beta} + \tilde{\sigma}\tilde{z}_1 V^{1/2}$ as before.
3. Calculate the imputed value $\tilde{y}$ as $\tilde{y} = X_{\text{miss}}\tilde{\beta} + \tilde{z}_2\tilde{\sigma}$, where z_2 is a $j = 1, 2, \ldots, n_0$ vector of $N(0, 1)$ random variates for those with missing data on y.
4. A new $\tilde{y}$ is obtained by drawing a new $\tilde{\sigma}^2$. This can be repeated m times.

Proportional Odds Logistic Regression

For PISA, most of the items in the CQ are ordered categorical (e.g., Likert scales). We wish to take into account the correct probability model, and so for ordered categorical items, we use the *proportional odds logistic regression model*. The proportional odds logistic regression approach is similar to Bayesian imputation except the following:

1. obtain $\hat{\beta}$ by iteratively reweighted least squares,
2. obtain $\tilde{p} = 1/\left(1 + \exp(-X_{\text{miss}}\tilde{\beta})\right)$, and
3. use the cumulative logit to assign K categorical responses:

$$\log\left[\frac{p(\tilde{y}_j \leq k)}{p(\tilde{y}_j > k)}\right] = \log\left[\frac{p(\tilde{y}_j \leq k)}{1 - p(\tilde{y}_j \leq k)}\right] = \log\left[\frac{\tilde{p}_{1,j} + \ldots + \tilde{p}_{k,j}}{\tilde{p}_{k+1,j} + \ldots + \tilde{p}_{K,j}}\right].$$

Validating Imputation Methods

For any rigorous study of missing data imputation, validity criteria must also be established. For this article, we follow the work of Rässler (2002) on levels of validation for imputation procedures. We discuss two of these levels—preserving marginal distributions and preserving correlation/covariance structure—that are directly relevant to concerns about matrix sampling and PV generation.[1] These levels of validity (described next) have also been examined in the context of statistical matching of two large-scale education surveys—PISA and the OECD Teaching and Learning International Survey—TALIS (Kaplan & McCarty, 2013; OECD, 2009a).

Preserving Marginal Distributions

The lowest level of validity and a minimum requirement for statistical matching is that the marginal distributions of the individual variables in the original surveys be preserved after imputation. Because the omitted responses in the rotated part of the CQ are missing completely at random (MCAR), we can compare the marginal distributions of the imputed responses to the marginal distribution of the observed responses. In this article, we examine the preservation of the marginal distributions of the CQ scales in Study 1. In Studies 2 and 3, we examine the impact of the imputation of the CQ on preserving the marginal distributions of the PVs.

Preserving Covariance/Correlation Structures

In Rässler's work, preserving the covariance/correlation structure was focused on data fusion problems. In this article, we will examine the preservation of the covariance/correlation structure among the variables in a different way. Specifically, we compute the pairwise correlations among items within each rotated form. Due to the random assignment of questionnaire forms to students, we expect the correlation to be similar across forms. Study 3 will examine an alternative matrix sampling design in terms of how it preserves correlation structure among the scales.

Study 1: Imputation Using the PISA 2012 Assessment Design

The 2012 cycle of PISA (OECD, 2014) implemented a matrix sampling design for the CQ. The design consisted of three forms. The forms contained a systematic combination of clusters.[2] Each form contained a common part and a rotated part. Each form contained two rotated clusters of questions so that there was an overlapping rotation cluster between any two of the rotated forms. Table 1 shows the PISA 2012 CQ matrix sampling design. For detailed information on the partition of questionnaire items into clusters, please refer to Table 6.4 in the PISA 2012 assessment and analytical framework (OECD, 2013, p. 194).

The PISA 2012 CQ design resulted in students missing one third of CQ items. Thus, without listwise deletion it would not be possible to jointly analyze the data from the questionnaires. Because the forms were randomly assigned to students, the missing data due to the design have the mechanism of MCAR. However, when the missing data are not MCAR (e.g., the random assignment was not perfectly implemented), results from listwise deletion may be biased. Instead, multiple imputation (MI) can produce unbiased results under MCAR or missing at random (MAR) (Enders, 2010; Little & Rubin, 2002; Schafer, 1997).

In this study, we compare the three imputation methods described in the previous section: predictive mean matching (*pmm*), Bayesian imputation under the normal model (*norm*), and proportional odds logistic regression (*polr*) model

TABLE 1.
PISA 2012 CQ Matrix Sampling Design

Form A	Form B	Form C
Common part	Common part	Common part
Cluster 1	Cluster 1	Cluster 1 missing
Cluster 2	Cluster 2 missing	Cluster 2
Cluster 3 missing	Cluster 3	Cluster 3

Note. CQ = context questionnaire.

using U.S. data from PISA 2012. In order to impute all the questionnaire data which contain both numeric items and categorical items, we use either *pmm* alone or use the combination of *norm* and *polr*.

We implemented two imputation procedures. In the first procedure, we imputed all the questionnaire items using *pmm*. The items in the imputation model are school ID, questionnaire form ID, and all the questionnaire items including common part and clusters. In the second imputation procedure, we imputed numeric items using *norm* and ordered categorical items using *polr*. However, due to the limitation that *polr* is not able to impute a large amount of categorical items (e.g., more than 70 items), especially with more than four levels in each categorical item, we had to partition the items in the scales into three sets which were imputed separately. The three sets of items correspond to Cluster 1, Set 2, and Set 3 in the PISA 2012 CQ rotation design. We partition the items into these three clusters because the number of items in each cluster is not too large for *polr*, and by design each set of questions with similar themes can be imputed together. The imputation models for each cluster contain questionnaire form ID and the questionnaire items in the corresponding question set. We had to exclude the school ID variable from the imputation model because it contains 162 levels which are too many for *polr* to impute with other items together. We conducted both a single imputation and five MIs for the two procedures. All the missing data which includes missing by design as well as item missing data were imputed using the "mice" package (van Buuren & Groothuis-Oudshoorn, 2010) in R (R Core Team, 2014). The mice package contains, among many other imputation methods, *pmm*, *norm*, and *polr*. All R codes for Studies 1 through 3 are available at http://bise.wceruw.org/publications.html.

Results of Study 1

Recall that Rässler's (2002) minimum requirement for imputation validity is preserving the observed marginal distributions. Thus, in order to assess how well the three methods reproduce the originally observed marginal distributions of the CQ items, we compare the item densities based on the imputed values with the densities based on the original values. Under the assumption of MCAR, we expect the densities based on the imputed values to be very close to the densities based on the originally observed values. To present the density plots, we chose 4 items that are representative of items that come from each of the three question sets and of different types of items. Similar results hold for the remaining items. The Item ST48Q01 from the construct math intentions (in Question Set 1) is a binary item. The Item ST57Q01 from out-of-school study time (in Question Set 2) is a numeric variable. The Item ST37Q01 from math self-efficacy (in Question Set 1) and the Item ST80Q01 from cognitive activation (in Question Set 3) are ordered categorical items with four levels.

Figures 1 and 2 display the kernel density plots of the 4 items using *pmm* for single imputation and five imputations, respectively. Note that the plots for the categorical variables are smoothed version of histograms. The coding of the discrete values is noted under the figures. Figures 3 and 4 show the kernel density plots of the 4 items using *polr* or *norm* for single imputation and five imputations. Comparing Figures 1 and 2 with Figures 3 and 4, all the densities of imputed values using *pmm* are closer to the densities of the observed values compared to using *polr* and *norm*. For the numeric variable, out-of-school study time, *norm* did not reproduce the observed marginal density which is highly right skewed. Additionally, the method *norm* can generate values outside the data range. For this analysis, we used the "squeeze" function in the "mice" package to ensure positive imputations for out-of-school study time. Regardless, we conjecture that the result for *norm* is likely due to the fact that *norm* assumes a normal model, and out-of-school study time is likely nonnormal in the population.

For assessing how well the methods preserve the correlations, Rässler's (2002), third level of validity, we compute the pairwise correlations among the 3 items ST37Q01 (in Question Set 1), ST57Q01 (in Question Set 2), and ST80Q01 (in Question Set 3) within each rotated form. Due to the random assignment of forms to students, we expect the correlation between the 2 items to be similar across forms. For example, Form A has both Items ST37Q01 and ST57Q01, so we are able to compute the observed correlation between the 2 items within Form A. However, in Form B, Item ST37Q01 is not observed, so we compute the correlation of the imputed values of Item ST37Q01 with the observed values of Item ST57Q01 within Form B. The same with Form C, Item ST57Q01 is not observed, so we compute the correlation of the imputed values of Item ST57Q01 with the observed values of Item ST37Q01 within form C. Table 2 shows the pairwise correlations among the 3 items using *pmm* with single imputation and with five imputations.

Compared to the observed correlations, all of the correlations that involve imputed values are attenuated; however, *pmm* preserved the direction of all the correlations. With five imputations, the correlations are only slightly better preserved than using a single imputation. Table 3 shows the pairwise correlations among the 3 items using *polr* and *norm* with a single imputation and with five

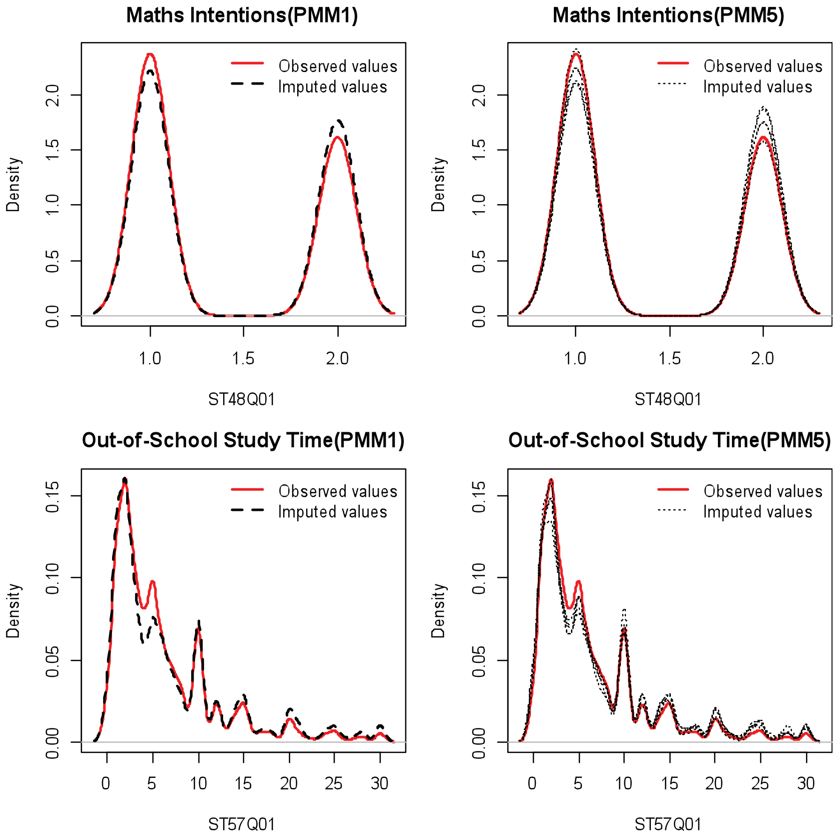


FIGURE 1. *Kernel density plots of imputed and observed values for Items ST48Q01 (1 = courses after school are maths and 2 = courses after school are test language) and ST57Q01 using* predictive mean matching *(*pmm*) under single imputation and five imputations.*

imputations. Here, we can see that all the correlations are strongly attenuated toward zero. The directions of the correlations are not preserved, and the results from five imputations are not an improvement over using a single imputation.

The results of Study 1 suggest that *pmm* does a better job of preserving the marginal densities and correlations than *polr* and *norm*. In terms of operational limitations, *pmm* is more flexible because it can be applied to all types of items and it is able to impute missing data on many items at once. The *polr* method is restricted by the total number of levels of all the categorical items in the imputation model, while *norm* assumes normality, has restrictions on the distribution of numeric scales, and requires postprocessing of the imputed values to ensure they stay in the plausible range. All three procedures do an inadequate job of

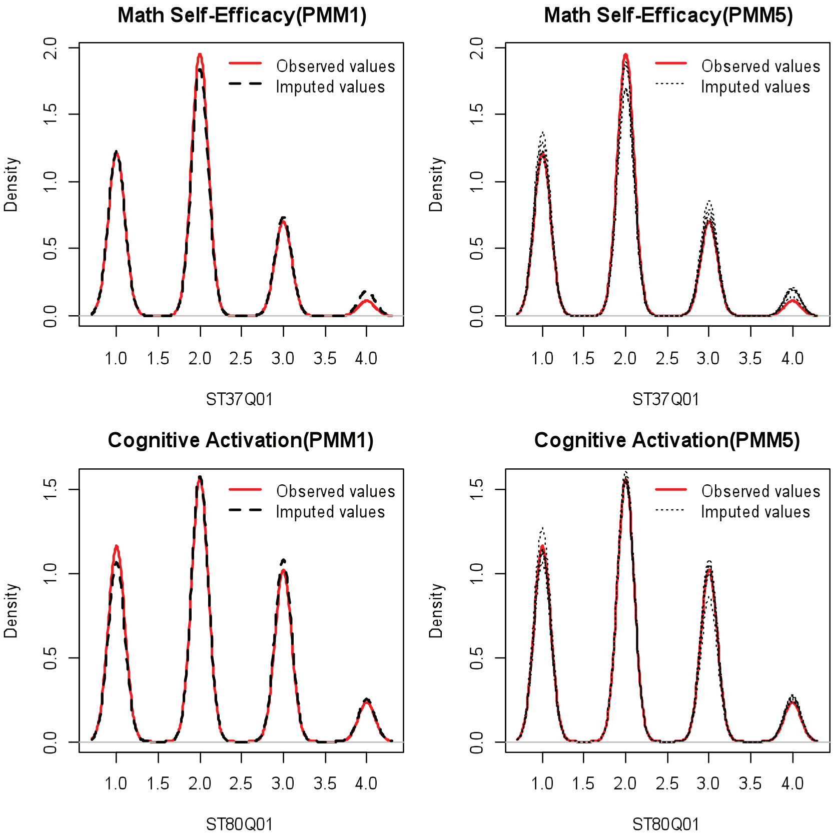


FIGURE 2. *Kernel density plots of imputed and observed values of Items ST37Q01 (1 =* very confident *to 4 =* not at all confident*) and ST80Q01 (1 =* always or almost always *to 4 =* never or rarely*) using* predictive mean matching *(*pmm*) under the single imputation and five imputations.*

reproducing observed correlations. For these reasons, we chose *pmm* as our imputation method in the following studies.

Study 2: Simulation of Matrix Sampling and Imputation Using PISA 2006

In Study 2, we focus our attention on the main issue in this article—namely the consequences of CQ rotation and imputation on PV generation. We study this problem under three different design conditions using the U.S. data from PISA 2006 (OECD, 2006). The first design condition uses the original questionnaire without rotation and generates the PVs according to the same procedures described in the PISA 2006 technical report (OECD, 2009).

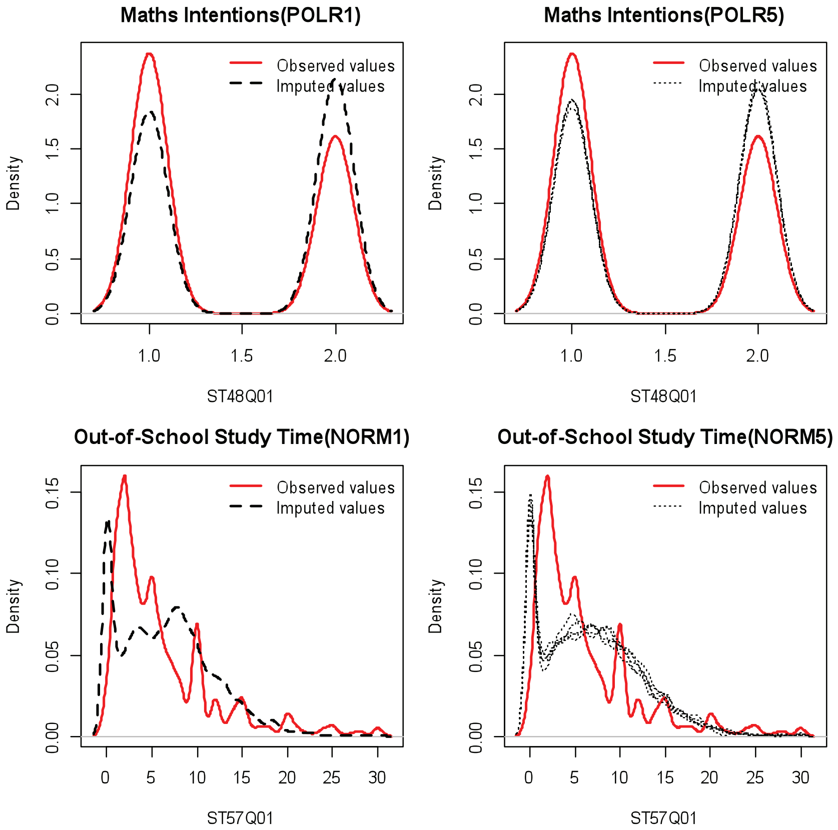


FIGURE 3. *Kernel density plots of imputed and observed values of Items ST80Q01 (1 =* always or almost always *to 4 =* never or rarely*) and ST57Q01 using* polr *or* norm *under single imputation and five imputations.*

The second design condition generates PVs by conditioning on the matrix sampled questionnaire without imputation. This approach replicates the Adams et al. (2013) paper using a joint conditioning approach with two questionnaire forms. In this design, three mutually exclusive clusters of scales in the questionnaire are created. The first cluster, the common part, was assigned to both questionnaire forms. The remaining clusters are assigned to each of the questionnaire forms, respectively. Thus, each of the questionnaire forms contains the common part of the scales and one of the two rotated clusters. The scales are allocated to the clusters according to the principle that the average correlation between the scales from Set 1 with the science performance is similar with the average correlation between the scales from Set 2 with the science performance. PISA 2012 also implemented a similar rotation design but with three clusters instead of two.

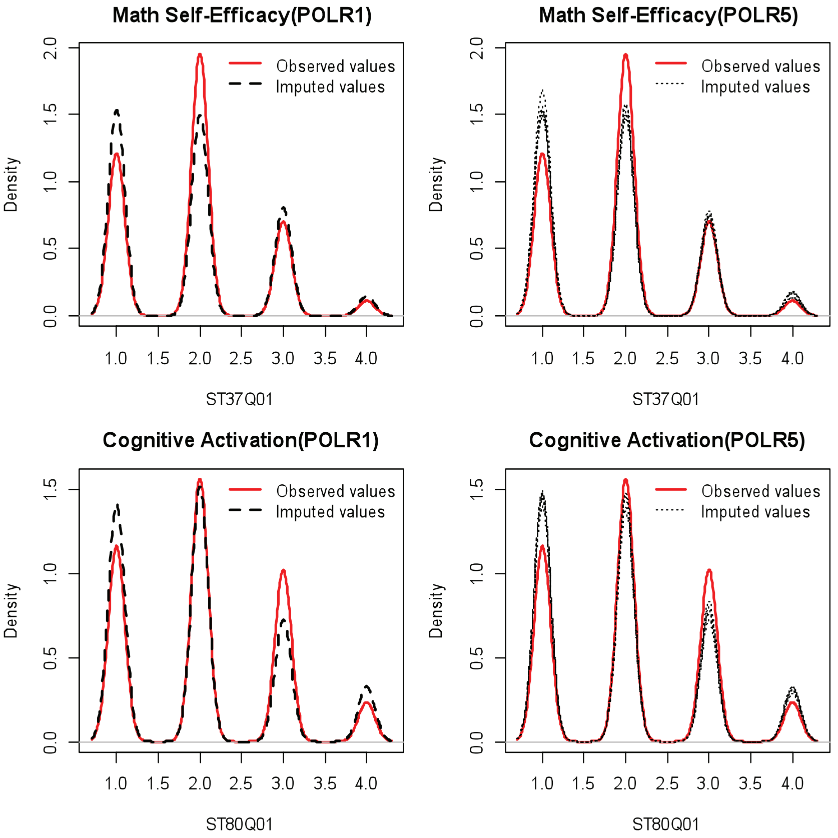


FIGURE 4. *Kernel density plots of imputed and observed values of Items ST37Q01 (1 =* very confident *to 4 =* not at all confident*) and ST80Q01 (1 =* always or almost always *to 4 =* never or rarely*) using* polr *under single imputation and five imputations.*

The third design condition uses a matrix sampled questionnaire with *pmm* imputation and generates PVs conditioned on the matrix sampled questionnaire with all missing data in the questionnaire imputed. This matrix sample design is the same as Adams et al. (2013) described above. All the missing data are imputed using *pmm* comparing the results of a single imputation to five MIs. Note that in contrast to Study 1, in order to replicate the Adams et al. rotation design, imputation is conducted on the scale level.

Procedure

In this section, we describe the procedures for the approach of rotation alone and the approach of rotation with imputation. The procedures for scaling

TABLE 2.
Pairwise Correlations Among Items ST37Q01, ST57Q01, and ST80Q01 Within Each Booklet Using pmm *Under the Single Imputation and Five Imputations*

	pmm 1			*pmm* 5		
Correlations	Form A	Form B	Form C	Form A	Form B	Form C
ST37Q01 with ST57Q01	**−0.21**	−0.12	−0.11	**−0.21**	−0.14	−0.13
ST37Q01 with ST80Q01	0.02	**0.15**	0.13	0.08	**0.15**	0.11
ST57Q01 with ST80Q01	−0.03	−0.04	**−0.05**	−0.02	−0.03	**−0.05**

Note. Boldface numbers are the correlations between the 2 observed items within the same form. The correlations under *pmm* 5 are the averaged correlations across five imputations. *pmm = predictive mean matching*.

TABLE 3.
Pairwise Correlations Among Items ST37Q01, ST57Q01, and ST80Q01 Within Each Form Using polr *and* norm *Under the Single Imputation and Five Imputations*

	polr and *norm* 1			*polr* and *norm* 5		
Correlations	Form A	Form B	Form C	Form A	Form B	Form C
ST37Q01 with ST57Q01	**−0.21**	−0.03	0.02	**−0.21**	−0.01	0.01
ST37Q01 with ST80Q01	0.03	**0.15**	−0.01	0.01	**0.15**	0.00
ST57Q01 with ST80Q01	−0.03	−0.01	**−0.05**	−0.01	0.01	**−0.05**

Note. Boldface numbers are the correlations between the 2 observed items within the same form. The correlations under *polr* and *norm* 5 are the averaged correlations across five imputations.

cognitive data and generating PVs were done using “TAM” package (Kiefer, Robitzsch, and Wu (2014; see also Adams, Wilson, & Wu, 1997; Adams & Wu, 2007) in the R software environment (R Core Team, 2014).

Creating the rotated questionnaire. Table 4 shows the scales contained in each of the two questionnaire forms. From the table, we can see that all students have data for the common part. One half of the students who were assigned to Form A have data from Cluster 1 and the other half of students have data from Cluster 2. In order to simulate the matrix sampling design, we simulate the case that one half of the students were randomly assigned to Form A and the other half to Form B. We partition the students into two subsets by randomly sampling one half of students within each school to the first set and assign them to Form A. The rest of students are assigned to Form B. In other words, data from Set 2 are deleted for those students who receive Form A. Similarly, data from Set 1 are deleted for those students who receive Form B.

TABLE 4.
Adams et al. (2013) Matrix Sampling Design for PISA 2006 Simulation Study

Form A	Form B		
Common part			
Scale name	Scale description		
PROGN	Country study program		
GRADE	Grade		
AGE	Age of the student		
GENDER	Gender		
BMMJ	Occupation of mother		
BFMJ	Occupation of father		
BSMJ	Occupation of self at 30		
MISCEDN	Educational level of mother		
FISCED	Educational level of father		
IMMIG	Immigration status		
LANG	Language at home		
DEFFORT	Difference in effort		
CULTPOSS	Classic literature, books of poetry, works of art		
HEDRES	Study desk, quiet place to study, computer for school work, educational software, own calculator, books to help with school work, dictionary		
WEALTH	Own room, Internet link, dishwasher, DVD/VCR, three country-specific wealth items, number of cellphones, TVs, computers, cars		
Cluster 1		Cluster 2	
CARINFO	Student information on science-related careers	ENVOPT	Environmental optimism
CARPREP	School preparation for science-related careers	ENVPERC	Perception of environmental issues
ENVAWARE	Awareness of environmental issues	GENSCIE	General value of science
INSTSCIE	Instrumental motivation in science	INTSCIE	General interest in learning science
JOYSCIE	Enjoyment of science	PERSIE	Personal value of science
SCIEFUT	Future-oriented science motivation	RESPDEV	Responsibility for sustainable development
SCINTACT	Science teaching: interaction	SCAPPLY	Science teaching: focus on applications or models
SCINVEST	Science teaching: student investigations	SCHANDS	Science teaching: hands-on activities
SCSCIE	Science self-concept	SCIEACT	Science activities
		SCIEEFF	Science self-efficacy

Note. HIGHCONF, INTCONF, PRGUSE, and INTUSE were excluded from the design because of no U.S. data in these four scales.

Impute missing data in the rotated questionnaire. We impute the questionnaire data under the Adams et al. (2013) rotation design using *pmm* in the "mice" package (van Buuren & Groothuis-Oudshoorn, 2010). The items included in the imputation model are all the items in Table 4, school ID, and form ID. The complete questionnaire data after imputation is then prepared for generating the PVs as described subsequently.

Preparing the conditioning items. The conditioning items for scaling cognitive data include *direct conditioning* items and *indirect conditioning* items. The direct conditioning items are form ID (deviation coded), school ID (dummy coded), and gender (dummy coded). Indirect conditioning items include all the other items listed in Table 4. Categorical items in the indirect conditioning items are dummy coded. For the rotation with imputation condition, after the imputation procedure we have the complete questionnaire data which can be directly prepared as conditioning items. However for the approach of rotation alone, in order to deal with the missing values in the questionnaire, missing indicators are created for all the items and missing values in continuous items are replaced with means.

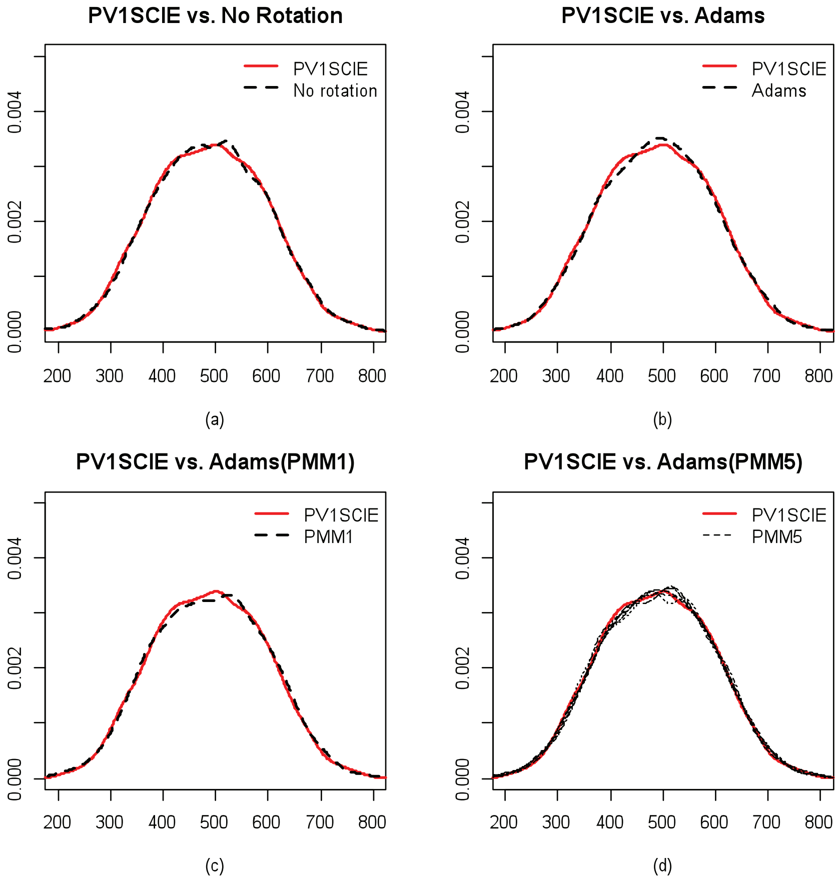
In order to reduce the dimensionality of the conditioning items, a principal components analysis is performed on the indirect conditioning items. The components that account for 95% of the total variance in the indirect conditioning items are then used together with the direct conditioning items for scaling.

Scaling the cognitive data and drawing PVs. The item response model for scaling the cognitive data is the mixed coefficients multinomial logit model (Adams & Wu, 2007). A five-dimensional scaling model which is composed of one reading, one science, one mathematics, and two attitudinal dimensions was used in this simulation study. Due to the fact that U.S. data do not contain the reading assessment, we implemented a four-dimensional scaling model. In the "TAM" package, we apply an option that yields a multidimensional one parameter partial credit model with ConQuest parametrization (Adams, Wu, & Wilson, 2015).

Before running the partial credit model, we first fixed the item parameters at the international values given in Appendix 1 of PISA 2006 technical report (OECD, 2009). Second, we specified the loading structure of items on dimensions. Third, we fixed to zero the regression coefficients between mathematics performance and those form contrasts that yield forms without a mathematics test. Finally, we specified a matrix of covariates for the latent regression conditioning model. The IRT models are the same for the three approaches except for the conditioning items. We use the "TAM" program to draw five normally approximated PVs for each of the four dimensions.

Results of Study 2

In Figure 5a, we display the kernel density plot of the first science PV under the no rotation design condition and the density of the first PV (PV1SCIE) in the



(a) (b) (c) (d)

FIGURE 5. *Kernel density plots of the first plausible value under the three approaches compared to the density of the first plausible values in the original data.*

original questionnaire data. We see that the two densities almost overlap completely. This plot validates our procedure of generating PVs using the "TAM" package. We then compare the PVs under the two other approaches with the observed PV1SCIE. Figure 5b shows the comparison of the density plots of the first PV under the Adams et al. (2013) rotation design. Figure 5c and d displays the comparisons of the kernel density plots of the first PV under the approach of rotation with single imputation and five imputations, respectively. We can see that the densities of the first PV under all three approaches are very close to the original PV1SCIE. Table 5 provides the descriptive statistics for the densities shown in the plots. The results show that the PVs generated under the three approaches replicate PV1SCIE very well. Thus, we found no evidence of bias in the PVs using rotation alone or with imputation for U.S. data.

TABLE 5.
Descriptive Statistics on the First Plausible Value of the Science Dimension Generated Under Three Conditions

	Mean	*SD*	10%	25%	75%	90%
No rotation	489.16	106.13	351.21	413.23	565.52	626.39
Adams et al. (2013) rotation	488.97	106.26	351.20	412.70	564.19	626.81
Adams et al. (2013) rotation (*pmm* 1)	489.67	107.43	350.74	410.56	567.69	628.89
Adams et al. (2013) rotation (*pmm* 1_5)	489.68	107.00	350.23	412.73	566.48	629.04
Adams et al. (2013) rotation (*pmm* 2_5)	489.05	108.04	348.54	409.68	566.65	628.56
Adams et al. (2013) rotation (*pmm* 3_5)	488.60	107.27	348.53	410.64	565.91	628.29
Adams et al. (2013) rotation (*pmm* 4_5)	489.43	107.33	349.29	412.45	565.65	627.44
Adams et al. (2013) rotation (*pmm* 5_5)	489.57	107.02	353.08	410.95	565.28	628.41

Note. pmm 1_5 is the first imputation of the five imputations using *pmm*. *pmm* = *predictive mean matching*.

For the approach of matrix sampling with imputation, it is important that the questionnaire data maintain the original covariance structure after imputation. Thus, we calculate the pairwise correlations among the rotated sets of items after single imputation using *pmm* and compare them with the correlations calculated using the original data (not matrix sampled). We also compared the pairwise correlations between the imputed items and original items with the first PVs on science. The horizontal axis of Figure 6 displays the distribution of the difference between the correlations after imputation and the original correlations. We see that the correlations after imputation are not well preserved. We find that 40% of the correlations (in total 190 correlations) have more than 0.1 absolute difference compared to the original correlations. Nine correlations are even more extreme, exceeding 0.5 absolute difference. Although we did not find a pattern indicating why the correlations among some items are more biased than others, the amount and magnitude of the correlation difference deserve our attention. Furthermore, 85% of the correlations have negative difference indicating that the correlations generated under the Adams et al. (2013) matrix sampling design with imputation are strongly attenuated; the number of negative correlation differences is far more than the number of positive correlation differences. This result is consistent with the attenuation of the correlations in Study 1. Finally, for the correlations between the PVs and items, 5 correlations out of 19 exceed 0.1 absolute difference. Of 19 correlations, 14 have negative difference. In the following study, we suggest

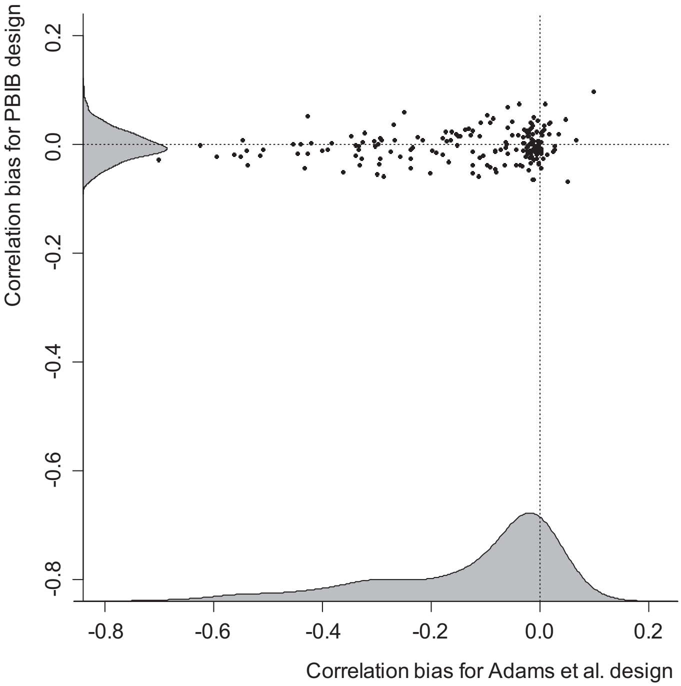


FIGURE 6. *Comparison of bias in correlations between the Adams et al. (2013) design and partially balanced incomplete block matrix sampling design (PBIB) design.*

a different matrix sampling design that may help better preserve the correlation structure among the items and PVs.[3]

Study 3: Alternative Matrix Sampling Design Using PISA 2006

In this simulation study, we examine the properties of a partially balanced incomplete block matrix sampling design (PBIB) of the CQ and its impact on PV generation. In this design, we keep the common part of questionnaire items the same as in the Adams et al. (2013) design. However, here we arrange the 19 scales according to a partially balanced incomplete block design with three associate classes (Montgomery, 2012). The distribution of the associate classes for each pair of scales is shown in Table 6. For example, we find that Scales 1 and 3 appear together 3 times, Scales 1 and 2 appear together 4 times, and Scales 1 and 7 appear together 5 times. The design (without the common part) is shown in Table 7, where we see that the 19 scales (excluding the common part) are arranged in 19 clusters.

TABLE 6.
Partially Balanced Incomplete Block Design With 3, 4, and 5 Associate Classes

	Scales									Scales									
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	9	4	3	4	4	4	5	3	4	4	3	4	4	4	4	5	3	5	5
2	4	9	3	5	4	4	4	4	4	4	3	4	4	4	5	4	4	4	4
3	3	3	9	3	4	4	4	5	5	3	5	5	4	4	4	4	4	4	4
4	4	5	3	9	4	4	3	4	4	4	4	4	4	4	5	4	4	4	4
5	4	4	4	4	9	5	4	4	4	4	4	4	4	4	4	4	4	4	3
6	4	4	4	4	5	9	4	4	4	4	4	4	4	3	4	3	4	4	5
7	5	4	4	3	4	4	9	4	4	4	4	3	5	4	4	4	4	4	4
8	3	4	5	4	4	4	4	9	4	5	5	4	4	4	3	4	4	4	3
9	4	4	5	4	4	4	4	4	9	4	4	3	4	4	4	4	4	4	4
10	4	4	3	4	4	4	4	5	4	9	4	5	4	4	4	4	4	4	3
11	3	3	5	4	4	4	4	5	4	4	9	4	4	4	4	4	4	4	4
12	4	4	5	4	4	4	3	4	3	5	4	9	4	4	4	4	4	4	4
13	4	4	4	4	4	4	5	4	4	4	4	4	9	4	3	4	4	4	4
14	4	4	4	4	4	3	4	4	4	4	4	4	4	9	4	4	4	4	5
15	4	5	4	5	4	4	4	3	4	4	4	4	3	4	9	4	4	4	4
16	5	4	4	4	4	3	4	4	4	4	4	4	4	4	4	9	4	4	4
17	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	9	4	5
18	5	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	9	3
19	5	4	4	4	3	5	4		4	3	4	4	4	5	4	4	5	3	9

In order to simulate the partially balanced incomplete design with U.S. data from PISA 2006, we randomly assign the clusters to the students and delete the data that students should not have due to the design shown in Table 6. Then we impute the questionnaire data using *pmm* with single imputation and MIs. The items included in the imputation model are school ID, block ID, the common items, and the 19 rotated items. We then scale the cognitive items and draw PVs as described in Study 2.

Results of Study 3

Figure 7 compares the kernel density plot of the first PV with one and five imputations under the PBIB rotation design to the density of the first PVs PV1SCIE in the data set. The overlapping densities suggest that there is no evidence that the PBIB design produces bias when generating the PVs using the U.S. data. The vertical axis of Figure 6 shows the distribution of the differences in the correlations under the PBIB design. We note that the distribution of the differences is much narrower compared to the distribution of the differences under the Adams et al. design. We find that among the items themselves, no correlation exceeds 0.1 in absolute difference compared to the original correlations.

TABLE 7.
Partially Balanced Incomplete Block Design for the 19 Questionnaire Scales

Cluster ID	Scales																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	0	1	1	0	1	1	0	1	0	1	1	1	0	1	0	0	0	0	0
2	1	1	0	0	0	1	1	0	1	1	0	0	0	0	1	0	1	1	0
3	1	1	0	1	0	0	0	1	0	0	0	1	0	1	1	1	0	1	0
4	0	1	0	0	0	0	0	1	1	1	0	0	1	1	0	1	1	0	1
5	0	1	1	0	1	0	1	0	0	0	0	1	1	0	1	1	1	0	0
6	0	1	0	1	0	1	0	0	0	0	1	1	1	0	0	0	1	1	1
7	1	0	0	0	1	1	1	1	0	0	1	0	0	0	0	1	1	0	1
8	0	0	1	1	0	0	1	1	1	0	1	0	1	1	0	0	1	0	0
9	0	0	1	1	0	1	0	1	0	1	0	1	0	0	1	0	1	0	1
10	1	0	0	1	1	1	0	0	1	1	0	1	1	0	0	1	0	0	0
11	1	0	1	0	1	0	0	0	1	0	0	1	0	1	0	0	1	1	1
12	0	1	0	1	1	1	1	0	1	0	0	0	0	1	1	0	0	0	1
13	1	1	0	1	1	0	1	1	0	1	0	0	1	0	0	0	0	1	0
14	1	0	0	0	0	0	1	0	0	1	1	1	1	1	1	0	0	0	1
15	0	0	1	0	0	0	1	1	1	1	1	1	0	0	0	1	0	1	0
16	1	0	1	0	0	1	1	0	0	0	0	0	1	1	0	1	0	1	1
17	0	0	1	0	1	1	0	1	1	0	1	0	1	0	1	0	0	1	0
18	0	0	0	1	1	0	0	0	0	1	1	0	0	1	1	1	1	1	0
19	1	1	1	1	0	0	0	0	1	0	1	0	0	0	1	1	0	0	1

Note. A “1” denotes the presence of the scale in the cluster, “0”, other size.

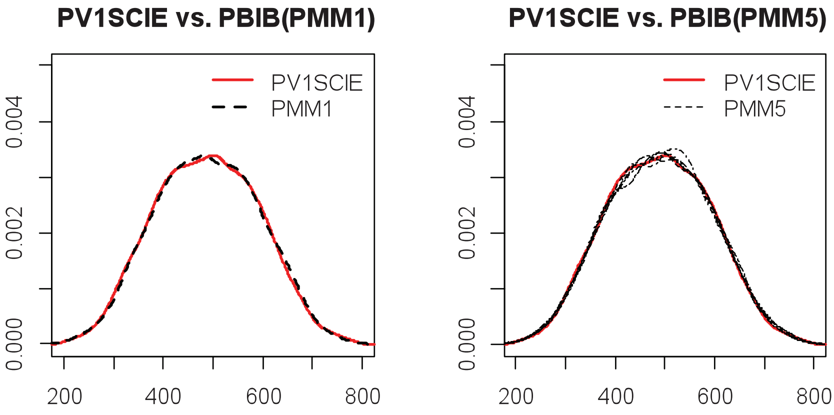


FIGURE 7. *Kernel density plots of the first plausible value under the partially balanced incomplete block matrix sampling design (PBIB) rotation design compared to the density of the first plausible value in the original data.*

Furthermore, 96% of the correlation differences are less than .05. We find significant improvement in the number and magnitude of bias in correlations among the items compared to Adams et al. (2013) rotation design. Regarding the direction of the difference, 55% correlation differences have negative signs. Because the number of negative correlation differences is close to the number of positive correlation differences, we conclude that there is no evidence of extensive attenuation in the correlations among the items after the single imputation using *pmm*.

The correlation difference between the PVs and items is smaller compared to the difference in Adams et al. (2013) rotation design. Only one correlation was observed that had a 0.1 absolute difference. However, the attenuation between the PVs and items is still an issue, since 14 of the 19 correlations are negatively biased. The results suggest that it is possible to have a matrix sampling design that preserves the marginal distributions of the PVs and the covariance structure among the items and PVs after imputation. Further studies on improving the correlations between PVs and rotated items are still needed.

Conclusions

This article presented three interrelated studies focusing on the issue of CQ rotation and its implications for the generation of PVs in large-scale assessments. Study 1 focused on the performance of three imputation algorithms and concluded that *pmm* performed best with respect to generating marginal distributions of imputed scales that match closely those of the observed data. Study 2 examined the quality of PV generation using *pmm*-imputed CQ data and served as a partial replication of Adams et al. (2013). We found that regardless of whether one uses a fully imputed CQ or the CQ without imputation, excellent recovery of the marginal distributions of the PVs is observed, at least for the U.S. data. However, we find that correlations among the imputed items are attenuated. Finally, Study 3 explored a partially balanced incomplete block design for the CQ to examine whether it could reproduce the marginal distributions and correlation structure of the data. We found that the design examined in this article was successful in reproducing the correlation structure of the data and, in addition, reproduced the marginal distributions of the PVs as expected.

The evidence presented in this article argues for seriously considering questionnaire rotation as an option for increasing contextual information in large-scale assessments. However, a set of issues, beyond the scope of this article, must still be resolved in order to fully warrant questionnaire rotation (see also von Davier, 2013). First, additional studies of imputation methods need to be examined. As noted, we found that *pmm* was fast and accurate, whereas polytomous logistic regression and Bayesian linear regression did not perform as well. However, the methods examined in this study are by no means exhaustive of the range of missing data imputation methods now readily available. In a similar example, a study of data fusion for large-scale assessments was presented in Kaplan and McCarty (2013) who examined a much large number of imputation methods. In their study, Kaplan and McCarty (2013) focused on creating a synthetic cohort of data from PISA and TALIS and presented an experimental evaluation of a representative group of data fusion methods using data from Iceland—the only OECD country that implemented both PISA and TALIS to all members of the relevant populations. On the basis of Rässler's (2002) criterion, Kaplan and McCarty (2013) found that Bayesian bootstrap predictive mean matching (Meinfelder, 2011; Rubin, 1981) and the EM-bootstrap (Honaker, King, & Blackwell, 2010) performed best with respect to creating a usable synthetic data file for research purposes.

A second issue concerns attenuation in relationships among the CQ imputed values and the PVs. This issue relates to the general advice given in the MI literature regarding the use of all available information when imputing missing data. As pointed out by Rubin (1987) and emphasized in the context of imputation in large-scale assessments by von Davier (2009, 2013), it is essential that all available information be used when imputing missing data. In the context of our article, if the PVs are not included as part of the imputation model for the CQ, then analyses involving the CQ and the PVs (such as regression models) will likely be biased. However, to address this problem would require a massive imputation of the PVs and the CQ simultaneously. Such an analysis was beyond the scope of this article and possibly not feasible from the point of view of a large-scale assessment operation. Nevertheless, we do believe that future research should examine simultaneous imputation of the PVs and CQ. We predict that with simultaneous imputation of the PVs and CQ that we will see little improvement in reproducing the marginal distributions of the PVs for purposes of policy reporting nor will we see much improvement in reproducing marginal and joint distributions among the CQ variables themselves. We predict that we will see improvement in correlations as well as model-based analyses involving the CQ and the PVs.

A third issue concerns the impact of CQ rotation in the context of small area estimation. Specifically, policy makers may be interested in obtaining reliable estimates of proficiency in cognitive domains for small geographic areas or demographic subgroups that were not directly accessed in the sampling frame. The topic of small-area estimation is beyond the scope of this article, but suffice to say that methods for small area estimation involve "borrowing strength" from similar areas with sufficient data to be used to provide predictions of performance for small demographic subgroups or areas (see e.g., Ghosh & Rao, 1994; Rao, 2004). Clearly, the use of matrix sampling alone or with imputation can have a impact on the prediction of proficiency in small areas, and this topic requires additional work before matrix sampling of the CQ can be confidently recommended. Concerns associated with the second issue raised above are relevant here as well.

Finally, it is important to examine alternative rotation designs beyond those presented in this article. The PISA 2012 rotation design provides excellent recovery of marginal distributions regardless of how the missing data are imputed, but with imputation, the correlation structure can be considerably biased. The partially balanced incomplete block design proposed in Study 3 seems to resolve this issue, but it was by no means exhaustive of the range of matrix sampling designs available.

The additional research questions raised in the previous paragraphs can, in principle, be addressed during the field trial stage of a large-scale assessment or through supplementary research supported by the organizing bodies of these large-scale assessments (e.g., OECD, International Association for the Evaluation of Educational Achievement [IEA], U.S. Department of Education). It is during the field trial stage that cognitive and noncognitive assessment items are trialed and their psychometric properties studied. We argue the field trial stage should also be used to examine different matrix sampling designs along with different imputation algorithms. In this way, a cumulative body of evidence regarding matrix sampling of CQs can be developed and discussed during the design of phase of large-scale educational assessments.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

[1] The other two levels are “preserving individual values” and “preserving joint distributions.”

[2] In this article, we use the PISA nomenclature of “forms” and “clusters.” In some large-scale assessments these are referred to as “booklets” and “blocks,” respectively.

[3] Correlation matrices from which Figure 6 is derived are available on request.

References

Adams, R. J., Lietz, P., & Berezner, A. (2013). On the use of rotated context questionnaires in conjunction with multilevel item response models. *Large-Scale Assessments in Education*, *1*, 5. Retrieved from http://www.largescaleassessmentsineducation.com/content/1/1/5

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

Adams, R. J., & Wu, M. (2007). The mixed-coefficients multinomial logit model. A generalized form of the Rasch model. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 55–76). New York, NY: Springer.

Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). *ACER conquest 4.0*. Melbourne, Australia: ACER.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: The Guilford Press.

Frey, A., Hartig, J., & Rupp, A. 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*, 39–53.

Ghosh, M., & Rao, J. N. K. (1994). Small area estimation: An appraisal. *Statistical Science*, *9*, 55–93.

Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. *IEA-ETS Research Institute Monograph*, *3*, 125–156.

Honaker, J., King, G., & Blackwell, M. (2010). Amelia II: A program for missing data [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=Amelia (R package version 1.2-18)

Kaplan, D., & McCarty, A. T. (2013). Data fusion with international large scale assessments: A case study using the OECD PISA and TALIS surveys. *Large-Scale Assessments in Education*, *1*, 6. Retrieved from http://www.largescaleassessmentsineducation.com/content/1/1/6

Kiefer, T., Robitzsch, A., & Wu, M. (2014). *TAM: Test analysis modules* [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=TAM (R package version 1.0-3.18-1).

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

Meinfelder, F. (2011). *BaBooN: Bayesian bootstrap predictive mean matching – multiple and single imputation for discrete data* [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=BaBooN (R package version 2.14.0).

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

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*, 133–161.

Montgomery, D. C. (2012). *Design and analysis of experiments* (8th ed.). Hoboken, NJ: Wiley.

Organization for Economic Cooperation and Development. (2006). *Assessing scientific, reading, and mathematical literacy: A framework for PISA 2006*. Paris, France: Author.

Organization for Economic Cooperation and Development. (2009a). *Creating effective teaching and learning results: First results from TALIS*. Paris, France: Author.

Organization for Economic Cooperation and Development. (2009b). *PISA 2006 technical report*. Paris, France: Author.

Organization for Economic Cooperation and Development. (2013). *The PISA 2012 assessment and analytic framework: Mathematics, reading, science, problem solving, and financial literacy*. Paris, France: Author.

Organization for Economic Cooperation and Development. (2014). *PISA 2012 technical report*. Paris, France: Author.

R Core Team. (2014). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Rao, J. N. K. (2004). *Small area estimation*. Hoboken, NJ: John Wiley & Sons.

Rässler, S. (2002). *Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches*. New York, NY: Springer.

Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics*, *9*, 130–134.

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

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York, NY: Chapman & Hall/CRC.

Shoemaker, D. M. (1973). *Principles and procedures of multiple matrix sampling*. Oxford, England: Balinger.

van Buuren, S. (2012). *Flexible imputation of missing data*. New York, NY: Chapman & Hall.

van Buuren, S., & Groothuis-Oudshoorn, K. (2010, January). *Multivariate imputation by chained equations, version 2.3*. Retrieved from http://www.multiple-imputation.com/

von Davier, M. (2009). Mixture distribution item response theory, latent class analysis, and diagnostic mixture models. In S. Ebretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 11–34). Washington, DC: APA Press.

von Davier, M. (2013). Imputing proficiency data under planned missingness in population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 175–201). Boca Raton: Chapman Hall/CRC.

Authors

DAVID KAPLAN is the Patricia Busk Professor of Quantitative Methods in the Department of Educational Psychology at the University of Wisconsin - Madison. Email: david.kaplan@wisc.edu. His research interests are in Bayesian statistical methods with applications to quasi-experimental and observational research.

DAN SU is a doctoral student in Quantitative Methods in the Department of Educational Psychology at the University of Wisconsin - Madison. Email: dsu4@wisc.edu. Her current research interests are planned missing data, causal inference and experimental designs.

Manuscript received July 3, 2015
Revision received October 2, 2015
Accepted November 20, 2015