JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education

Journal of Educational Measurement
Fall 2011, Vol. 48, No. 3, pp. 293–312

The Impact of Missing Background Data on Subpopulation Estimation

Leslie Rutkowski
Indiana University

Although population modeling methods are well established, a paucity of literature appears to exist regarding the effect of missing background data on subpopulation achievement estimates. Using simulated data that follows typical large-scale assessment designs with known parameters and a number of missing conditions, this paper examines the extent to which missing background data impacts subpopulation achievement estimates. In particular, the paper compares achievement estimates under a model with fully observed background data to achievement estimates for a variety of missing background data conditions. The findings suggest that subpopulation differences are preserved under all analyzed conditions while point estimates for subpopulation achievement values are influenced by missing at random conditions. Implications for cross-population comparisons are discussed.

Large-scale assessment programs such as the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and the Programme for International Student Assessment (PISA) are tasked with measuring what populations of students know and can do in a number of learning areas including mathematics, science, and reading. Given the formidable scope of these programs, it is necessary that creative assessment designs are employed to ensure sufficient content coverage and that assessed groups are measured with appropriate precision. To that end, large-scale assessment programs generally use a sophisticated assessment design whereby each individual student is administered just a small number of the total possible items, yet all items are administered throughout each of the reporting groups. This approach to item administration is often referred to as item sampling (Lord, 1962) or, more commonly in current large-scale assessment literature, as multiple-matrix sampling (Shoemaker, 1973).

Although this method of item delivery is efficient from an administration perspective, the approach poses currently intractable challenges for precisely estimating individual student achievement. Because only a fraction of the students in the population take any one item, and because any selected student takes only a fraction of the total available items, the actual distribution of student ability cannot be approximated by its empirical estimate (Mislevy, Johnson, & Muraki, 1992). In fact, traditional methods of estimating individual achievement introduce an unacceptable level of uncertainty and the possibility of serious aggregate-level bias (Little & Rubin, 1983; Mislevy, Beaton, Kaplan, & Sheehan, 1992). To overcome the methodological challenges associated with multiple-matrix sampling, large-scale assessment programs adopted a population or latent regression modeling approach that uses marginal estimation techniques to generate population- and subpopulation-level achievement estimates (Mislevy, 1991; Mislevy, Beaton, Kaplan & Sheehan, 1992; Mislevy, Johnson, & Muraki, 1992).

Under the population modeling approach, consistent population- and subpopulation-level ability estimates are achieved by treating achievement as missing (latent) data. These data points are missing for all examinees and are "filled in" using an approach analogous to multiple imputation (Rubin, 1976, 1987). As in multiple imputation methods, an imputation model (called a "conditioning model") is developed to predict individual student achievement *values* (from the posterior population model). This model uses all available student data (cognitive as well as background information) to generate a conditional proficiency distribution for each student from which to draw a number of plausible values (usually five) for each student on each latent trait (e.g., mathematics, science and associated subdomains).

Population modeling methods, used operationally for three decades, have been well established theoretically and empirically and subpopulation estimates of achievement derived from the conditioning models used in this approach are less biased than those estimated via traditional item response theory (IRT) methods (Mislevy, 1991; Mislevy, Beaton, Kaplan, & Sheehan, 1992; von Davier, Gonzalez, & Mislevy, 2009). Despite the sound foundation on which population modeling methods rest, a paucity of literature appears to exist regarding the effect of poor-quality background data on subpopulation achievement estimates. Further, in generating plausible values it is assumed that background data, used to derive posterior achievement distributions, are measured without error (Direct Estimation Software Interactive [DESI], 2009).

In an effort to better understand the effect of less-than-optimal-quality background data, the current paper seeks to examine the impact of missing background data used in the conditioning model to estimate subpopulation achievement. Using simulated data with known parameters and a conditioning model with a number of missing conditions (discussed in detail in the "Methods" section), this paper examines the extent to which subgroup estimation is biased. In particular, plausible value estimates under a model with fully observed background data are compared to plausible value estimates for a variety of missing background data conditions based on a population model that uses three simulated background characteristics and responses to a 70-item multiple-choice test.

Motivation for this line of inquiry rests in the following: It is well established that plausible values generated without using information about group membership underestimate group differences as the estimated subpopulation means shrink toward the overall population mean (Mislevy, 1991; von Davier et al., 2009). With this in mind, it is reasonable to imagine that group differences also might be biased when background variables used in the conditioning model have relatively high rates of *systematically* missing data. That is, background data with high missing rates can be thought of as unused (and unusable) information regarding actual group membership. Further, the effect on achievement estimates could differ from the limited case of underestimation given that group information is both unknown and subject to missing mechanisms known to cause biased parameter estimates (Rubin, 1976).

Background

Multiple-matrix sampling. NAEP, TIMSS, and PISA use multiple-matrix sampling in conjunction with a rotated booklet design that ensures that each item receives sufficient exposure and that each examinee receives an adequate number of items to estimate population-level achievement in (possibly) several domains and subdomains. For example, more than 10 hours of testable material was available for the TIMSS 2007 assessment (Olson, Martin, and Mullis, 2008). To minimize individual examinee burden, test developers used an assessment design that distributed 429 total mathematics and science items across 14 nonoverlapping mathematics blocks and 14 nonoverlapping science blocks. That is, the blocks exhaustively and mutually exclusively contained all available testing material. The blocks subsequently were arranged into 14 booklets containing two science and two mathematics blocks each, with no block-wise overlap within a booklet. That is, no block would appear more than once within a booklet. This design ensured linking across booklets because each block (and therefore each item) appeared in two different booklets. Further, the total assessment material was divided into more reasonable 90-minute periods of testing time for each student. It is important to note that this is just one of many possible designs that might fall under the umbrella of multiple-matrix sampling.

Population modeling. The population modeling approach is very briefly reviewed here. For a detailed explication of this method, interested readers are directed to Mislevy (1991), Mislevy, Johnson, and Muraki (1992), or Mislevy, Beaton, Kaplan, and Sheehan (1992). For an accessible primer on these methods, see von Davier et al. (2009). Because achievement (θ) is a latent, unobserved variable for every examinee, it is reasonable to treat θ as a missing value and to approximate statistics involving θ by its expectation. That is, for any statistic t, $\hat{t}(X,Y) = E\left[t(\theta,Y)\,|X,Y\right] = \int t(\theta,Y)\,p(\theta|X,Y)\,d\theta$, where X is a matrix of item responses for all examinees and Y is the matrix of responses of all examinees to the set of administered background questions. Because closed-form solutions typically are not available, random draws from the conditional distributions $p(\theta\,|\,\boldsymbol{x}_i, \boldsymbol{y}_i)$ are drawn for each sampled examinee, i (Mislevy, Johnson, & Muraki, 1992). In line with missing data practices (Rubin, 1987), values for each examinee are drawn multiple times. These values typically are referred to as *plausible values* in large-scale assessment terminology or *multiple imputations* in missing data literature.

The conditional distribution of θ is derived in the following way. Using first Bayes' theorem and then the IRT assumption of conditional independence, whereby the latent variable accounts for all associations among responses to various items in a specified domain (i.e., $P(x_i|\theta, y_i) = P\,(x_i|\theta)$), we have

$$p(\theta|\boldsymbol{x}_{i,}\boldsymbol{y}_i) \propto P(\boldsymbol{x}_i|\theta,\boldsymbol{y}_i)p(\theta|\boldsymbol{y}_i) = P(\boldsymbol{x}_i|\theta)p(\theta|\boldsymbol{y}_i), \tag{1}$$

where $p(\boldsymbol{x}_i\,|\,\theta)$ is the likelihood function for θ induced by observing x_i and $p(\theta\,|\,\boldsymbol{y}_i)$ is the distribution of θ for a given vector of background variables. In other words, the posterior distribution of a student with observed responses, $\boldsymbol{x}_i$, and vector of background characteristics, $\boldsymbol{y}_i$, is proportional to the product of the likelihood of θ induced by x_i through the response model and the population density (Mislevy, Beaton, Kaplan, & Sheehan, 1992). The distribution of θ, that is, $p(\theta\,|\,\boldsymbol{y}_i)$, is assumed normal with a mean given by the following linear model such that $\boldsymbol{y}^c$ is the vector of (usually assumed) *complete* background variables,

$$\theta = \mathbf{\Gamma}' \boldsymbol{y}^c + \boldsymbol{\varepsilon}, \tag{2}$$

where $\boldsymbol{\varepsilon} \sim N(0, \mathbf{\Sigma})$ and $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$ are estimated from the data (Mislevy, Johnson, & Muraki, 1992). Given modern IRT methods, a sufficient number of items and examinees, and a well-fitting measurement model, it is a fairly straightforward matter to estimate multiple plausible achievement values for every examinee. These estimates can then be used as ordinary values in subsequent statistical computations (e.g., estimates of group differences or more general models).

Missing data mechanisms. The current paper examines the impact of missing at random (MAR) or missing not at random (MNAR) background data on subpopulation estimates—two subtypes from the classical missing data taxonomy (Rubin, 1976). The third subtype in the taxonomy, missing completely at random (MCAR), is not considered in the current paper, because under MCAR omitting the cause of missingness from the model does not cause biased parameter estimates (Graham, Cumsille, & Elek-Fisk, 2003). Further, MAR and MNAR conditions are untestable and more difficult to ameliorate (Peugh & Enders, 2004; Schafer & Graham, 2002), which makes their impact more deserving of attention.

Under an MAR mechanism, the probability that a variable is observed is related to other variables; however, the missing variables are unrelated to the underlying values of the variables themselves. In other words, the mechanism of missingness is conditionally random. For example, when responses to items that elicit information about a respondent's socioeconomic status (SES) are systematically missing for respondents with parents who have low education levels, the data on this item are said to be MAR. More formally, Rubin's (1976) missing data typology defines MAR in the following way. Denote the complete data set as $\boldsymbol{Y}_{\text{com}}$ with dimensions $r \times c$. We can also consider the complete data set as composed of the observed and missing parts: $\boldsymbol{Y}_{\text{com}} = (\boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}})$ where $\boldsymbol{Y}_{\text{obs}}$ and $\boldsymbol{Y}_{\text{mis}}$ are the observed and missing parts, respectively. The, missing data are said to be MAR if

$$P(R|\boldsymbol{Y}_{\text{com}}) = P(R|\boldsymbol{Y}_{\text{obs}}), \tag{3}$$

where R is an $r \times c$ matrix of binary "missingness" indicators. Entries in R are typically 1 if a variable in $\boldsymbol{Y}_{\text{com}}$ is missing and 0 if the variable is present. Equation 3 specifies that the conditional distribution of missingness does not depend on $\boldsymbol{Y}_{\text{mis}}$.

When (3) is violated and the distribution of R depends on $\boldsymbol{Y}_{\text{mis}}$, data are said to be MNAR. In other words, under an MNAR mechanism, the probability that a variable is missing depends on the underlying value of the variable. For example, responses to questions regarding SES that are systematically missing for respondents from low SES families are said to be MNAR. Data that are MAR or MNAR are known to cause biased parameter estimates (e.g., Rubin, 1987; Schafer & Graham, 2002).

Population models and missing background data. As previously mentioned, the conditional distribution of θ is given by a linear model with the vector of background variables, $\boldsymbol{y}_i$, serving as predictors. This approach generates regression coefficient estimates, $\hat{\boldsymbol{\Gamma}}$ that relate the vector of predictor variables to the latent response.

Current operational procedures for developing this *conditioning* (or imputation) model use dummy codes for missing observations that then are used in the conditioning model (this is discussed in more detail subsequently). Schafer and Graham (2002) refer to this sort of missing data treatment as a "trick" that merely redefines the parameters or the population but does not deal with missing data in a principled way. Further, when observations are missing on $\boldsymbol{y}_i$, the vector of estimated regression coefficients for the complete data, $\hat{\boldsymbol{\Gamma}}_{\text{complete}}$ can differ from the vector of estimated regression coefficients for the observed data, $\hat{\boldsymbol{\Gamma}}_{\text{observed}}$.[1] The magnitude of the difference $(\hat{\boldsymbol{\Gamma}}_{\text{complete}} - \hat{\boldsymbol{\Gamma}}_{\text{observed}})$ depends in large part on the mechanism that generated the missing data as well as the amount of missing data (Rubin, 1987). Given the linear relationship between $\boldsymbol{\Gamma}$ and θ, it follows that distortions in $\hat{\boldsymbol{\Gamma}}_{\text{observed}}$ due to missing data on $\boldsymbol{y}_i$ will also manifest as distortions in $\hat{\boldsymbol{\theta}}_{\text{observed}}$ and any associated statistics, $\hat{t}(\boldsymbol{X}, \boldsymbol{Y}_{\text{observed}})$ The degree to which statistics, $\hat{t}(\boldsymbol{X}, \boldsymbol{Y}_{\text{complete}}) = E\left[t\left(\theta, \boldsymbol{Y}_{\text{complete}}\right) | \boldsymbol{X}, \boldsymbol{Y}_{\text{complete}}\right]$ and $\hat{t}(\boldsymbol{X}, \boldsymbol{Y}_{\text{observed}}) = E\left[t\left(\theta, \boldsymbol{Y}_{\text{observed}}\right) | \boldsymbol{X}, \boldsymbol{Y}_{\text{observed}}\right]$ differ under a number of conditions is the focus of the current study. Notable differences in the complete and observed solutions can have significant implications, particularly if policy-relevant subpopulation differences are incorrectly estimated.

Methods

To investigate the impact of missing background data on subpopulation achievement estimates, an assessment was simulated according to a number of known parameters. To mimic a reasonable multiple-matrix sampled assessment design, 70 multiple-choice TIMSS 2007 8th grade mathematics items were selected with their associated item parameter estimates (Olson et al., 2008). To estimate item parameters, the TIMSS 2007 measurement model used for these data was a three-parameter logistic IRT model (see Embretson & Reise, 2000 for an example of this model). Using these 70 items, seven booklets containing three blocks with 10 multiple-choice items each were assembled. This rotated booklet design is illustrated in Table 1, where cells marked with a "1" indicate that a particular block is contained in a given booklet. For example, booklet 1 is comprised of blocks A, B, and D. Also, block A can be found in booklets 1, 5, and 7. Here, we can see that every examinee attempts 30 items. And by randomly assigning booklets to students in a systematic rotation, every item is attempted by 43% of the sample while each block (and therefore item) appears three times per booklet rotation. Average item parameters for each of the booklets are presented in Table 2. As is typical in IRT notation, *a*, *b*, and *c* correspond to the discrimination, difficulty, and guessing parameters, respectively. This arrangement (of several considered) provided a reasonable balance of difficulty and discrimination across booklets.

In an attempt to maintain a fairly simple analysis, three uncorrelated, arbitrary background variables with two levels (high and low) and a proficiency mean and variance associated with each level were used as generating ability distributions for each of the subpopulations. Given three background variables with two levels each, this resulted in 2 × 2 × 2 = 8 fully conditional subpopulation membership possibilities, with 1,000 examinees in each subpopulation and an associated generating ability

Table 1
Simulated Assessment Design

		Booklet						
		1	2	3	4	5	6	7
Block	A	1	0	0	0	1	0	1
	B	1	1	0	0	0	1	0
	C	0	1	1	0	0	0	1
	D	1	0	1	1	0	0	0
	E	0	1	0	1	1	0	0
	F	0	0	1	0	1	1	0
	G	0	0	0	1	0	1	1

Table 2
Average Item Parameters by Booklet

		Booklet						
		1	2	3	4	5	6	7
Average Item Parameter	*a*	1.04	1.00	1.03	1.10	1.05	1.01	.95
	b	.22	.40	–.03	.44	.33	.34	.17
	c	.15	.15	.13	.14	.13	.13	.11

Table 3
Simulated Examinee Ability Means for Each Major Subgroup

	Category		
	High (*N*)	Low (*N*)	Overall (*N*)
Background Variable 1	.81 (4,000)	−.81 (4,000)	.00 (8,000)
Background Variable 2	.44 (4,000)	−.44 (4,000)	.00 (8,000)
Background Variable 3	.19 (4,000)	−.19 (4,000)	.00 (4,000)

Note. Each major subgroup has a variance fixed at 1.50.

distribution. The subpopulations are exhaustive (in this study) and mutually exclusive. A concrete example of such an arrangement might include the background variables gender (male, female); SES (high, low); and education level (high, low). Then, one of eight fully conditional subpopulations could be males from a low SES background and a high education level. The marginal generating ability distributions for each major subpopulation are presented in Table 3. The fully conditional generating ability distributions for the eight subpopulations are presented in Table 4.

Using the simulated sample of 8,000 students with generating ability distributions specified by subgroup membership, booklets were randomly assigned to examinees

Table 4
Simulated Examinee Ability Distribution for Each Fully Conditional Subgroup

	Background Variable 1			
	High Background Variable 2		Low Background Variable 2	
Background Variable 3	High (*N*)	Low (*N*)	High (*N*)	Low (*N*)
High	1.50 (1,000)	.50 (1,000)	–.25 (1,000)	–1.00 (1,000)
Low	1.00 (1,000)	.25 (1,000)	–.50 (1,000)	–1.50 (1,000)

Note. Each fully conditional subgroup has a variance fixed at 1.50.

in a rotated fashion to ensure that every block (and therefore every item) was administered an approximately equal number of times. Using known item parameters and specified generating examinee ability distributions, responses to the 70 items were subsequently simulated with the probability of a correct answer determined by an examinee's ability. Individual probabilities were compared with a random draw from a uniform distribution. If an examinee's probability of a correct answer was greater than the value from the random draw, the item was marked correct; otherwise, the item was marked incorrect. To assess the stability of the results, the test administration with complete background data was replicated 500 times. Three-parameter IRT models then were fit to the resulting 500 examinee by item response matrices to estimate item parameters.

The next step in the process of data simulation and preparation was to create patterns of missingness in the background data. In this analysis, the investigation included the cases of MAR and MNAR data. To model reasonable MAR and MNAR missing data situations, several missing patterns for the 500 examinee-by-item response matrices were generated according to the mechanisms and missing data percentages outlined in Table 5. For the MAR condition, 10%, 15%, and 20% of a background variable dependent on the level of one of the other two background variables was set as missing. For instance, missing data on background variable 1 was determined by a low level of background variable 2 with varying proportions of missing data. For the MNAR condition, 10%, 15%, and 20% of a background variable dependent on the level of that background variable was set as missing. For example, missing data on background variable 1 was determined by a low level of background variable 1. These data generations resulted in 500 replications $\times 3$ background variables $\times 3$ missing data rates $=$ 4,500 data matrices for the MAR and MNAR conditions each. These data matrices, including background variables, were subsequently used to generate plausible values via a latent regression model. Details for the latent regression model fitting are discussed next.

Population and subpopulation achievement were initially estimated using a conditioning model with the specified background variables, responses to the items, and item parameters that were estimated in an earlier step. These results served as a baseline against which to compare the results of the missing data conditions. To

Table 5
Missing Data Mechanisms and Percentages

Missing Data on:	Missing Mechanism	Percent of Missing Data Per Condition		
	Missing at Random			
BV1	If BV2 = Low	10	15	20
BV2	If BV1 = Low	10	15	20
BV3	If BV2 = High	10	15	20
	Missing Not at Random			
BV1	If BV1 = Low	10	15	20
BV2	If BV2 = Low	10	15	20
BV3	If BV3 = High	10	15	20

assess the impact of missingness, subpopulation achievement based on the MAR and MNAR background data was estimated and compared to the fully observed data condition. In line with operational approaches to estimating achievement in many large-scale assessment programs, a separate dummy code for missing responses on each background variable was assigned (Martin, Mullis, & Kennedy, 2007; OECD, 2009; Olson et al., 2008). For example, when data were specified as missing on background variable 1, two variables were used to capture the presence or absence of a response on this variable. That is,

$$BV1_{\text{Low}} = \begin{cases} 1 & \text{for background variable } 1 = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$BV1_{\text{Missing}} = \begin{cases} 1 & \text{for background variable } 1 = \text{missing} \\ 0 & \text{otherwise} \end{cases} .$$

Notice that, similar to models that use discrete variables as predictors, it was not necessary to include a code for high values of background variable 1 as this was redundant information. In this way, missing values were included as predictor variables in the conditioning model and, regardless of missing responses on the background questionnaires, plausible values were estimated for every examinee. To clarify, consider an example with only one background variable; then the conditioning model would be of the form:

$$\theta = \gamma_0 + \gamma_1(BV1_{\text{Low}}) + \gamma_2(BV1_{\text{Missing}}) + \varepsilon, \tag{4}$$

where γ_0, the intercept, corresponds to the mean ability level for respondents who endorsed a high level of background variable 1, γ_1 corresponds to the regression coefficient for the effect of being in the low level of background variable 1, and γ_2 corresponds to the regression coefficient for the effect of a missing response on background variable 1.

For the current analysis, data were generated using a modified macro to simulate a multiple-matrix sampled design (Gonzalez, 2009). The measurement models were fit to the data using Parscale 4.1 (Scientific Software International, Inc., 2003) and the population model used to estimate achievement was generated with DESI (2009), a freely available software package for population estimation.[2] DESI was originally developed by Educational Testing Service for estimating latent regression models with NAEP and other large-scale assessment data. In line with current large-scale assessment practice, five plausible values were generated for each examinee under each of the missing data conditions, including the condition where all background data are fully observed.

Results

To summarize the results of 500 replications for each missing data condition, subgroup estimate averages are presented for each of the plausible values under the fully observed condition and for each of the various missing data conditions.

Missing at Random

Only results for which subgroup data are MAR are presented. For example, when data are MAR on background variable 1, results for background variable 2 and background variable 3 are omitted. This choice was made partly in the interest of space and partly because MAR on one variable had no discernible impact on subgroup estimation for the other two variables, regardless of missing data rates. Table 6 presents the plausible value estimation results under the MAR condition for each of the subgroups when data are MAR for that subgroup. The first five rows of the table correspond to plausible value estimates for both categories of each background variable when data are fully observed. The remaining rows in Table 6 correspond to plausible value estimates for high and low levels of each background variable across various rates of missing data on the relevant variable, subject to the missing mechanism specified in the table.

To evaluate the impact of missing data, consider first the subpopulation average plausible value estimates for background variable 1, which suggest that as the rate of missing data increases, achievement estimates for both low and high levels of background variable 1 increase. The missing data effect is most apparent when comparing the results for background variable 1 in the fully observed condition to results for the 20% MAR condition. Specifically, achievement estimates increase by about 8% (calculated as a percentage changed and averaged across the five plausible values) for the low category and by about 17% for the high category as missing data increase. Under this particular MAR condition, subgroup point estimates are shifted slightly in the positive direction and by approximately equal amounts for both categories of the background variable. The magnitude and direction of the shift is reasonable, given the missing mechanism which caused missing data on background variable 1 only for examinees who had a low level of background variable 2 (and a relatively low mean generating ability of –.44). That is, 10%, 15%, and 20% of relatively low performers were not included in comparisons between high and low levels of background variable 1. An examination of subgroup achievement estimates

Table 6
Results for the MAR Mechanism

		BV1 (Missing if BV2 Low)						BV2 (Missing if BV1 Low)						BV3 (Missing BV2 High)					
		Low		High		Missing		Low		High		Missing		Low		High		Missing	
		Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Fully Observed	pv1	−.66	.83	.43	.84			−.41	.95	.17	.96			−.24	.99	.01	.98		
	pv2	−.67	.83	.42	.83			−.41	.95	.16	.95			−.25	.98	.00	.98		
	pv3	−.68	.83	.42	.85			−.42	.96	.16	.97			−.25	1.00	−.01	1.00		
	pv4	−.67	.83	.42	.83			−.41	.95	.16	.95			−.25	.99	.00	.98		
	pv5	−.66	.83	.42	.83			−.41	.95	.17	.95			−.24	1.00	−.01	.97		
	N	4,000		4,000					4,000		4,000			4,000		4,000			
10% MAR	pv1	−.63	.84	.45	.84	−.36	.92	−.35	.95	.23	.96	−.65	.84	−.28	.99	−.03	.99	.16	.95
	pv2	−.66	.84	.44	.84	−.32	.89	−.37	.96	.21	.95	−.61	.8	−.29	.99	−.04	.99	.15	.93
	pv3	−.65	.84	.45	.84	−.34	.89	−.36	.95	.21	.96	−.65	.81	−.28	.99	−.04	.98	.13	.97
	pv4	−.65	.83	.45	.85	−.35	.91	−.36	.95	.22	.96	−.64	.83	−.28	.99	−.03	.99	.14	.94
	pv5	−.64	.84	.46	.84	−.36	.90	−.36	.95	.23	.95	−.66	.82	−.27	1.00	−.02	.98	.13	.94
	N	3,623		3,601		776		3,623		3,614		763		3,599		3,621		780	
15% MAR	pv1	−.62	.84	.47	.85	−.35	.91	−.33	.95	.27	.95	−.64	.84	−.30	.99	−.04	.98	.15	.97
	pv2	−.64	.84	.46	.84	−.33	.90	−.34	.96	.25	.94	−.64	.81	−.31	.98	−.05	.99	.14	.95
	pv3	−.64	.84	.47	.85	−.35	.89	−.33	.95	.25	.95	−.65	.82	−.30	.99	−.05	.98	.13	.98
	pv4	−.64	.83	.47	.85	−.35	.90	−.33	.95	.26	.95	−.64	.82	−.30	.98	−.05	.99	.14	.96
	pv5	−.63	.84	.47	.84	−.36	.90	−.33	.95	.27	.95	−.66	.82	−.30	.99	−.04	.98	.14	.96
	N	3,441		3,417		1,142		3,441		3,408		1,151		3,392		3,431		1,177	

(Continued)

Table 6
Continued

		BV1 (Missing if BV2 Low)						BV2 (Missing if BV1 Low)						BV3 (Missing BV2 High)					
		Low		High		Missing		Low		High		Missing		Low		High		Missing	
		Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
20% MAR	pv1	−.60	.84	.50	.85	−.38	.91	−.28	.94	.31	.95	−.64	.83	−.32	.99	−.06	.98	.15	.97
	pv2	−.62	.84	.48	.84	−.37	.91	−.30	.95	.29	.94	−.63	.82	−.33	.98	−.07	.99	.14	.95
	pv3	−.62	.84	.49	.85	−.38	.90	−.28	.95	.30	.95	−.65	.82	−.31	.99	−.07	.98	.13	.97
	pv4	−.62	.83	.49	.85	−.37	.90	−.29	.95	.30	.95	−.64	.82	−.32	.98	−.07	.99	.14	.97
	pv5	−.61	.84	.50	.84	−.38	.90	−.28	.94	.32	.94	−.66	.83	−.31	.99	−.06	.98	.13	.97
	N	3,215		3,226		1,559		3,215		3,187		1,598		3,200		3,229		1,571	

Note. Each row marked pv1 represents the average of that plausible value and its standard deviation for the relevant subgroup across 500 replicates.

Table 7
MAR Subgroup Differences across Average Plausible Value Estimates and Varied Missingness

		BV1[a]		BV2[a]		BV3[a]	
Missing Data on:		Difference[b] High – Low	*SE*	Difference[b] High – Low	*SE*	Difference[b] High – Low	*SE*
Fully Observed	pv_1	1.09	.02	.58	.02	.25	.02
	pv_2	1.09	.02	.57	.02	.25	.02
	pv_3	1.10	.02	.58	.02	.24	.02
	pv_4	1.09	.02	.57	.02	.25	.02
	pv_5	1.08	.02	.58	.02	.23	.02
10% MAR	pv_1	1.08	.02	.58	.02	.25	.02
	pv_2	1.10	.02	.58	.02	.24	.02
	pv_3	1.10	.02	.57	.02	.23	.02
	pv_4	1.10	.02	.58	.02	.24	.02
	pv_5	1.10	.02	.59	.02	.24	.02
15% MAR	pv_1	1.09	.02	.6	.02	.25	.02
	pv_2	1.10	.02	.59	.02	.24	.02
	pv_3	1.11	.02	.58	.02	.23	.02
	pv_4	1.11	.02	.59	.02	.24	.02
	pv_5	1.10	.02	.60	.02	.24	.02
20% MAR	pv_1	1.10	.02	.59	.02	.25	.02
	pv_2	1.10	.02	.59	.02	.25	.02
	pv_3	1.11	.02	.58	.02	.23	.02
	pv_4	1.11	.02	.59	.02	.24	.02
	pv_5	1.11	.02	.60	.02	.24	.02

[a] The variable for which differences are estimated and data are missing.

[b] This difference is between high and low values of the relevant background variable across 500 replicates of five plausible value estimates.

for the missing category on background variable 1 confirms this finding: the estimated marginal ability (–.32 to –.38) is similar in magnitude to the generating mean ability for low levels of background variable 2. This uniform shift serves to preserve subgroup differences, as can be confirmed in the column of Table 7 labeled “BV1.” Again, the observed shift is reasonable given that within a level of background variable 1, data are MAR subject to a low level of background variable 2.

The middle columns of Table 6 summarize results for MAR data on background variable 2 with the missing mechanism determined by low levels of background variable 1. These findings suggest that plausible value estimates for background variable 2 are quite susceptible to MAR data on background variable 2. Similar to the results for background variable 1, as the level of missing data increases, the plausible value estimates for the low group improve. This apparent improvement is matched by an increase of a similar magnitude in the plausible value estimates for the high group. The tandem shift in both levels of background variable 2 serve to preserve subgroup differences, as can be seen in Table 7. As was the case for background variable 1 results, the observed patterns are easily explained by the missing data mechanism which caused missingness on background variable 2 only for examinees with low levels of background variable 1 (and, consequently, relatively low generating ability: –.81); this resulted in an overall increase in achievement for both levels of background variable 2. Again, an examination of the mean achievement estimates for the *missing* subgroup confirms this assertion, where the average, by plausible value, ranges from –.63 to –.66. Although subgroup differences are preserved when data are MAR, it is notable that at the highest levels of missingness, the plausible value estimates for the low level of background variable 2 are about 30% higher than when the data for background variable 2 are fully observed. Similarly, at the highest levels of missingness examined here, the high level of background variable 2 has plausible value estimates that are about 82% higher than when the data are fully observed.

The results for the condition where data on background variable 3 are missing with probability subject to high values of background variable 2 are presented in the far right columns of Table 6. Again, similar findings emerge for high and low levels of background variable 3. As in the results for background variables 1 and 2, shifts in subgroup estimates occur in similar directions for both subgroups, which serves to leave intact the differences between high and low levels of background variable 3. Difference results can be found in Table 7. In contrast to previous results, the shift observed for background variable 3 is in the opposite direction than it was for background variables 1 and 2 when data were MAR. In other words, the plausible value estimates are lower for the low level of background variable 3 as missing rates increase, whereas estimates for the high group are reduced by a similar magnitude. Specifically, the plausible value estimates under 20% MAR data are about 29% smaller for the low level of background variable 3 compared to fully observed data for background variable 3. And for the high level of background variable 3, plausible values are estimated to be six times smaller for the highest levels of missingness when compared to fully observed background data. As with the previous two results, the direction and approximate magnitude of the shift in achievement estimates is in line with the missing data mechanism, which specified as missing observations with a high level of background variable 2 (and relatively high generating ability). Finally, an inspection of the column of mean achievement estimates for the *missing* subgroup in Table 6 confirms that examinees with relatively high achievement (.13 to .16) were omitted from the achievement comparison between high and low levels of background variable 3.

Missing Not at Random

Similar to the MAR condition, only results for those background variables that have MNAR data are presented. The results for the MNAR condition that specifies background variable 1 missing with probability subject to low values on background variable 1 are located in Table 8 under the columns labeled *BV1 (Missing if BV1 Low)*. Identical to the MAR results, the first five rows are the fully observed results to compare against increasing missingness on each of the background variables. Unlike the MAR conditions, there are no discernible changes in the plausible value

Table 8

Results for the MNAR Mechanism

		BV1 (Missing if BV1 Low)						BV2 (Missing if BV2 Low)						BV3 (Missing if BV3 High)					
		Low		High		Missing		Low		High		Missing		Low		High		Missing	
		Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Fully Observed	pv1	−.66	.83	.43	.84			−.41	.95	.17	.96			−.24	.99	.01	.98		
	pv2	−.67	.83	.42	.83			−.41	.95	.16	.95			−.25	.98	.00	.98		
	pv3	−.68	.83	.42	.85			−.42	.96	.16	.97			−.25	1.00	−.01	1.00		
	pv4	−.67	.83	.42	.83			−.41	.95	.16	.95			−.25	.99	.00	.98		
	pv5	−.66	.83	.42	.83			−.41	.95	.17	.95			−.24	1.00	−.01	.97		
	N	4,000		4,000					4,000		4,000			4,000		4,000			
10% MNAR	pv1	−.66	.83	.42	.84	−.67	.86	−.41	.94	.17	.96	−.39	.99	−.24	.99	.00	.98	−.02	1.01
	pv2	−.69	.83	.41	.83	−.62	.82	−.44	.96	.16	.95	−.35	.95	−.25	.99	−.01	.99	−.01	.98
	pv3	−.68	.83	.42	.84	−.66	.83	−.42	.95	.16	.96	−.37	.95	−.24	.99	−.01	.98	−.03	1.00
	pv4	−.68	.82	.42	.84	−.66	.85	−.42	.95	.16	.96	−.38	.98	−.25	.99	.00	.99	−.04	1.00
	pv5	−.66	.83	.42	.83	−.67	.84	−.42	.94	.17	.96	−.39	.97	−.24	1.00	.01	.98	−.04	.98
	N	3,237		4,000		763		3,224		4,000		776		4,000		3,210		790	
15% MNAR	pv1	−.66	.83	.42	.84	−.66	.86	−.42	.94	.17	.96	−.38	.98	−.24	.99	.01	.98	−.02	1.00
	pv2	−.69	.83	.41	.83	−.65	.83	−.44	.96	.16	.95	−.36	.96	−.25	.99	0	.99	−.02	.98
	pv3	−.68	.83	.42	.84	−.67	.83	−.42	.95	.16	.96	−.38	.95	−.24	.99	−.01	.98	−.03	.99
	pv4	−.68	.82	.42	.84	−.66	.84	−.43	.95	.16	.96	−.38	.96	−.25	.99	.01	.99	−.03	.99
	pv5	−.66	.83	.42	.83	−.68	.84	−.42	.94	.17	.96	−.39	.96	−.24	1.00	.01	.98	−.04	.99
	N	2,849		4,000		1,151		2,858		4,000		1,142		4,000		2,830		1,170	
20% MNAR	pv1	−.67	.82	.42	.84	−.65	.85	−.41	.94	.17	.96	−.41	.96	−.24	.99	.02	.98	−.03	1.00
	pv2	−.69	.83	.41	.83	−.65	.84	−.43	.95	.16	.95	−.39	.96	−.25	.99	.01	.99	−.03	.99
	pv3	−.68	.82	.42	.84	−.66	.84	−.41	.95	.16	.96	−.41	.95	−.24	.99	.01	.98	−.04	.99
	pv4	−.69	.82	.42	.84	−.65	.84	−.42	.95	.16	.96	−.4	.95	−.25	.99	.02	.99	−.04	1.00
	pv5	−.66	.82	.42	.83	−.67	.85	−.41	.94	.17	.96	−.41	.96	−.24	1.00	.03	.97	−.05	.99
	N	2,402		4,000		1,598		2,441		4,000		1,559		4,000		2,407		1,593	

Note. Each row marked pv1 represents the average of that plausible value and its standard deviation for the relevant subgroup across 500 replicates.

estimates or associated standard deviations for background variable 1, regardless of the level of missingness or the subpopulation under examination compared to the fully observed data. Instead, the average point estimate for each of the five plausible values is quite stable for both high and low levels of background variable 1. It follows that differences between high and low levels of the relevant subpopulation also are preserved. Further, the only noticeable impact of missing data on both subpopulations within background variable 1 is a steady increase in the standard error for each difference estimate—a predictable function of decreased sample size for the low subgroup of background variable 1. Both of these observations can be confirmed in Table 9. These and subsequent MNAR results are reasonable when we consider that for a given missing mechanism (e.g., data are missing on background variable 1 only for low levels of background variable 1), actual missing data are MCAR. That is, missing data on background variable 1 can come from anywhere within the generating ability distribution of low levels of background variable 1. As such, achievement point estimates and differences are generally preserved, while standard errors for the relevant background variable level increase with higher missing data rates.

The results for MNAR data on background variable 2 and background variable 3 are consistent with the findings for background variable 1. That is, average point estimates for each subpopulation within a background variable are stable, regardless of the level of missing data or the subpopulation. Results for background variable 2 and background variable 3 can be found in the middle and far right columns of Table 8, respectively. Further, differences between subpopulations on both background variables 2 and 3 are preserved across all analyzed levels of missing data on both background variables. Finally, as the level of missing data on background variables 2 and 3 increases (and sample sizes for the relevant subpopulation decrease), standard errors associated with the achievement estimates increase slightly. The last four columns of Table 9 summarize these results.

Discussion and Conclusion

Given the political sensitivities associated with making comparisons across subgroups on variables such as SES and gender, it is important to understand the impact that less-than-optimal background instruments might have in the estimation of achievement in large-scale assessment. The current paper attempted to investigate, in a limited but controlled context, the impact of missingness on subpopulation achievement. In particular, test data were simulated under a rotated multiple-matrix sampled design whereby each "examinee" was administered only part of the total test items. This approach closely matched many large-scale assessment programs administered nationally and internationally. Further, generating ability distributions associated with background characteristics were used to generate item responses to a 70-item test. Based on the fully observed background data, a number of missing data conditions were created: from 10% to 20% on each of three background variables, subject to both an MAR and an MNAR mechanism. Using an IRT model in conjunction with a latent regression approach, subpopulation level achievement was estimated for each of the missing data conditions. Finally, the results of

Table 9
MNAR Subgroup Differences across Average Plausible Value Estimates and Varied Missingness

	BV1[a]		BV2[a]		BV3[a]	
	Difference[b]		Difference[b]		Difference[b]	
Missing data on:	High – Low	*SE*	High – Low	*SE*	High – Low	*SE*
Fully Observed	1.09	.019	.58	.021	.25	.022
	1.09	.019	.57	.021	.25	.022
	1.10	.019	.58	.022	.24	.022
	1.09	.019	.57	.021	.25	.022
	1.08	.019	.58	.021	.23	.022
10% MAR	1.08	.020	.58	.022	.24	.023
	1.10	.020	.60	.023	.24	.023
	1.10	.020	.58	.023	.23	.023
	1.10	.020	.58	.023	.25	.023
	1.08	.020	.59	.022	.25	.023
15% MAR	1.08	.020	.59	.023	.25	.024
	1.10	.020	.60	.023	.25	.024
	1.10	.020	.58	.023	.23	.024
	1.10	.020	.59	.023	.26	.024
	1.08	.020	.59	.023	.25	.024
20% MAR	1.09	.021	.58	.024	.26	.025
	1.10	.021	.59	.024	.26	.025
	1.10	.021	.57	.024	.25	.025
	1.11	.021	.58	.024	.27	.025
	1.08	.021	.58	.024	.27	.025

[a] The variable for which differences are estimated and data are missing.
[b] This difference is between high and low values of the relevant background variable across 500 replicates of five plausible value estimates.

each condition were compared with the fully observed estimates. The findings from 500 replications of each condition suggest that subgroup differences are preserved despite varied levels of missing data of up to 20% MAR and MNAR. Further, point estimates within a background variable were found to be quite stable for all analyzed MNAR conditions. Results for subgroup estimates with fully observed data (when one other background variable was MAR or MNAR) suggest that plausible value estimates were stable, regardless of the level of missing data on the other background variables.

Notable in all of the MNAR analyses is that the average plausible value estimate for the missing background variable closely matches the subpopulation estimate from which the missing group came when data are fully observed. This result can be expected, given that data were randomly missing within a particular subpopulation, subject to the relevant MNAR mechanism. For example, from the distribution of 4,000 examinees that were classified as *low* on background variable 1, 10%, 15%, and 20% of those examinees were randomly selected as missing on background variable 1 for subsequent analyses. Such a missing mechanism, consistent with the definition of MNAR, served to maintain the point estimates, data spread, and subgroup differences of all background variables, including that variable for which data were missing. This finding suggests that under the conditions used to simulate and analyze these data, plausible value estimates appear relatively robust to simple MNAR data and subpopulation differences are stable under both MNAR and MAR conditions when up to 20% of the data are missing.

Although the current methods used to estimate subgroup differences appear fairly resistant to missing background data, this analysis found that shifts in achievement estimates occurred for all three background variables' levels for which data were missing in the MAR condition. In other words, the process of setting as missing (at random, dependent on the value of another background variable in the data set) a fixed proportion of responses to a single background variable had the effect that achievement for subgroups within that background variable was shifted. These shifts generally occurred in similar amounts and directions across levels of a background variable along the achievement continuum and could be readily explained by the nature of the missing mechanism. Given the uniformity of the shifts that occurred, comparisons between high and low levels of a background variable were consistent across all levels of missing data. Conversely, cross-population comparison might be impacted by these effects. The following example illustrates this potentially important point.

Consider a scenario where performance of boys in two countries, country A and country B, is of policy relevance for a researcher. Further, assume that gender data are fully observed and that country A is estimated to have significantly higher achievement than country B. Now, consider a slightly different scenario where many responses to the question asking about examinee gender are MAR in country B and fully observed in country A. Further, assume that the missing data cause a large positive shift along the theta continuum for boys and girls in country B who responded to the gender question. In this second case, researchers and policy makers might come to unfounded conclusions regarding cross-country performance of boys in the face of high levels of missing gender data in country B. That is, a positive shift in achievement due to missing data on the gender variable would cause our hypothetical researcher to conclude that boys from country A were achieving at lower levels than would be found if the background data were fully observed. This scenario is illustrated in Figure 1. Particularly problematic in this context is that, according to results from this analysis, shifts can occur in both directions; the magnitude depends on the missing mechanism and the amount of missing data. Further, the magnitude of the shift, should it occur, also is variable. And, given the state of the art in achievement estimation, these quantities and directions are, at present, unknown. It should be noted, too, that comparable populations over time conceivably also could be affected by missing background data (e.g., comparing achievement for girls across two testing cycles).

Findings from this preliminary analysis into the issue of less-than-optimal background data beg caution on the part of testing organizations, both within countries and internationally. Until methods are developed that can ameliorate or at least detect the degree to which subgroup achievement estimates are shifted due to MAR data,

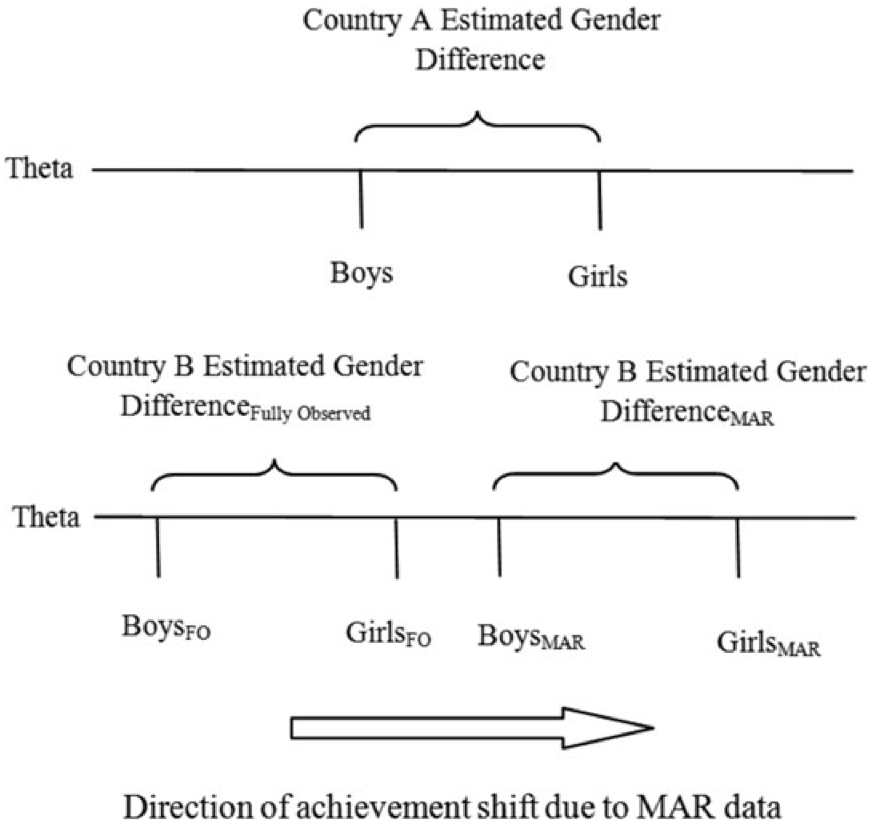
Country A Estimated Gender Difference
Theta
Boys
Girls
Country B Estimated Gender Difference$_{\text{Fully Observed}}$
Country B Estimated Gender Difference$_{\text{MAR}}$
Theta
Boys$_{\text{FO}}$
Girls$_{\text{FO}}$
Boys$_{\text{MAR}}$
Girls$_{\text{MAR}}$
Direction of achievement shift due to MAR data

Figure 1. Hypothetical shift in achievement due to MAR data on gender in country B.

it seems reasonable that, at the very least, study reports should publish missing data rates along with subpopulation achievement estimates and comparisons. Finally, this paper should serve to remind instrument developers and test administrators of the importance of high-quality instruments and careful data collection to minimize the occurrence of missing data.

This study examined the impact on subpopulation achievement when one background variable in the conditioning model had varied levels of MAR or MNAR data. As with any simulation, the conditions, and therefore the generalizability of the study, are somewhat limited. In addition, it often is the case in large-scale assessment data that multiple background variables are missing. Although findings from the current study suggest that there is an impact of missing data on subpopulation achievement, findings may differ under a more complex missing data structure. Also, in this study sample sizes were equal within and across subgroups. It is reasonable to expect that large differences in the proportion of examinees that belong in a particular background variable category might result in different findings, particularly with respect to standard errors or when all or nearly all of a subgroup is missing. Further work in this area is necessary to better understand the impact of poor-quality background data in general and missing data in particular on subpopulation estimation. Increasing the rates of missingness or including more variables with missing data in the conditioning model would provide a clearer and perhaps more realistic picture of the impact of missing data. It might also prove useful and result in different findings to examine the impact on subpopulation achievement of data that are MNAR on a background variable that is closely related to particular levels of achievement. It is reasonable to hypothesize meaningful impacts of missing data given that under this particular condition, background variables are missing due to levels of that background variable and due to certain levels of achievement. For example, imagine a situation where missing data on an SES variable is systematically associated with low achievement and a low level of SES. If low SES levels are also associated with poorer achievement, subgroup differences might be misestimated and the shifts in achievement observed in this study might intensify or manifest in entirely different ways. Finally, the principled treatment of missing background data subsequent to achievement estimation (e.g., via multiple imputation) also might serve to reduce the impact of biased subgroup achievement estimates and could be a viable option for correcting at least MAR biases *post hoc*; however, validity of this argument remains to be investigated.

Acknowledgments

The author thanks Ted Blew for his patient help with DESI; Eugene Gonzalez for feedback on an earlier draft of this manuscript; and the journal editor and three anonymous reviewers for their thoughtful feedback and constructive criticism. Any errors are the author’s own.

Notes

[1] In this instance, the dimension of $\hat{\boldsymbol{\Gamma}}_{\text{observed}}$ will be greater than $\hat{\boldsymbol{\Gamma}}_{\text{complete}}$ due to additional coefficients that correspond to dummy coded missing responses for each variable in y; however, our interest is only in those regression coefficients that are in common between the observed and complete solutions.

[2] Interested researchers should email desi@ets.org for more information.

References

Direct Estimation Software Interactive (DESI) (2009). Version 3.23 [Computer software and manual]. Princeton, NJ: Educational Testing Service.

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

Gonzalez, E. (2009). GenItmDat macro for SAS. [SAS macro]. Princeton, NJ: Educational Testing Service.

Graham, J. W., Cumsille, P. E., & Elek-Fisk, E. (2003). Methods for handling missing data. In J. A. Schinka & W. F. Velicer (Eds.), *Comprehensive handbook of psychology: Vol. 2. Research methods in psychology* (pp. 87–114). New York, NY: Wiley.

Little, R., & Rubin, D. (1983). On jointly estimating parameters and missing data by maximizing the complete-data likelihood. *The American Statistician*, *37*, 218–220.

Lord, F. (1962). Estimating norms by item-sampling. *Educational and Psychological Measurement*, *22*, 259–267.

Martin, M., Mullis, I., & Kennedy, A. (Eds.) (2007). *PIRLS 2006 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

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.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, *17*, 131–154.

Olson, J., Martin, M., & Mullis, I. (Eds.) (2008). *TIMSS 2007 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Organisation for Economic Co-operation and Development (OECD) (2009). *PISA 2006 technical report*. Paris, France: OECD.

Peugh, J., & Enders, C. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, *74*, 525–556.

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

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

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

Scientific Software International, Inc. (2003). Parscale for Windows (Version 4.1). Chicago: Scientific Software International, Inc.

Shoemaker, D. M. (1973). *Principles and procedures of multiple matrix sampling*. Cambridge, MA: Ballinger Publishing Company.

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). Plausible values: What are they and why do we need them? *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *2*, 9–36.

Author

LESLIE RUTKOWSKI is Assistant Professor of Inquiry Methodology in the Department of Counseling and Educational Psychology at Indiana University, 201 N. Rose Avenue, Bloomington, IN, 47405; e-mail: lrutkows@indiana.edu. Her research is focused in the area of international large-scale assessment from both methodological and applied perspectives. Her interests include the impact of background questionnaires on assessment results and integrating survey methods with statistical modeling to better exploit large-scale assessment data.