The Impact of Missing at Random Data on Subgroup Estimation

Leslie Rutkowski

Indiana University

International large-scale assessment programs such as the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Literacy Study (PIRLS) 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 novel assessment designs are employed to ensure sufficient content coverage and that groups are measured with appropriate precision. To this end, TIMSS and PIRLS use a sophisticated assessment design whereby each individual student is only administered a small number of the total possible cognitive 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 LSA literature, as multiple-matrix sampling (Shoemaker, 1973). TIMSS and PIRLS use multiple-matrix sampling in conjunction with a rotated booklet design that ensures that each cognitive item receives sufficient exposure and that each examinee receives a sufficient number of items to estimate population-level achievement in a number of domains and sub-domains.

The following example illustrates multiple-matrix sampling using the TIMSS 2007 design. In total, more than 10 hours of testing time was available for the TIMSS 2007 assessment. To minimize individual examinee burden, test developers used an assessment design that distributed 429 total mathematics and science items across 14

non-overlapping mathematics blocks and 14 non-overlapping 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. This design ensured linking across booklets since each block (and therefore each item) appeared in two different booklets. Booklets were then administered such that 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 one of many possible designs that might fall under the umbrella of multiple-matrix sampling. The TIMSS 2007 design is represented in Table 1.

Table 1

TIMSS 2007 Booklet Design

	Part 1		Part 2	
Booklet	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	M01	M02	S01	S02
2	S02	S03	M02	M03
3	M03	M04	S03	S04
4	S04	S05	M04	M05
5	M05	M06	S05	S06
6	S06	S07	M06	M07
7	M07	M08	S07	S08
8	S08	S09	M08	M09
9	M09	M10	S09	S10
10	S10	S11	M10	M11
11	M11	M12	S11	S12
12	S12	S13	M12	M13
13	M13	M14	S13	S14
14	S14	S01	M14	M01

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 item, and 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, & Sheahan, 1992). To overcome the methodological challenges associated with multiple-matrix sampling, international LSA programs adopted a population or latent regression modeling approach that uses marginal estimation techniques to generate population level achievement estimates (Mislevy, 1991; Mislevy, Beaton, Kaplan & Sheehan, 1992; Mislevy, Johnson & Muraki, 1992).

Under the population modeling approach, consistent population-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 similar 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 sub-domains). The population modeling approach is very briefly reviewed next. For a detailed

explication of this method, interested readers are directed to Mislevy (1991), Mislevy, Johnson, and Muraki (1992), or Mislevy, Kaplan, Beaton, and Sheehan (1992).

Because 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*, $\hat{t}(\mathbf{X}, \mathbf{Y}) = \mathrm{E}\left[t(\theta, \mathbf{Y})|\mathbf{X}, \mathbf{Y}\right] = \int t(\theta, \mathbf{Y}) p(\theta|\mathbf{X}, \mathbf{Y}) d\theta$, 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 are typically not available, random draws from the conditional distributions $p(\theta|\mathbf{x}_i, \mathbf{y}_i)$ 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 are typically referred to as *plausible values* in LSA terminology or *multiple imputations* in missing data literature.

Using Bayes' theorem and the IRT assumption of conditional independence,

$$\begin{aligned} p(\theta|\mathbf{x}_i, \mathbf{y}_i) &\propto P(\mathbf{x}_i|\theta, \mathbf{y}_i) p(\theta|\mathbf{y}_i) \\ &= P(\mathbf{x}_i|\theta) p(\theta|\mathbf{y}_i), \end{aligned}$$

where $P(\mathbf{x}_i|\theta)$ is the likelihood function for θ and $p(\theta|\mathbf{y}_i)$ is the distribution of θ for a given vector of response variable. The distribution of θ is assumed normal with a mean given by the following linear model such that $\mathbf{y}^c$ is the vector of (usually assumed) *complete* background variables,

$$\theta = \mathbf{\Gamma}' \mathbf{y}^c + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{\Sigma})$ and $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$ are estimated. 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).

Although population modeling methods have been well established theoretically and empirically and subpopulation estimates of achievement derived from the conditioning models are less biased than those estimated via traditional IRT methods (Mislevy, 1991; Mislevy, Beaton, Kaplan, & Sheehan, 1992; von Davier, Gonzalez, & Mislevy, 2009), 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 is measured without error (Direct Estimation Software Interactive, 2009).

The current paper seeks to examine the impact of missing background data used to estimate subpopulation achievement. In particular, I propose to examine one quality aspect of background data. First, 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, Gonzalez, & Mislevy, 2009). It is reasonable to imagine that group differences might also be underestimated when background variables used in the conditioning model have relatively high rates of missing at random (MAR) data – one subtype from the classical missing data taxonomy (Rubin, 1976). Under MAR, the probability that a data point is observed depends on the value of another variable. 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. Data that are MAR

are known to cause biased parameter estimates (e.g. Rubin, 1987; Schafer & Graham, 2002).

Using simulated data with known parameters and a number of missing conditions, discussed in detail in the Methods section, the extent to which subgroup estimation is biased is examined. In particular, I compare plausible value estimates under a model with fully observed background data 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 cognitive test.

Methods

To investigate the impact of randomly missing background data on sub-population achievement estimates, an assessment was simulated according to a number of known parameters. To mimic a reasonable multiple-matrix sampled assessment design, I selected 70 multiple choice TIMSS 2007 8th grade mathematics items and associated item parameter estimates. The TIMSS 2007 measurement model used for these data was a three-parameter logistic item response theory (3PL IRT) model (see Embretson & Reise, 2000 for an example of this model). Using these 70 items, I then assembled seven booklets containing three blocks with ten multiple choice items. The rotated booklet design is illustrated in Table 2, 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 Booklet 1, 5, and 7. Here, we can see that every examinee attempts 30 items and that, by randomly assigning booklets to students in a systematic rotation, every item is attempted by 43% of the sample. Average item

parameters for each of the booklets are presented in Table 3. This arrangement, of several considered, provided a reasonable balance of difficulty and discrimination across booklets.

Table 2

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 3

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	0.95
	b	0.22	0.40	-0.03	0.44	0.33	0.34	0.17
	c	0.15	0.15	0.13	0.14	0.13	0.13	0.11

In an attempt to maintain a fairly simple analysis, three arbitrary background variables (BVs) with two levels (high and low) were used to generate known proficiency distributions. This approach yielded eight associated subgroups with 1000 students in

each group. Fully conditional subgroup ability distributions are represented in Table 4. Collapsing over the other two categories, the ability distributions for each major subgroup are presented in Table 5.

Table 4

Simulated Examinee Ability Distributions (Fully Conditional)

Background variable 1	Background variable 2	Background variable 3	N	Average Ability
Low	Low	Low	1000	-1.50
Low	Low	High	1000	-1.00
Low	High	Low	1000	-0.50
Low	High	High	1000	-0.25
High	Low	Low	1000	0.25
High	Low	High	1000	0.50
High	High	Low	1000	1.00
High	High	High	1000	1.50

Table 5

Simulated Examinee Ability Distributions for Each Major Subgroup

	Category		
	High	Low	Overall
Background variable 1	0.8125	-0.8125	0
Background variable 2	0.4375	-0.4375	0
Background variable 3	0.1875	-0.1875	0

Based on the simulated sample of 8000 students with varied abilities, booklets were randomly assigned to examinees 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 known examinee ability distributions, responses to the 70 cognitive items were subsequently simulated. In order to assess the stability of the results, the test administration with complete background data was replicated 500 times.

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 was limited to the case of MAR data. To model reasonable MAR missing data situations, I generated several missing patterns for the 500 examinee-by-item response matrices according to the mechanisms and missing data percentages outlined in Table 6. For the MAR condition, 10, 15, and 20 percent of a BV dependent on the level of one of the other two BVs was set as missing. For instance, missing data on BV 1 was determined by a low level of BV 2, with varying proportions of missing data. The resultant 1500 data matrices were classified as MAR for subsequent IRT model fitting and plausible value generation.

Table 6

Missing Data Mechanisms and Percentages Under MAR

Missing data on:	Missing mechanism	Percent of missing data per condition		
BV 1	If BV 2 = Low	10	15	20
BV 2	If BV 1 = Low	10	15	20
BV 3	If BV 2 = High	10	15	20

Using a conditioning model with the specified background variables and responses to the cognitive items, population and subpopulation achievement based on complete background data was estimated. To assess the impact of missingness, population and subpopulation achievement based on the MAR data was subsequently

estimated and compared to the fully observed data condition. In line with operational approaches to estimating achievement in TIMSS and PIRLS, a dummy code for missing background variables was assigned. In this way, missing values are included as background variables in the conditioning model. For example, when data are specified as missing on BV 1, two variables are used to capture the presence or absence of a response on this variable. That is,

$$\mathrm{BV1}_{Low} = \begin{cases} 2 & \text{for BV1} = 2 \\ 1 & \text{otherwise} \end{cases}$$

$$\mathrm{BV1}_{Missing} = \begin{cases} 2 & \text{for BV1} = \text{Missing} \\ 1 & \text{otherwise} \end{cases}.$$

Notice, that it is not necessary to include a code for high values of BV1, as this is redundant information.

For the current analysis, data was 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 (2003) and the population model used to estimate achievement used the Direct Estimation Software Interactive (2009). 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, 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.

Table 7 presents the plausible value estimation under the missing mechanism such that low values of BV 2 were predictive of missing values for BV 1. The first five rows of the table correspond to estimates that used fully observed background variables. The remaining rows correspond to various rates of missing data on BV 1 subject to low values of BV 2, from 10% to 20%. When comparing plausible value estimates for high and low levels of BV 2 and BV 3, there appear to be no discernable difference depending on missing data on BV 1. More interestingly, there are also few differences across the mean plausible value estimates for high and low levels of BV 1, even at the highest level of missingness for that background variable. In fact, for the 20% MAR condition, the mean difference in plausible value estimates between high and low levels of BV 1 are, in the worst cases, just a few hundredths of a point different than the estimates for the fully observed data. These results are confirmed in the first column of Table 10.

Table 7

Results for the Missing Mechanism: If BV 2 Is Low, Then BV 1 Is Missing

		BV 1						BV2				BV 3			
		Low		High		Missing		Low		High		Low		High	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Fully Observed	pv_1	-0.66	0.83	0.43	0.84			-0.41	0.95	0.17	0.96	-0.24	0.99	0.01	0.98
	pv_2	-0.67	0.83	0.42	0.83			-0.41	0.95	0.16	0.95	-0.25	0.98	0.00	0.98
	pv_3	-0.68	0.83	0.42	0.85			-0.42	0.96	0.16	0.97	-0.25	1.00	-0.01	1.00
	pv_4	-0.67	0.83	0.42	0.83			-0.41	0.95	0.16	0.95	-0.25	0.99	0.00	0.98
	pv_5	-0.66	0.83	0.42	0.83			-0.41	0.95	0.17	0.95	-0.24	1.00	-0.01	0.97
	N	4000			4000				4000		4000		4000		4000
10% MAR	pv_1	-0.63	0.84	0.45	0.84	-0.36	0.92	-0.41	0.94	0.17	0.96	-0.24	0.99	0.01	0.98
	pv_2	-0.66	0.84	0.44	0.84	-0.32	0.89	-0.42	0.95	0.16	0.96	-0.25	0.99	-0.01	0.99
	pv_3	-0.65	0.84	0.45	0.84	-0.34	0.89	-0.41	0.95	0.16	0.97	-0.24	0.99	-0.01	0.98
	pv_4	-0.65	0.83	0.45	0.85	-0.35	0.91	-0.41	0.95	0.16	0.96	-0.24	0.99	0.00	0.99
	pv_5	-0.64	0.84	0.46	0.84	-0.36	0.90	-0.41	0.94	0.17	0.96	-0.24	1.00	0.00	0.98
	N	3623		3601		776		4000		4000		4000		4000	
15% MAR	pv_1	-0.62	0.84	0.47	0.85	-0.35	0.91	-0.41	0.94	0.17	0.97	-0.24	0.99	0.01	0.98
	pv_2	-0.64	0.84	0.46	0.84	-0.33	0.90	-0.41	0.95	0.16	0.96	-0.25	0.99	-0.01	0.99
	pv_3	-0.64	0.84	0.47	0.85	-0.35	0.89	-0.41	0.94	0.16	0.97	-0.24	1.00	-0.01	0.98
	pv_4	-0.64	0.83	0.47	0.85	-0.35	0.90	-0.41	0.94	0.16	0.97	-0.24	0.99	0.00	0.99
	pv_5	-0.63	0.84	0.47	0.84	-0.36	0.90	-0.41	0.94	0.17	0.97	-0.24	1.00	0.00	0.98
	N	3441		3417		1142		4000		4000		4000		4000	
20% MAR	pv_1	-0.60	0.84	0.50	0.85	-0.38	0.91	-0.41	0.94	0.17	0.97	-0.24	0.99	0.01	0.98
	pv_2	-0.62	0.84	0.48	0.84	-0.37	0.91	-0.41	0.95	0.16	0.96	-0.25	0.98	0.00	0.99
	pv_3	-0.62	0.84	0.49	0.85	-0.38	0.90	-0.41	0.94	0.16	0.97	-0.24	0.99	-0.01	0.98
	pv_4	-0.62	0.83	0.49	0.85	-0.37	0.90	-0.41	0.94	0.16	0.97	-0.24	0.99	0.00	0.99
	pv_5	-0.61	0.84	0.50	0.84	-0.38	0.90	-0.40	0.94	0.17	0.97	-0.24	1.00	0.00	0.98
	N	3215		3226		1559		4000		4000		4000		4000	

Table 8 presents the plausible value estimates for the missing condition such that low levels of BV 1 were predictive of missing data on BV 2. Similar to the previous results, plausible value estimates are consistent for high and low levels of the two groups for whom data are fully present (BV 1 and BV 3). Slightly different results emerge for BV 2 – the background variable for whom data are missing. In fact, as the level of

missing data increases, the plausible value estimates for the “low” group are higher. This growth is offset, however, by an increase of a similar magnitude in the plausible value estimates for the “high” group. The tandem shift in both levels of BV 2 serve to preserve subgroup differences, as can be seen in the second column of Table 10. 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 BV 2 are about 30% higher than when the data for BV 2 are fully observed. Similarly, at the highest levels of missingness examined here, the “high” level of BV 2 has plausible value estimates that are about 60% higher than when the data are fully observed.

Table 8

Results for the Missing Mechanism: If BV 1 Is Low, Then BV 2 Is Missing

		BV 1				BV 2						BV 3			
		Low		High		Low		High		Missing		Low		High	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Fully Observed	pv_1	-0.66	0.83	0.43	0.84	-0.41	0.95	0.17	0.96			-0.24	0.99	0.01	0.98
	pv_2	-0.67	0.83	0.42	0.83	-0.41	0.95	0.16	0.95			-0.25	0.98	0.00	0.98
	pv_3	-0.68	0.83	0.42	0.85	-0.42	0.96	0.16	0.97			-0.25	1.00	-0.01	1.00
	pv_4	-0.67	0.83	0.42	0.83	-0.41	0.95	0.16	0.95			-0.25	0.99	0.00	0.98
	pv_5	-0.66	0.83	0.42	0.83	-0.41	0.95	0.17	0.95			-0.24	1.00	-0.01	0.97
	N	4000		4000			4000		4000			4000		4000	
10% MAR	pv_1	-0.66	0.83	0.42	0.84	-0.35	0.95	0.23	0.96	-0.65	0.84	-0.24	0.99	0.01	0.98
	pv_2	-0.67	0.83	0.41	0.84	-0.37	0.96	0.21	0.95	-0.61	0.8	-0.25	0.98	-0.01	0.99
	pv_3	-0.67	0.83	0.42	0.84	-0.36	0.95	0.21	0.96	-0.65	0.81	-0.25	0.99	-0.01	0.98
	pv_4	-0.67	0.82	0.42	0.84	-0.36	0.95	0.22	0.96	-0.64	0.83	-0.24	0.99	-0.01	0.99
	pv_5	-0.66	0.83	0.42	0.84	-0.36	0.95	0.23	0.95	-0.66	0.82	-0.24	1.00	0.00	0.98
	N	4000		4000		3623		3614		763		4000		4000	
15% MAR	pv_1	-0.66	0.83	0.43	0.84	-0.33	0.95	0.27	0.95	-0.64	0.84	-0.24	0.99	0.01	0.98
	pv_2	-0.67	0.83	0.41	0.84	-0.34	0.96	0.25	0.94	-0.64	0.81	-0.25	0.98	-0.01	0.99
	pv_3	-0.67	0.83	0.42	0.84	-0.33	0.95	0.25	0.95	-0.65	0.82	-0.24	0.99	-0.01	0.98
	pv_4	-0.67	0.82	0.42	0.84	-0.33	0.95	0.26	0.95	-0.64	0.82	-0.24	0.99	-0.01	0.99
	pv_5	-0.66	0.83	0.42	0.84	-0.33	0.95	0.27	0.95	-0.66	0.82	-0.24	1.00	0.00	0.98
	N	4000		4000		3441		3408		1151		4000		4000	
20% MAR	pv_1	-0.66	0.83	0.43	0.84	-0.28	0.94	0.31	0.95	-0.64	0.83	-0.24	0.99	0.01	0.98
	pv_2	-0.67	0.82	0.41	0.84	-0.30	0.95	0.29	0.94	-0.63	0.82	-0.25	0.98	-0.01	0.99
	pv_3	-0.67	0.82	0.42	0.84	-0.28	0.95	0.30	0.95	-0.65	0.82	-0.24	0.99	-0.01	0.98
	pv_4	-0.67	0.82	0.42	0.84	-0.29	0.95	0.30	0.95	-0.64	0.82	-0.24	0.99	0.00	0.99
	pv_5	-0.66	0.83	0.42	0.84	-0.28	0.94	0.32	0.94	-0.66	0.83	-0.24	1.00	0.00	0.98
	N	4000		4000		3215		3187		1598					

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

Finally, the results for the condition where data on BV 3 are missing when BV 2 is a high value are presented in Table 9. In line with the two previous results, plausible values for subgroups that have fully observed data are stable across the missing data conditions. That is, plausible values are consistently estimated for high and low levels of

both BV 1 and BV 2 despite varied levels of missingness on BV 3. Again, different findings emerge for high and low levels of BV 3. As in the results for BV 2, shifts in the subgroup estimates occur, but in similar directions, which serves to leave intact the subgroup differences. Difference results can be found in the Table 10. In other words, the plausible value estimates are lower for the low level of BV 3 as missing rates increase while estimates for the high group are reduced by a similar degree. Further, the shift observed for BV 3 is in the opposite direction than BV 2 when data were MAR. In particular, the plausible value estimates under 20% MAR data are about 38% smaller for the low level of BV 3 compared to fully observed data for BV 3. And for the high level of BV 3, plausible values are estimated to be about 87% lower for the highest levels of missingness compared to fully observed background data.

Table 9

Results for the Missing Mechanism: If BV 2 Is High, Then BV 3 Is Missing

		Group 1				Group 2				Group 3					
		Low		High		Low		High		Low		High		Missing	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Fully Observed	pv1	-0.66	0.83	0.43	0.84	-0.41	0.95	0.17	0.96	-0.24	0.99	0.01	0.98		
	pv2	-0.67	0.83	0.42	0.83	-0.41	0.95	0.16	0.95	-0.25	0.98	0.00	0.98		
	pv3	-0.68	0.83	0.42	0.85	-0.42	0.96	0.16	0.97	-0.25	1.00	-0.01	1.00		
	pv4	-0.67	0.83	0.42	0.83	-0.41	0.95	0.16	0.95	-0.25	0.99	0.00	0.98		
	pv5	-0.66	0.83	0.42	0.83	-0.41	0.95	0.17	0.95	-0.24	1.00	-0.01	0.97		
	N	4000		4000			4000		4000	4000		4000			
10% MAR	pv1	-0.66	0.83	0.42	0.84	-0.41	0.95	0.17	0.96	-0.28	0.99	-0.03	0.99	0.16	0.95
	pv2	-0.68	0.83	0.42	0.83	-0.42	0.96	0.16	0.95	-0.29	0.99	-0.04	0.99	0.15	0.93
	pv3	-0.67	0.83	0.42	0.84	-0.41	0.95	0.15	0.96	-0.28	0.99	-0.04	0.98	0.13	0.97
	pv4	-0.67	0.83	0.42	0.84	-0.41	0.95	0.16	0.95	-0.28	0.99	-0.03	0.99	0.14	0.94
	pv5	-0.67	0.83	0.42	0.83	-0.41	0.95	0.17	0.95	-0.27	1.00	-0.02	0.98	0.13	0.94
	N	4000		4000		4000		4000		3599		3621		780	
15% MAR	pv1	-0.66	0.83	0.42	0.84	-0.41	0.95	0.17	0.96	-0.30	0.99	-0.04	0.98	0.15	0.97
	pv2	-0.68	0.83	0.42	0.83	-0.42	0.96	0.16	0.95	-0.31	0.98	-0.05	0.99	0.14	0.95
	pv3	-0.67	0.83	0.42	0.84	-0.41	0.95	0.15	0.96	-0.30	0.99	-0.05	0.98	0.13	0.98
	pv4	-0.67	0.83	0.42	0.84	-0.41	0.95	0.16	0.95	-0.30	0.98	-0.05	0.99	0.14	0.96
	pv5	-0.67	0.83	0.42	0.83	-0.41	0.95	0.17	0.95	-0.30	0.99	-0.04	0.98	0.14	0.96
	N	4000		4000		4000		4000		3392		3431		1177	
20% MAR	pv1	-0.66	0.83	0.42	0.84	-0.41	0.95	0.17	0.96	-0.32	0.99	-0.06	0.98	0.15	0.97
	pv2	-0.68	0.83	0.42	0.83	-0.42	0.96	0.16	0.95	-0.33	0.98	-0.07	0.99	0.14	0.95
	pv3	-0.67	0.83	0.42	0.84	-0.41	0.95	0.15	0.96	-0.31	0.99	-0.07	0.98	0.13	0.97
	pv4	-0.67	0.83	0.42	0.84	-0.41	0.95	0.16	0.95	-0.32	0.98	-0.07	0.99	0.14	0.97
	pv5	-0.67	0.83	0.42	0.83	-0.41	0.95	0.17	0.95	-0.31	0.99	-0.06	0.98	0.13	0.97
	N	4000		4000		4000		4000		3200		3229		1571	

Table 10

Subgroup Differences across Average Plausible Value Estimates and Varied Missingness

		BV 1[a]		BV 2[a]		BV 3[a]	
		Difference[b] High – Low	SE	Difference[b] High – Low	SE	Difference[b] High – Low	SE
Fully Observed	pv_1	1.09	0.02	0.95	0.02	0.25	0.02
	pv_2	1.09	0.02	0.95	0.02	0.25	0.02
	pv_3	1.10	0.02	0.96	0.02	0.24	0.02
	pv_4	1.09	0.02	0.95	0.02	0.25	0.02
	pv_5	1.08	0.02	0.95	0.02	0.23	0.02
10% MAR	pv_1	1.08	0.02	0.94	0.02	0.25	0.02
	pv_2	1.10	0.02	0.95	0.02	0.24	0.02
	pv_3	1.10	0.02	0.95	0.02	0.23	0.02
	pv_4	1.10	0.02	0.95	0.02	0.24	0.02
	pv_5	1.10	0.02	0.94	0.02	0.24	0.02
15% MAR	pv_1	1.09	0.02	0.94	0.02	0.25	0.02
	pv_2	1.10	0.02	0.95	0.02	0.24	0.02
	pv_3	1.11	0.02	0.94	0.02	0.23	0.02
	pv_4	1.11	0.02	0.94	0.02	0.24	0.02
	pv_5	1.10	0.02	0.94	0.02	0.24	0.02
20% MAR	pv_1	1.10	0.02	0.94	0.02	0.25	0.02
	pv_2	1.10	0.02	0.95	0.02	0.25	0.02
	pv_3	1.11	0.02	0.94	0.02	0.23	0.02
	pv_4	1.11	0.02	0.94	0.02	0.24	0.02
	pv_5	1.11	0.02	0.94	0.02	0.24	0.02

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

Discussion and Conclusion

Given the political sensitivities associated with making comparisons across subgroups such as SES levels and gender, it is important to understand the impact that less-than-optimal background instruments may have in the estimation of achievement in international large-scale assessment. The current paper attempted to investigate, in a

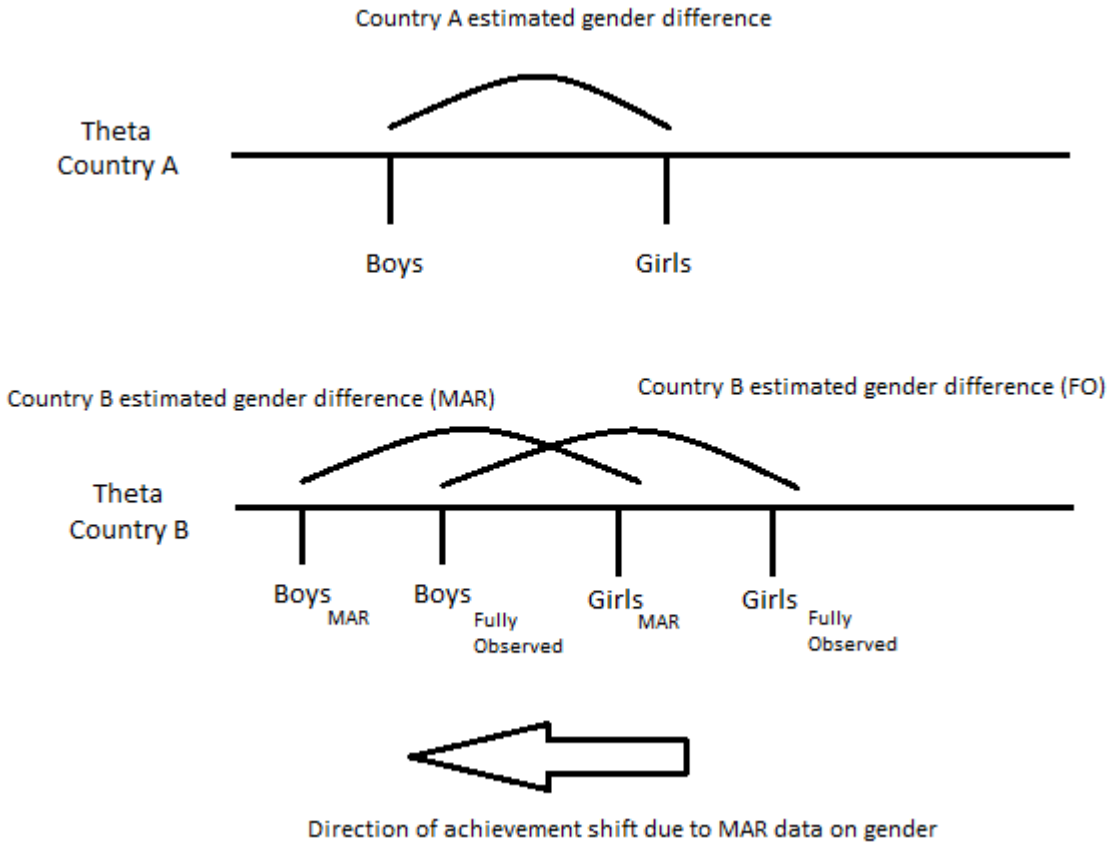
limited but controlled context, the impact of missingness on subpopulation achievement. In particular, test data were simulated under a balanced-incomplete block spiraled design, whereby each "examinee" was administered only part of the total test items. Further, background characteristics and ability values for each examinee were also generated. Based on the fully observed background data, a number of MAR conditions were created, from 10 to 20% on each of three background variables. Using a population model and an IRT model, population- and sub-population level achievement was estimated for each of the data conditions. Finally, the results of 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 missingness of up to 20% MAR.

Although the current methods used to estimate subgroup differences appear quite robust to MAR background data, this analysis found that shifts in achievement estimates occurred for two of three background variable's levels. In other words, the process of setting as missing a set number of responses to a single background variable had the effect that achievement for subgroups within that background variable were shifted, in generally similar amounts and similar directions across levels of a background variable, along the theta continuum. 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. On the other hand, 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 B, is of policy

relevance for a researcher. The researcher might come to unfounded conclusions in the face of high levels of missing data in a given educational system, say Country B. That is, an unspecified shift in performance due to missing data on the gender variable would cause our hypothetical researcher to conclude that boys from Country A were performing at higher or 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 the current analysis showed that these shifts can occur in both directions or not at all. Further, the magnitude of the shift, should it occur, is also variable. And, given the state-of-the-art in achievement estimation, these quantities are, at present, unknown.

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



Findings from this preliminary analysis into the issue of less-than-optimal background data beg caution on the part of testing organizations, 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, it seems reasonable that, at the very least, international 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, respectively.

This study examined the impact on subpopulation achievement when one background variable in the conditioning model had varied levels of missing at random data. As with any simulation, the conditions, and therefore the generalizability of the study, are limited. Further work in this area is necessary to understand better the impact of poor quality background data on subpopulation estimation. Increasing the rates of missingness or including more variables with missing data in the conditioning model would provide a clearer and 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 varied levels of data that are missing not at random. It is reasonable to expect that the impacts might be more severe given that under this condition background variables are missing due to levels of that background variable. For example, missing rates on the SES variable might be systematically missing for students from low SES backgrounds. If low SES levels are associated with poorer achievement, subgroup

differences might be misestimated and the shifts in achievement observed in this study might persist or manifest in entirely different ways.

References

SPSS (Version 16.0) [Computer software and manual]. (2007). Chicago: SPSS.

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

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

Little, R., & Rubin, D. (1983). On jointly estimating parameters and missing data by maximizing the complete-data likelihood. *The American Statistician, 37*(3), 218-220.

Lord, F. (1962). Estimating norms by item-sampling. *Educational and Psychological Measurement, 22*(2), 259-267.

Master, G. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

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*(2), 133-161.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, (17)*2, 131-154.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm, *Applied Psychological Measurement, 16*(2), 159-176.

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

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

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

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

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.