Article

SAGE Open
January-March 2018: 1–12

DOI: 10.1177/2158244018757584
journals.sagepub.com/home/sgo
SAGE

Multiple Imputation for Dichotomous MNAR Items Using Recursive Structural Equation Modeling With Rasch Measures as Predictors

Celeste Combrinck[1], Vanessa Scherman[2], David Maree[1], and Sarah Howie[1]

Abstract

Missing Not at Random (MNAR) data present challenges for the social sciences, especially when combined with Missing Completely at Random (MCAR) data for dichotomous test items. Missing data on a Grade 8 Science test for one school out of seven could not be excluded as the MNAR data were required for tracking learning progression onto the next grade. Multiple imputation (MI) was identified as a solution, and the missingness patterns were modeled with IBM Amos applying recursive structural equation modeling (SEM) for 358 cases. Rasch person measures were utilized as predictors. The final imputations were done in SPSS with logistic regression MI. Diagnostic checks of the imputations showed that the structure of the data had been maintained, and that differences between MNAR and non-MNAR missing data had been accounted for in the imputation process.

Keywords

Missing Not at Random (MNAR) data, multiple imputation (MI), Rasch person measures, structural equation modeling (SEM), dichotomous or binary items, social science methods, modeling missing data

Perfect data sets do not exist in the real world, and missing data are an authentic challenge facing social science analysts and researchers. Missing values can bias analyses, especially when high percentages are missing or there are patterns in the missingness (Allison, 2002; Osborne, 2013; Wang, Bartlett, & Ryan, 2017). The higher the percentage of missing values, the greater the potential problems (Bennett, 2001; Osborne, 2013). Consequently, the handling of missing data has been a topical issue in social sciences and methods dealing with missing values have grown exponentially (Enders, 2010; Li, Stuart, & Allison, 2015; Little & Rubin, 2002; Rubin, 1987; van Buuren, 2012). Treating missing data as incorrect responses and excluding cases, which is a common practice in large-scale assessments, could lead to significantly biased item parameter estimates (Hohensinn & Kubinger, 2011; Rose, von Davier, & Xu, 2010). Model-based approaches are recommended for handling missing data that provide the opportunity to consider the likelihood of responding and ability (Peugh & Enders, 2004; Mayer, Muche, & Hohl, 2012; Yucel, 2011).

As methods for handling missing data become easier to access, their limitations, including the evaluation of imputations and reporting, should be given more attention (Cox, McIntosh, Reason, & Terenzini, 2014; Graham, 2012; Little & Rubin, 2002; van Buuren, 2012). Some forms of missing data, such as Missing Not at Random (MNAR) data, warrant further research, which complicates the use of missing data handling techniques as data missing randomly is an assumption of many imputation methods (Galimard, Chevret, Protopopescu, & Resche-Rigon, 2016). In longitudinal research, large-scale assessments, high-stakes studies, and research designs that gather sensitive data, missing data are particularly problematic, and could affect statistical validity (Mallinckrodt, Roger, et al., 2013; Peng, Harwell, Liou, & Ehman, 2003). The most popular and recommended methods for handling missing data, such as multiple imputation (MI) and maximum likelihood (ML) estimation, were originally developed for continuous variables with the assumption of a normal distribution for data Missing at Random (MAR) or data Missing Completely at

[1]University of Pretoria, South Africa
[2]University of South Africa, Pretoria, South Africa

Corresponding Author:
Celeste Combrinck, Centre for Evaluation and Assessment, Faculty of Education at the University of Pretoria, Library Building Room 2-13, Corner of George Storrar & Leyds, Groenkloof, Pretoria 0002, South Africa.
Email: celeste.combrinck@up.ac.za

Random (MCAR). Some studies have found that when missingness mechanisms are investigated, MI can be used for MNAR data (Baraldi & Enders, 2010).

This article reports on a practical application of MI using structural equation modeling (SEM) to model the missingness of MNAR dichotomous data using Rasch person measures as predictors. The data contained a combination of MNAR and MCAR dichotomous test anchor items. Data are MNAR when missingness on a variable is directly related to the outcome variable (e.g., science proficiency; Enders, 2010; Graham, 2012; Kim & Shao, 2014; van Buuren, 2012). When data are missing due to underlying patterns and variables in the model, the mechanism of missingness is nonignorable (Wang et al., 2017). The percentage of missing data can be used to guide the selection of methods to handle the missingness: With less than 5% of missing values, listwise or pairwise deletion is an option as long as MAR data are present (Allison, 2002). Greater percentages of missing data may cause bias in analyses and should be investigated, and other options such as MI should be considered (Mallinckrodt, Lin, & Molenberghs, 2013; McPherson et al., 2015; Roberts, Sullivan, & Winchester, 2017).

Current Study

The study included seven independent high schools in South Africa. The schools are parts of a coalition and are sponsored by a funding agent. The funding agent required a set of year-end assessments to evaluate curriculum knowledge and to ensure that all students in the schools had achieved the same standards. Science assessment instruments were designed for Grades 8 to 11. During the yearly assessment of the eighth-grade students, one school received a copy of the Science test that did not contain the anchor (common) items. This was due to a printing error. The missing data for the anchor items could not be classified as missing randomly as they were completely missing only for that school. The missing data can be classified as MNAR because their absence from the test was unintended and not part of a planned design. Furthermore, the school with the missing data were different from the other six schools, as it had consistently higher score averages in all subjects. Using data from the other schools in the sample to predict the missing data for the seventh school would have led to the underestimation of achievement.

Treating data that are missing due to a specific variable, in this case one school, as MAR could have a biased effect on the imputation (Cleophas & Zwinderman, 2012; Fielding et al., 2008). The assumption of missingness has to be carefully investigated and conducting sensitivity analysis to assess the accuracy of the imputations is essential (Fielding et al., 2008; Keene, Roger, Hartley, & Kenward, 2014; McPherson et al., 2015). According to Roberts et al. (2017), "Patterns of missingness dictate how data should be analyzed" (p. 10). Based on this specific case of missingness (MNAR), the study aimed to investigate and answer three research questions:

Research Question 1: How can MNAR missing data be imputed by modeling the missingness?
Research Question 2: Which type of model and variables would best predict the MNAR data and how would the variables be identified?
Research Question 3: What contribution could be made by Rasch scores in comparison with raw scores to build more accurate MI models for missing item responses?

Method

To link the assessments within each subject from one year to the next, common items (anchor items) were included in the assessment design. The eighth-grade tests were of particular importance as they served as a baseline assessment of student ability and knowledge. The nine anchor items were present in the tests completed by the other six schools but not the seventh school. This meant crucial items that would be used for anchoring within the cohort were missing for one school. For the school in which the items were not included in the test, the data were MNAR and were directly related to variables in the data set, namely, that of school and science proficiency for tracking (anchoring). It was not possible to exclude the missing data as it was crucial to have responses to the items for anchoring in the subsequent year's assessment. A hybrid approach was explored to handle the missing data. Hybrid approaches are recommended for strengthening methods for handling missing data (Aste, Boninsegna, Freno, & Trentin, 2015).

Participants and Ethical Considerations

A total of 358 Grade 8 students from seven independent high schools completed the Science assessment at the end of the academic year. Parents signed consent forms for testing participation, as well as for the results to be used for research purposes. The average age of the Grade 8 students was 15.53 years with a greater number of female participants (71.79%) than males. The sample included a girls-only school, which accounts for the larger proportion of females in the sample. The MNAR school, which received the test copies without the anchor items, accounted for 18.15% of the total sample (65/358).

Instruments and Procedures

The assessments were designed to cover the South African Curriculum and Assessment Plan Statement (CAPS) and measure the knowledge gained over the course of a year (Department of Basic Education, 2012). The Science assessments were administered at the seven schools at the end of each academic year and were conducted using standardized procedures, with external evaluators conducting the testing processes at each school. Examination conditions were maintained during the assessments.

Table 1. Variable Summary of Missing Data for Imputation Items MCAR and MNAR.

	Missing both MCAR and MNAR		Missing MCAR		
Test questions	*n*	%	*n*	%	Valid *n*
Anchor Item Q1	68	18.99	3	1.02	290
Anchor Item Q2	73	20.39	8	2.73	285
Anchor Item Q3	106	29.61	41	13.99	252
Anchor Item Q4	108	30.17	43	14.68	250
Anchor Item Q5	87	24.30	22	7.51	271
Anchor Item Q6	99	27.65	34	11.60	259
Anchor Item Q7	76	21.23	11	3.75	282
Anchor Item Q8	76	21.23	11	3.75	282
Anchor Item Q9	68	18.99	3	1.02	290

Note. MCAR = Missing Completely at Random; MNAR = Missing Not at Random.

Data Analysis

To address the problem of MNAR data for the dichotomous anchor items, this study investigated methods to handle missing data when the mechanism for missingness is known. IBM SPSS Version 23 and IBM Amos were used in the analysis, and a practical application of modeling the missingness and imputing missing values, based on the model, was demonstrated. Rasch person measures were identified as the most suitable predictors for the missing scores. Rasch theory uses logistic regression models to estimate the likelihood of answering a question correctly and creates an equal interval logit scale for persons and items (Andrich, 2011; Bond & Fox, 2015; Dunne, Long, Craig, & Venter, 2012; Linacre, 2016; Uebersax, 1993). The Rasch models are quite resilient to missing data in general (Bond & Fox, 2015; Boone, Staver, & Yale, 2014; Linacre, 2016). Winsteps 3.75.0 was used to produce items and person estimates before the imputation (Linacre, 2016). The measures for both persons and items were rescaled from 0 to 100 to make the outputs easier to interpret. The anchor items did not form a scale and consequently could not be used as predictors; using Rasch person measures addressed this challenge.

The data contained a combination of both MNAR data (one school did not have the anchor items) and MCAR data (the other schools had the items but some students elected not to answer some items). The composition of the two missingness mechanisms is shown in Table 1. MI was chosen as the method to handle the missing data because it uses multiple values to estimate parameters and explicitly accounts for the uncertainty associated with missing data by reflecting the underlying variability (Enders, 2010; Rubin, 1976, 1987; van Buuren, 2012). MI produces continuous imputations for categorical variables if multiple linear regression is used, as opposed to logistic imputation (Cox et al., 2014). Rounding off values so that illogical values fit the original variables' scale has to be done with caution and can be especially problematic for dichotomous items (Finch, 2010, 2011; Horton, Lipsitz, & Parzen, 2003). In addition, MI can be used for MNAR data when the missingness is modeled (Dong & Peng, 2013; Horton & Kleinman, 2007; van Buuren, 2012). For ordinal data with a monotone pattern, as was the case discussed in this article, logistic regression was the preferable method for imputation (Mayer et al., 2012; Schafer, 1999b). Mayer et al. (2012) recommend using IBM Amos when the researcher knows the reason for missingness; in this way, the missingness can be modeled explicitly with SEM and the imputations will be based on the model's structure. The model can also be evaluated for fitness and refined so that the MI will have more accurate imputations, which are based on the relationships within the data. It should be noted that both the percentage of missing data and the sample size have an impact on the MI model, and when sample sizes are small, such as $n < 50$ and missing greater than 20% of values, bias can be introduced into the imputation process and results (Hardt, Herke, Brian, & Laubach, 2013).

For the current study, only the nine anchor items in the Grade 8 Science test needed to be imputed. The highest percentage of missing values in this study was 30% for one of the variables, and the sample size of 358 was judged to be adequate to estimate the missing data. When utilizing IBM Amos, Bayesian analysis is conducted for ordered categorical data and the Markov Chain Monte Carlo (MCMC) algorithm is employed to draw random values of the parameters from joint posterior distributions (Arbuckle, 2014b; Grace, 2015; Poleto, Singer, & Paulino, 2011). When dichotomous variables are used in an Amos model, additional constraints must be added to identify the model (Arbuckle, 2014a; Grace, 2015; IBM, 2015). As MI uses regression to predict outcomes, SEM is the next natural step, and more complex relationships among imputation variables can be specified. All of the anchor items to be imputed were dichotomous and functioned as endogenous variables in the model. For each item, the residual mean and variance were fixed as 0 and 1, respectively (Arbuckle, 2014b). Dichotomous variables have only one boundary, and to determine the origin and underlying scale required for the variable, parameter constraints must be imposed. The constraints act as priors, restricting the dichotomous variables to a range of 0 to 1. Further priors were not added in this study, as the aim was to analyze the model for use in SPSS, and uninformative priors were recommended for this purpose (Grace, 2009).

Modeling the MI in Amos allowed for the testing of several possible models, as well as refining the model to obtain a model best suited for MI. SPSS was used to conduct the final MI after modeling was completed in Amos, and the most appropriate model was used to specify the imputation in SPSS. The main reason for using SPSS for the final imputations was that SPSS has logistic regression MI as an option, which produces categorical variables within the correct ranges. In contrast, Amos Bayesian multiple regression results produce variables on a wider, continuous scale and

this requires additional formulae for rescaling and rounding (Graham, 2012). Using Amos to assess the model produced statistics such as regression weights and the posterior predictive *p* value (Nguyen, Lee, & Carlin, 2015). SPSS outputs limited statistics once the MI process has been conducted, making it challenging for users of the MI function to assess the statistical validity of their model and imputations (IBM, 2012). In addition, SPSS does not provide an iteration history for categorical variables (IBM, 2014), which is where Amos proves to be a more advantageous tool. The variables for use in the MI model were identified using Pearson's correlation coefficient (*r*) to assess the strength of the relationships. Possible predictors for the MI process were investigated and included auxiliary variables such as gender, school, and age, as well as other items in the instrument, the imputation variables, and the composite (total test) scores. Only variables with small to large significant correlations found in the imputation variables were used in models for predictive power and improved model functioning.

After the MI process had been performed in SPSS, sensitivity analysis was conducted to compare nonimputed data with the imputations. In this step, the original anchor items were compared with the imputed variables using the McNemar and Kruskal–Wallis tests to determine whether the original items differed statistically from the imputed variables when the results were pooled (Schafer, 1999a). Using a hybrid approach by combining SEM in Amos with MI in SPSS led to a stronger imputation model, as the advantages of each program were utilized and their limitations were negated.

Results

Table 1 shows the percentage of missing data per item, first showing the percentage of missing values per item for all types of missingness (MCAR and MNAR), and then illustrating the percentage of missing only for MCAR. Anchor Item 4 has the largest percentage of missing data at 30% of values missing (valid n = 250/358), with 14.68% of those being MCAR data. All anchor items were dichotomous and for MCAR and MNAR data combined, 62.57% of cases and 82.29% of values were complete. For MNAR data only, 76.45% of cases and 93.33% of values were complete. The MCAR data accounted for 6.67% of missing values, whereas the MNAR mechanism explained 11.04% of missing data (overall 17.72% of the values were missing). A monotone pattern of missingness was identified due to data being missing for one school in particular (IBM, 2013; Rezvan, Lee, & Simpson, 2015). Little's MCAR test of the data for schools where the items were completed confirmed that the missing values were MCAR for the other schools, χ^2 = 195.269, df = 166, p = .06 (within SPSS Version 23; Little, 1988). This established that the data contained a combination of MCAR and MNAR data. A listwise deletion of all missing data would thus result in 37% of cases being excluded.

Auxiliary Variables

Literature on building missing value models indicates that including auxiliary variables could be very beneficial for imputation (Crameri, von Wyl, Koemeda, Schulthess, & Tschuschke, 2015; Manly & Wells, 2015; Nguyen et al., 2015). The advantages of auxiliary variables are dependent on significant correlations (>.40) with the imputation variables, as well as lower percentages of missing values for the auxiliary variables (Dong & Peng, 2013; Enders, 2010). In this study, three auxiliary variables were considered: gender, which did not correlate significantly with any of the imputation variables or predictor variables; then, age was considered but it correlated weakly with only one of the imputation variables; and finally, school membership was assessed, with membership in the seventh school functioning as constant as it was completely missing for the MNAR school. For reasons cited above, none of the demographic variables were included in the model. MI can be robust to application without auxiliary variables when viable alternative imputation variables are utilized (Mustillo & Kwon, 2015).

Predictor and Imputation Variables in the Model

The anchor items had correlations with one another, ranging from nonexistent (r = .002) to weak (r = –.273), with a principal components analysis showing that the nine anchor items did not form a factor. Using the anchor items to predict missing values on one another was not recommended, and when a saturated model was attempted, it failed to converge (Poleto et al., 2011, had similar findings when using a saturated model). The other test items in the assessment were also considered; however, they only correlated well with some of the anchor items to be imputed and thus could not be used as a group. Inclusion of all 79 items in the model may have led to the overcomplication of the model (Hardt et al., 2013). As a result, two types of composite (total test) scores were considered: raw composite scores and Rasch person measures. The use of item composite scores creates a variable that contains all the information from items without MNAR data to predict MNAR variables. Correlations among anchor items and raw composite variables showed small to moderate significant correlations to the raw composite scores with five out of the nine anchor items (correlations ranged from r = .128 to r = .424). The Rasch person measures of all items had small to moderate correlations with six out of the nine anchor items (r = .133-.360). The Rasch person measures variable, based only on anchor items, had small to large correlations with all anchor items (r = .173-.667).

The SEM Model of Missingness

The SEM of missingness, built for this particular study, was specified as a recursive model. The imputation variables

Table 2. Standardized Direct Effects on Imputation Items.

Test questions	Pred1 *M*[a]	Pred2 *M*[b]	Maximum	Minimum	Lower bound 50%	Upper bound 50%	*SD*	*SE*	Convergent statistic
Anchor_Q1	0.402	0.099	0.639	0.125	0.355	0.452	0.072	0.001	1.000
Anchor_Q2	0.468	0.051	0.748	0.072	0.414	0.525	0.081	0.001	1.000
Anchor_Q3	0.282	−0.128	0.605	−0.054	0.219	0.346	0.093	0.001	1.000
Anchor_Q4	0.427	0.136	0.688	0.099	0.379	0.480	0.077	0.001	1.000
Anchor_Q5	0.772	0.259	0.890	0.604	0.748	0.799	0.038	0.000	1.000
Anchor_Q6	0.873	0.069	0.941	0.732	0.857	0.891	0.026	0.000	1.000
Anchor_Q7	0.512	0.285	0.736	0.217	0.469	0.558	0.065	0.001	1.000
Anchor_Q8	0.440	−0.083	0.677	0.150	0.394	0.488	0.071	0.001	1.000
Anchor_Q9	0.549	−0.179	0.768	0.175	0.502	0.599	0.074	0.001	1.000

[a]Rasch person measures for anchor items only.
[b]Rasch person measures for all items.

were the anchor items, nine items which were all dichotomous, with 18% to 30% of the values missing in a combination of MCAR and MNAR mechanisms. The predictor variables included the Rasch person measures of the anchor items, as well as the Rasch person measures of all test items (excluding the anchor items). The standardized direct effects (regression weights) indicated good predictive power for each of the imputation variables, see Table 2. Both, Predictors 1 and 2, worked well in the model due to their moderate correlation with each other at $r = .352$ ($p = .000$), as well as small to large significant correlations with the imputation variables. In Amos, five imputations were generated to assess the model, and a total of 10 imputations were generated in SPSS (Dong & Peng, 2013; Schafer, 1999a; White, Royston, & Wood, 2011).

Figure 1 displays the visual representation of the recursive model, which converged after 10,000 observations. Van Buuren (2012) suggests using the most simplistic model for MNAR data, with the result that the model below is both the simplest and most accurate that could be devised with the data.

The posterior predictive *p* had a value of .02, which indicates a lack of good fit between the data and the model, as the ideal value should be closer to .5 (IBM, 2014; Nguyen et al., 2015). However, the *p* value is only one indication of model functioning and is subject to factors such as percentage of missing values and sample size and thus should be treated with caution (Gelman, 2013). Further checks of the model and convergence were done by examining the histograms with first and last distributions. A sample of the histograms, trace plots, and autocorrelation plots are displayed in Figures 2 to 7. The histograms show that the first and last distributions from the analyses are closely aligned and almost equal, an indication that the posterior distributions were successfully identified and modeled.

Figures 4 and 5 illustrate the trace plots or time-series plots and indicate that the MCMC procedures converged quickly. There were no long-term trends or drifts, only minimal fluctuations.

The autocorrelation plots, depicted in Figures 6 and 7, show high initial correlation and then small or no correlation by 100 iterations, the point at which the model converged.

A pseudo-R^2 was calculated by using the formula as recommended by Grace (2009) for use in MCMC models in Amos: $R^2 = 1 - (e1 /$ implied variance of predictor variables; see also Grace & Bollen, 2005). This yielded $R^2 = .737$ for the SEM model, showing that the overall model accounted for a large percentage of variance.

Diagnostic Checks of the Imputation Model

A comparison of the original data and the pooled data for the 10 imputations is presented in Table 3. No imputed item recorded a statistically significant difference between the original item and the pooled item, demonstrating the accuracy of the imputation. As recommended by Schafer (1999b), the McNemar test was used to compare preimputation with imputed binary variables. The Kruskal–Wallis test was also reported to use the imputation numbers as grouping variables, so that each imputation was compared with every other imputation in this analysis. Both the McNemar and Kruskal–Wallis tests showed that there was no statistically significant difference between the original data and the imputed values, or among the imputations ($p > .01$).

When the imputed items were imported into Winsteps, the item measures were found to have remained very stable, with each imputed anchor item correlating above .9 with the original anchor item measures (estimates; Suarez Enciso, 2016). Each imputation was imported into Winsteps separately, and the outputs were produced and compared. The school with data completely missing on the anchor items (MNAR data) tended to have higher performance on the test overall ($M = 49.91$, $SE = 0.48$, $n = 65$) when compared with the other six schools ($M = 43.84$, $SE = 0.36$, $n = 293$). In the MI model, this was accounted for by including the overall Rasch person estimates. The MNAR school had a higher

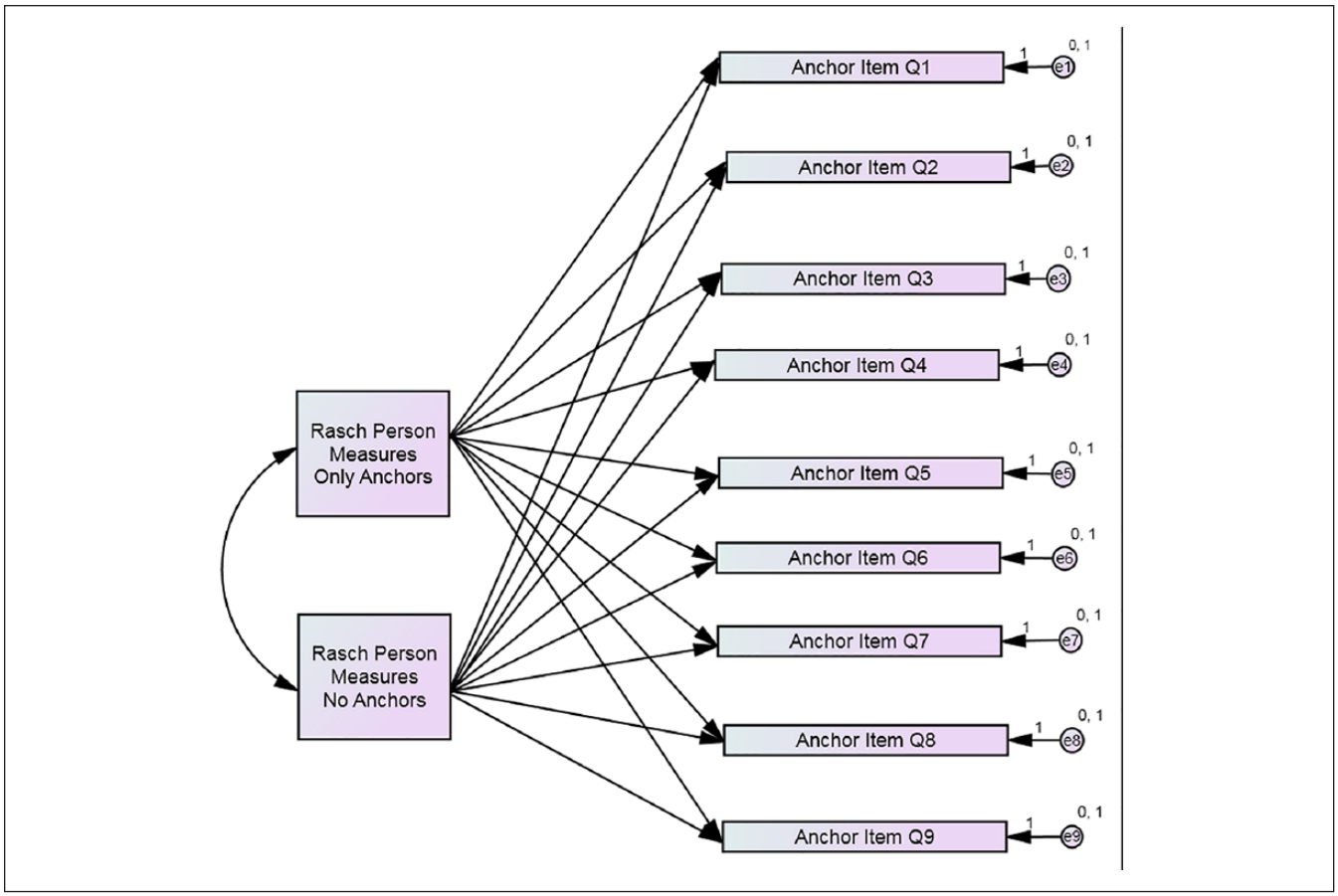

Figure 1. IBM Amos recursive model for imputing missing data.

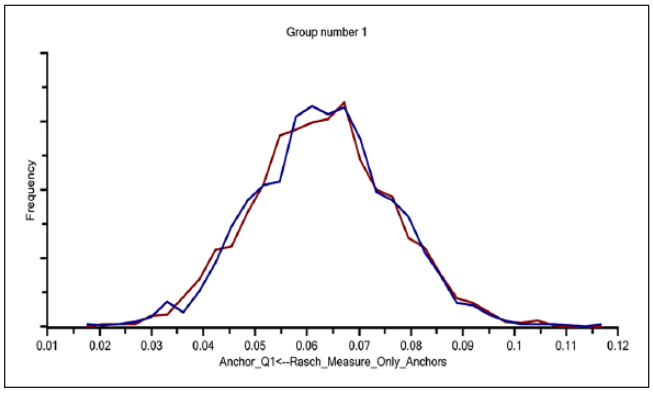

Figure 2. Q1 Histogram Predictor 1.

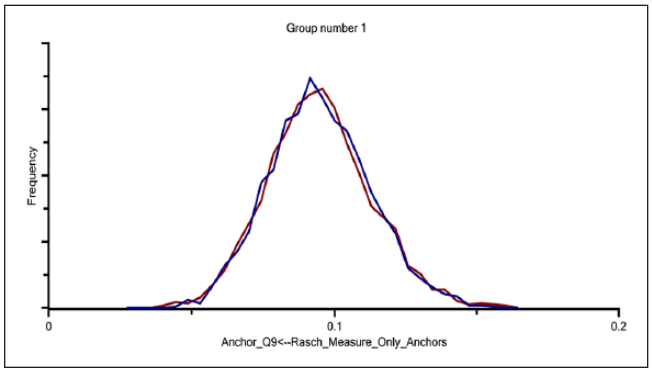

Figure 3. Q9 Histogram Predictor 1.

mean for the pooled MI anchor items ($M = 46.02$, $SE = 2.79$, $n = 65$), which was 4.5% higher than that of the other schools ($M = 41.45$, $SE = 0.44$, $n = 293$). Notably, the standard error was much higher for the MI data of the MNAR school than that of the other schools. Figure 8 also demonstrates graphically how the mean of Rasch person measures for all items with no imputation compares with the multiple imputed anchor items' person mean.

Discussion

The quality of analyses and findings improves when researchers acknowledge missing data, investigate the reasons thereof, and actively find ways to deal with the missing values (Carpenter, Bartlett, & Kenward, 2010; Manly & Wells, 2015; Peng et al., 2003). By building SEM models with Bayesian analysis to find the best model and assess the convergence, an MI model could be structured for

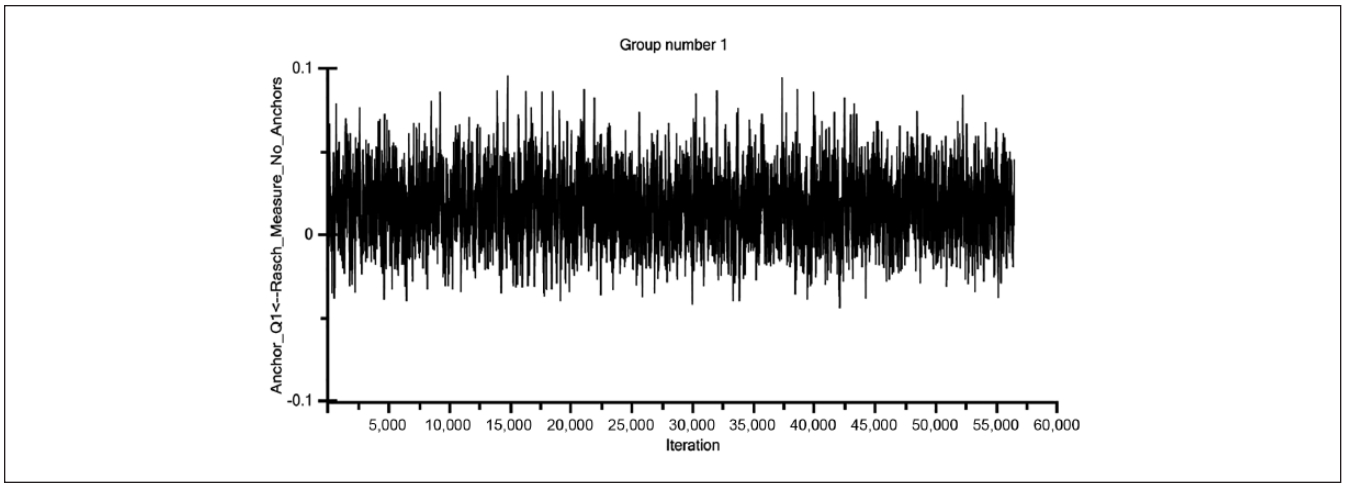

Figure 4. Q1 Trace Plot Predictor 1.

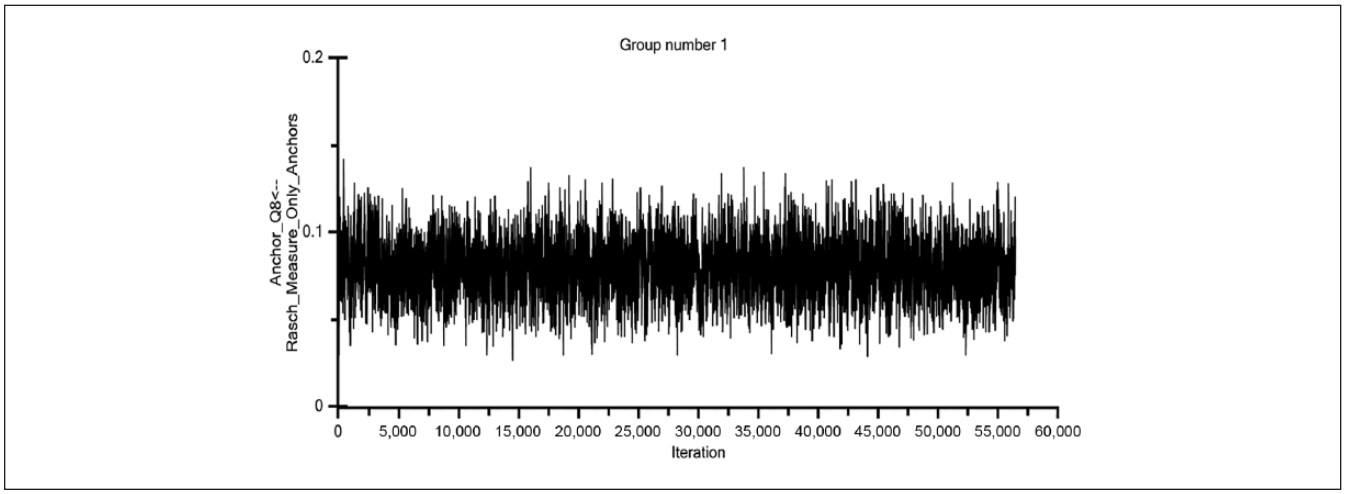

Figure 5. Q8 Trace Plot Predictor 1.

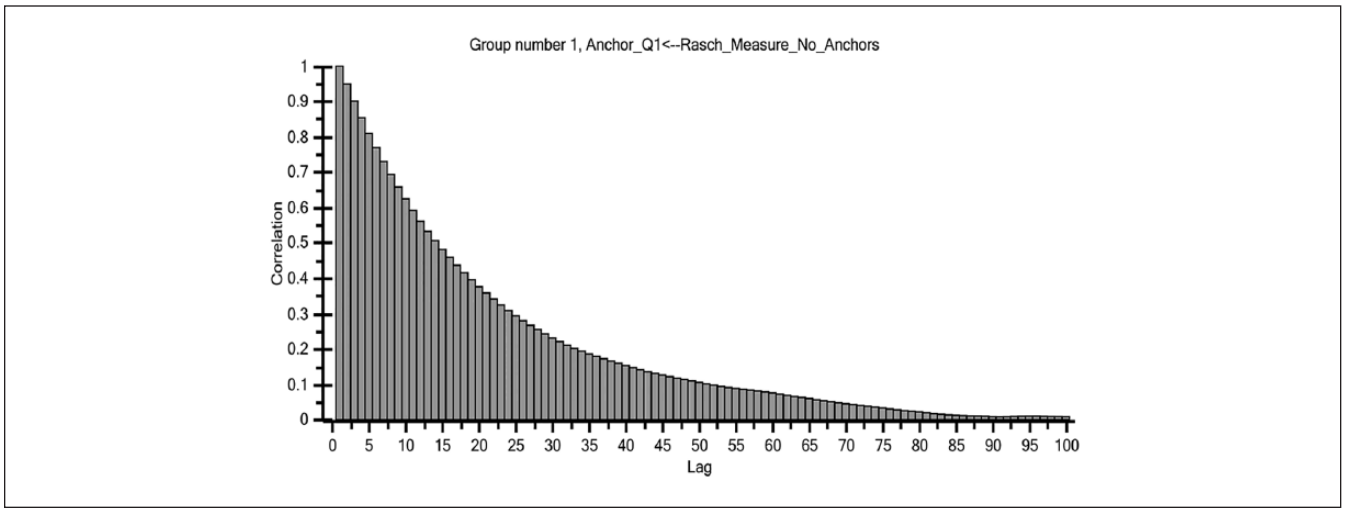

Figure 6. Q1 AutoCorrelation Predictor 1.

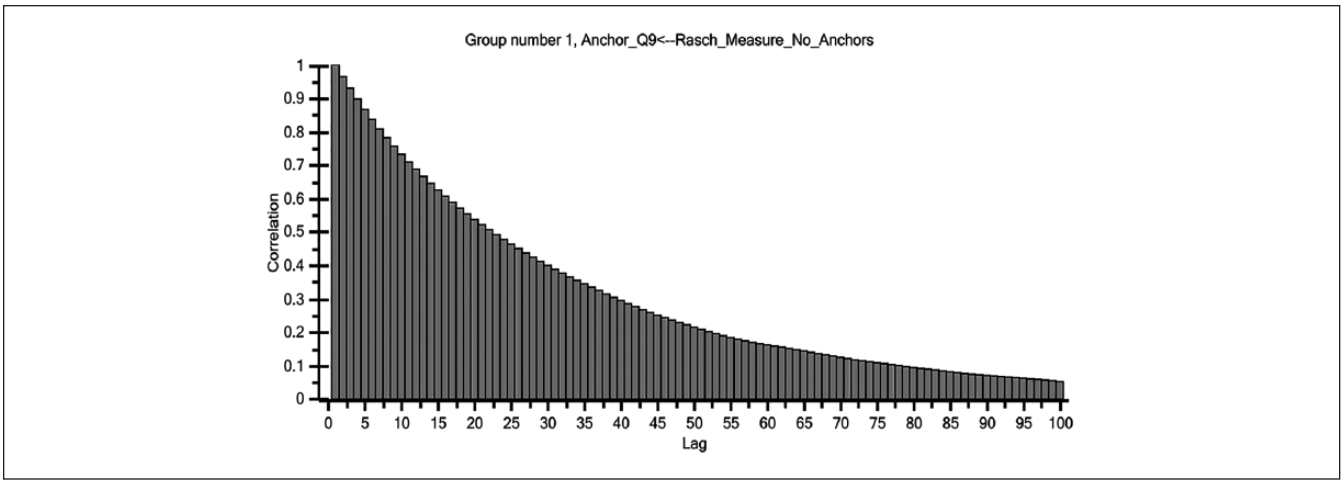

Figure 7. Q9 AutoCorrelation Predictor 1.

Table 3. Original Data Compared With Pooled Data.

Test questions	Original *n*	Original *M*	Original *SE*	Pooled *M*	Pooled *SE*	χ^2	Asymptotic significant[a]	Exact significant (two-tailed)[b]
Anchor Item Q1	290	0.272	0.026	0.297	0.027	2.828	.985	1.000
Anchor Item Q2	285	0.179	0.023	0.204	0.028	7.142	.712	1.000
Anchor Item Q3	252	0.202	0.025	0.215	0.032	9.938	.446	1.000
Anchor Item Q4	250	0.304	0.029	0.340	0.037	11.377	.329	1.000
Anchor Item Q5	271	0.472	0.030	0.489	0.030	2.488	.991	1.000
Anchor Item Q6	259	0.467	0.031	0.473	0.032	3.517	.967	1.000
Anchor Item Q7	282	0.298	0.027	0.332	0.031	5.493	.856	1.000
Anchor Item Q8	282	0.294	0.027	0.316	0.030	4.749	.907	1.000
Anchor Item Q9	290	0.172	0.022	0.207	0.030	9.886	.451	1.000

[a]Kruskal–Wallis test, $df = 10$.
[b]Binomial distribution used for McNemar.

imputations in SPSS using logistic regression. Using IBM Amos, a recursive model was built with Rasch person measures as predictors and had one predictor based on all items and the other on the anchor items only. The result was that the Rasch person measures were better predictors compared with using composite scores of the raw values, and as a result, the recursive model worked best. Other models were attempted, such as a saturated model and possible path models with either both or one of the Rasch person measures or with the raw composite scores, but these models either failed to converge or had poor model fit. Poor fit for the other models is attributed to the fact that the imputation variables did not form a scale. Finding good predictors for MI models should be done by identifying variables that correlate moderately to highly with the imputation variables. For this reason, imputation items alone could not be utilized in the model. This is where the Rasch scores were useful, because they created an estimate of persons, which was on an equal interval scale, thus providing a more accurate measure. The Rasch measures also provided significant correlations with the imputation variables while maintaining the pattern of performance between the MNAR school and the other schools with MCAR missing data.

IBM Amos provided a way to model the missingness and check model functioning. In addition, Amos produced continuous imputation variables for MI. Considering the problems caused by rounding off variables and the importance of using the correct MI method for the type of variables imputed, the model constructed in Amos was used as the guideline for imputation in SPSS (hybrid approach). Checks of the multiple imputed variables from SPSS showed that they maintained the structure of the original variables with similar means, standard deviations, and standard errors. The imputed variables were not statistically different from the original variables ($p = 1.000$). The MNAR school had a higher mean for all items in comparison with the other schools in the test. This pattern was maintained by the MI model, with the multiple imputed anchor items of the MNAR school having a 4.5% higher mean than that of the other schools after imputation. This is similar to the original pattern of the other items

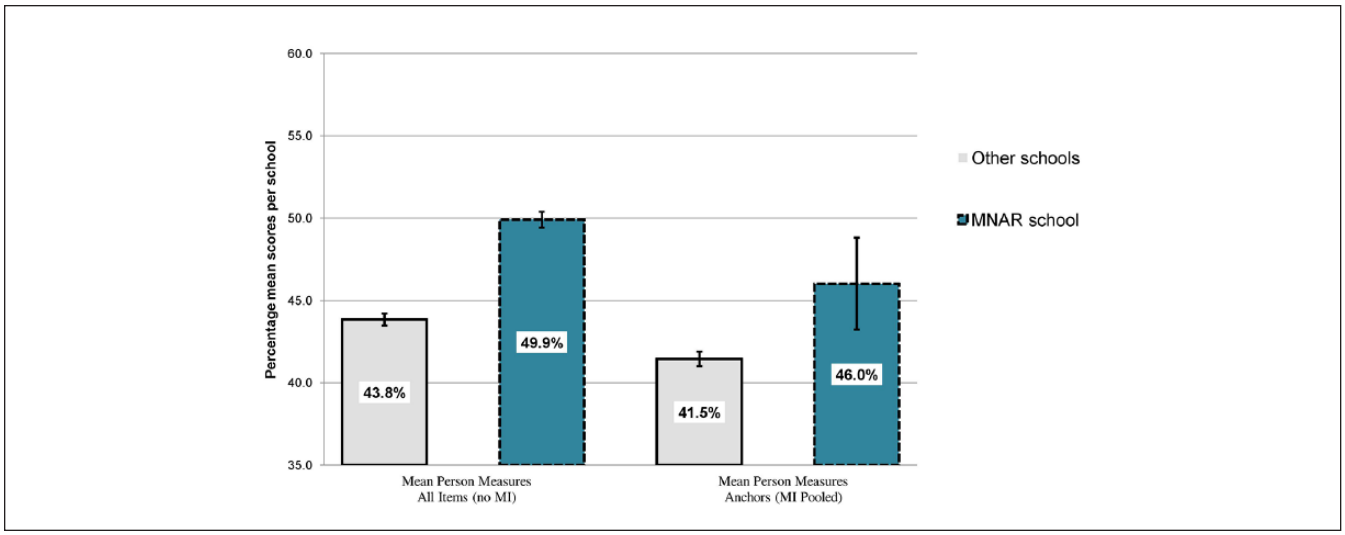

Figure 8. Mean person measures for all items (no MI) versus MI pooled.
Note. MI = multiple imputation; MNAR = Missing Not at Random.

in the assessment, which had on average 6.0% higher means for the MNAR school items (see Figure 8). The impact of the MI on the measurement model was also investigated, and it was found that the item and person parameters remained stable and highly correlated with the original estimates. However, for the MNAR school, it should be noted that the imputations increased the standard error.

Summary and Conclusion

MI is mainly conducted by assuming data are MAR or MCAR and the imputation variables are used as both predictors and imputed variables. Imputations are often conducted without checking the accuracy of the predictors (Kim & Shao, 2014; Osborne, 2013). Factors such as the missingness mechanism, the strength of the predictors in the MI model, variable types for imputation (such as dichotomous items), and ways to improve the MI model should be considered and the statistical validity should be strengthened. MNAR data are especially challenging to handle and this article is one demonstration of how to take important factors into consideration and to use MI for dichotomous MNAR items. Other applications have been carried out with different types of MNAR data and in various disciplines using a variety of approaches (see, for example, Galimard et al., 2016; Poleto et al., 2011; Wang et al., 2017). The study described in this article adds value due to its realistic set up, and demonstrates a single application of MI using SEM to model the missingness and Rasch scores as predictors. If missingness can be modeled, then the best identified model can be used to specify the imputation process. The usefulness of Rasch scores as predictors was also explored, as well as the impact of MI values on the measurement model. The following steps were used and could be considered for similar studies:

1. The missingness mechanism was known, and correlations of demographic and other variables with the missing values were calculated to find potential predictor variables.
2. The MNAR data were modeled with SEMs to find the model that best predicted the missingness mechanism.
3. Predictors were identified by calculating the correlations among imputation variables, as well as composite scores (outcome variables) and demographic variables. Only predictors that had significant correlations with the imputation variables were used in the model. Rasch scores were used as they had higher correlations with the anchor items than raw total test scores (imputation variables).
4. Logistic regression MI was utilized for the dichotomous anchor items.
5. The imputation model was checked statistically by comparing the imputed variables with the original. For the SEM, the convergence statistics, goodness of fit, and other indicators such as graphs and plots were checked for evidence of convergence and goodness of fit. A pseudo-R^2 was calculated for the model.
6. The measurement model was assessed by comparing how the imputations affected the item and person parameters.

MI has become less complicated to apply, particularly with the availability of statistical programs. Thus, the onus rests on the researcher to investigate the underlying assumptions before applying MI and finding the most accurate models with which to predict the missing data (Fielding et al., 2008). It also highlights the importance of strong predictors in MI models and checking the imputation model after imputations have been completed.

This study was conducted on a relatively small sample (n = 358), and it is suggested that larger studies with more dichotomously scaled items or ordered categorical variables could expand knowledge in this area. Several methods are available to deal with MNAR data, including many different software packages (Mayer et al., 2012). It is recommended that researchers learn how to handle missing data with software they are familiar with and that they should examine the advantages and disadvantages of their software for imputing missing data. Researchers should take into consideration assumptions of imputation models, limitations, and sensitivity analyses when handling missing data. More research is needed in educational and psychological disciplines so that guidelines can be established for imputing data for special cases, especially where anchor items are concerned, as well as for MNAR dichotomous test items.

Authors' Note

Opinions expressed and conclusions arrived at are those of the authors and are not necessarily to be attributed to the South African National Research Foundation (NRF).

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) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The financial assistance of the South African National Research Foundation (NRF) toward this research is hereby acknowledged.

ORCID iD

Celeste Combrinck https://orcid.org/0000-0002-8067-5299

References

Allison, P. D. (2002). *Missing data (Quantitative applications in the social sciences)*. Thousand Oaks, CA: Sage.

Andrich, D. (2011). *Rasch models for measurement* (A Sage university papers series: Quantitative Applications in the Social Sciences, No. 07-068). Newbury Park, CA: Sage.

Arbuckle, J. L. (2014a). Amos (Version 23.0) [Computer program]. Chicago, IL: IBM SPSS.

Arbuckle, J. L. (2014b). Amos 23*.0 user's guide*. Chicago, IL: IBM SPSS.

Aste, M., Boninsegna, M., Freno, A., & Trentin, E. (2015). Techniques for dealing with incomplete data: A tutorial and survey. *Pattern Analysis and Applications*, *18*, 1-29.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, *48*, 5-37.

Bennett, D. A. (2001). How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health*, *25*, 464-469.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (Vol. 3). New York, NY: Routledge.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. London, England: Springer.

Carpenter, J., Bartlett, J., & Kenward, M. (2010). *Guidelines for handling missing data in social science research*. Available from www.missingdata.org.uk

Cleophas, T., & Zwinderman, A. (2012). *Statistics applied to clinical studies* (5th ed.). Dordrecht, The Netherlands: Springer.

Cox, B. E., McIntosh, K. L., Reason, R. D., & Terenzini, P. T. (2014). Working with missing data in higher education research: A primer in real-world example. *Review of Higher Education*, *37*, 377-402.

Crameri, A., von Wyl, A., Koemeda, M., Schulthess, P., & Tschuschke, V. (2015). Sensitivity analysis in multiple imputation in effectiveness studies of psychotherapy. *Frontiers in Psychology*, *6*, Article 1042.

Department of Basic Education. (2012). *National Protocol for Assessment in Grades R-12*. Pretoria, South Africa: Government Printing Works.

Dong, Y., & Peng, C.-Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, *2*, Article 222.

Dunne, T., Long, C., Craig, T., & Venter, E. (2012). Meeting the requirements of both classroom-based and systemic assessment of mathematics proficiency: The potential of Rasch measurement theory. *Pythagoras*, *33*(3), 16-35.

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

Fielding, S., Fayers, P. M., McDonald, A., McPherson, G., Campbell, M. K., & RECORD Study Group. (2008). Simple imputation methods were inadequate for Missing Not at Random (MNAR) quality of life data. *Health and Quality of Life Outcomes*, *6*, Article 57.

Finch, W. H. (2010). Imputation methods for missing categorical questionnaire data: A comparison of approaches. *Journal of Data Science*, *8*, 361-378.

Finch, W. H. (2011). The use of multiple imputation for missing data in uniform DIF analysis: Power and type I error rates. *Applied Measurement in Educational Assessment*, *24*, 281-301.

Galimard, J. E., Chevret, S., Protopopescu, C., & Resche-Rigon, M. (2016). A multiple imputation approach for MNAR mechanisms compatible with Heckman's model. *Statistics in Medicine*, *35*, 2907-2920.

Gelman, A. (2013). Two simple examples for understanding posterior p-values whose distributions are far from uniform. *Electronic Journal of Statistics*, *7*, 2595-2602.

Grace, J. B. (2009). Intro to Amos Bayesian SEM and MCMC estimation: Markov Chain Monte Carlo (MCMC) algorithm approach to estimating models. In *Tutorials using the Amos software*. Retrieved from http://www.structuralequations.com/AmosTutorials.html

Grace, J. B. (2015). SE modeling when some response variables are categorical: The special case of binary (dichotomous) variables. In *Modeling with structural equations*. Retrieved from http://www.structuralequations.com/resources/BinaryResponseModeling.Mar30_2009.pps

Grace, J. B., & Bollen, K. (2005). Interpreting the results from multiple regression and structural equation models. *Bulletin of the Ecological Society of America*, *86*, 283-295.

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

Hardt, J., Herke, M., Brian, T., & Laubach, W. (2013). Multiple imputation of missing data: A simulation study on a binary response. *Open Journal of Statistics*, *3*, 370-378.

Hohensinn, C., & Kubinger, K. D. (2011). On the impact of missing values on item fit and the model validness of the Rasch model. *Psychological Test and Assessment Modeling*, *53*, 380-393.

Horton, N. J., & Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, *61*, 79-90.

Horton, N. J., Lipsitz, S. R., & Parzen, M. (2003). A potential for bias when rounding in multiple imputation. *The American Statistician*, *57*, 229-232.

IBM. (2012). *Method: Multiple imputation* (Statistical Package for the Social Sciences [SPSS] Support Centre). Retrieved from https://www.ibm.com/support/knowledgecenter/en/SSLVMB_21.0.0/com.ibm.spss.statistics.help/idh_idd_mi_method.htm?view=embed

IBM. (2013). Released 2013: IBM SPSS statistics for Windows (Version 23.0230). Armonk, NY: Author.

IBM. (2014). IBM SPSS missing values 23. Armonk, NY: Author.

IBM. (2015). Bayesian estimation. In *IBM SPSS Amos for structural equation modeling*. Retrieved from http://amosdevelopment.com/features/bayesian/index.html

Keene, O. N., Roger, J. H., Hartley, B. F., & Kenward, M. G. (2014). Missing data sensitivity analysis for recurrent event data using controlled imputation. *Pharmaceutical Statistics*, *13*, 258-264.

Kim, J. K., & Shao, J. (2014). *Statistical methods for handling incomplete data*. New York, NY: CRC Press.

Li, P., Stuart, E. A., & Allison, D. B. (2015). Multiple imputation: A flexible tool for handling missing data. *Journal of the American Medical Association*, *314*, 1966-1967.

Linacre, J. M. (2016). *Winsteps® Rasch measurement computer program user's guide*. Beaverton, OR: Winsteps.com.

Little, R. J. A. (1988). A test of Missing Completely at Random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*, 1198-1202.

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

Mallinckrodt, C., Lin, Q., & Molenberghs, M. (2013). A structured framework for assessing sensitivity to missing data assumptions in longitudinal clinical trials. *Pharmaceutical Statistics*, *12*, 1-6.

Mallinckrodt, C., Roger, J., Chuang-Stein, C., Molenberghs, G., Lane, P. W., O'Kelly, M., . . . Thijs, H. (2013). Missing data: Turning guidance into action. *Statistics in Biopharmaceutical Research*, *5*, 369-382.

Manly, C. A., & Wells, R. S. (2015). Reporting the use of multiple imputation for missing data in higher education research. *Research in Higher Education*, *56*, 397-409.

Mayer, B., Muche, R., & Hohl, K. (2012). Software for the handling and imputation of missing data: An overview. *Clinical Trials*, *2*(1), 1-8.

McPherson, S., Barbosa-Leiker, C., Mamey, M. R., McDonell, M., Enders, C. K., & Roll, J. (2015). A "Missing Not at Random" (MNAR) and "Missing at Random" (MAR) growth model comparison with a buprenorphine/naloxone clinical trial. *Addiction*, *110*, 51-58.

Mustillo, S., & Kwon, S. (2015). Auxiliary variables in multiple imputation when data are Missing Not at Random. *The Journal of Mathematical Sociology*, *39*, 73-91.

Nguyen, C. D., Lee, K. J., & Carlin, J. B. (2015). Posterior predictive checking of multiple imputation models. *Biometrical Journal/Biometrische Zeitschrift*, *57*, 676-694.

Osborne, J. (2013). *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. Thousand Oaks, CA: Sage.

Peng, C.-Y. J., Harwell, M., Liou, S.-M., & Ehman, L. H. (2003, April). *Advances in missing data methods and implications for educational research*. Paper presented at the 2002 Chinese American Educational Research and Development Association Conference, Taipei, Taiwan.

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

Poleto, F. Z., Singer, J. M., & Paulino, C. D. (2011). Missing data mechanisms and their implications on the analysis of categorical data. *Statistics and Computing*, *21*, 31-43.

Rezvan, P. H., Lee, K. J., & Simpson, J. A. (2015). The rise of multiple imputation: A review of the reporting and implementation of the method in medical research. *BMC Medical Research Methodology*, *15*, 30-60.

Roberts, M. B., Sullivan, M. C., & Winchester, S. B. (2017). Examining solutions to missing data in longitudinal nursing research. *Journal for Specialists in Pediatric Nursing*, *22*(2), Article e12179.

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with Item Response Theory (IRT)* (Report No. ETS RR-10-11). Princeton, NJ: Educational Testing Service.

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: John Wiley.

Schafer, J. L. (1999a). Multiple imputation: A primer. *Statistical Methods in Medical Research*, *8*, 3-15.

Schafer, J. L. (1999b). *NORM users' guide* (Version 2). University Park: The Methodology Center, Penn State. Available from http://methodology.psu.edu

Suarez Enciso, S. M. (2016). *The effects of missing data treatment on person ability estimates using IRT models* (Master's thesis). University of Nebraska–Lincoln, Nebraska. Retrieved from http://digitalcommons.unl.edu/cehsdiss/274

Uebersax, J. (1993). Rasch model software and FAQ. Retrieved from http://www.john-uebersax.com/stat/rasch.txt

van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: CRC Press.

Wang, J. J. J., Bartlett, M., & Ryan, L. (2017). Non-ignorable missingness in logistic regression. *Statistics in Medicine*, *36*, 3005-3021.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, *30*, 377-399.

Yucel, R. M. (2011). State of the multiple imputation software. *Journal of Statistical Software*, *45*(1), 1-7.

Author Biographies

Celeste Combrinck is a researcher at the Centre for Evaluation and Assessment (CEA), faculty of Education at the University of Pretoria. She is a social scientist who works in the field of education. Her academic interests include test design and development

(applying the Rasch model), structural equation modeling (SEM), and research design and methodology in the social sciences.

Vanessa Scherman is a registered research psychologist and professor at the Department of Psychology of Education at the University of South Africa (UNISA). She has worked in the fields of psychometrics and school effectiveness for 18 years. She supervises master's and doctoral students in the areas of assessment, quality assurance, school effectiveness research, psychometrics, and evaluation.

David Maree has worked in the field of psychometrics since 1995. In 1997, he became director of the unit for test development at the Human Sciences Research Council (HSRC). In 1999, he moved to the University of Pretoria as a full professor in the Department of Psychology. He is a specialist in many areas of psychology, including psychometrics, cognitive psychology, research methods, Rasch modeling, critical realism, and the philosophy of science.

Sarah Howie is a well-known leader in the field of large-scale educational assessment and was the National Research Coordinator (NRC) of the Progress in International Reading Literacy Study (PIRLS) in South Africa for three rounds: 2006, 2011, and 2016. She was previously NRC for the Trends in Mathematics and Science Study (TIMSS) 1995 and 1999, and contributed internationally to PISA 2015 and 2018.