8-2016

THE EFFECTS OF MISSING DATA TREATMENT ON PERSON ABILITY ESTIMATES USING IRT MODELS

Sonia Mariel Suarez Enciso
University of Nebraska-Lincoln, marielsuaren@gmail.com

THE EFFECTS OF MISSING DATA TREATMENT ON PERSON ABILITY ESTIMATES USING IRT MODELS

by

Sonia Mariel Suarez Enciso

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Master of Arts

Major: Educational Psychology

Under the Supervision of Professor Rafael De Ayala

Lincoln, Nebraska

August, 2016

THE EFFECTS OF MISSING DATA TREATMENT ON PERSON ABILITY ESTIMATES USING IRT MODELS

Sonia Mariel Suarez Enciso, M.A.

University of Nebraska, 2016

Adviser: Rafael De Ayala

Unplanned missing responses are common to surveys and tests including large scale assessments. There has been an ongoing debate on how missing responses should be handled and some approaches are preferred over others, especially in the context of the item response theory (IRT) models. In this context, examinees' abilities are normally estimated with the missing responses generally ignored or treated as incorrect. Most of the studies that have explored the performance of missing data handling approaches have used simulated data. This study uses the SERCE (UNESCO, 2006) dataset and missingness pattern to evaluate the performance of three approaches: treating missing as incorrect, midpoint imputation, and multiple imputation with and without auxiliary variables. Using the Rasch and 2PL models, the results showed that treating missing as incorrect had a reduced average error in the estimation of ability but tended to underestimate the examinee's ability. Multiple imputation with and without auxiliary variables had similar performances to one another. Consequently, the use of auxiliary variable may not harm the estimation, but it can become an unnecessary burden during the imputation process. The midpoint imputation did not differ much from multiple imputation in its performance and thus should be preferred over the latter for practical reasons. The main implication is that SERCE might have underestimated the student's ability. Limitations and further directions are discussed.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

ACER	Australian Council for Educational Research
AIC	Akaike Information Criterion
AIC3	Modified Index of AIC
BIC	Bayesian Information Criterion
CES	Classroom Environment Study
CIM	Corrected Item Mean Substitution
CIM-E	Corrected Item Mean Substitution with normally distributed error
CIVED	Civic Education Study
CM	Mean conditional on the covariates
COMPED	Computers in Education Study
EAP	Expected A Posterior
EM	Expectation-Maximization
EV	Expected Value
FCS	Fully Conditional Specification
FIML	Full Information Maximum Likelihood
FIMS	First International Mathematics Study
FISS	First International Science Study
FR	Fractional imputation
GPCM	Generalized Partial Credit Model
HDD	Hot-Deck Deterministic
HDNC	Hot-Deck Next Case
HDNN	Hot-Deck Nearest Neighbor
HDR	Hot-Deck Random
IAS	Incorrect Answer Substitution
ICS	Item Correlation Substitution
IMS	Item Mean Substitution
IRF	Item Response Function
IRT	Item Response Theory
JMLE	Joint Maximum Likelihood Estimation
LC MI	Latent-Class Multiple Imputation
LD	Listwise Deletion
LLECE	Latin American Laboratory for Assessment of the Quality of Education
MCM	Multiple-Choice Model
MCM-MI	Multiple Imputation with MCM
MCMC	Marcov Chain Monte Carlo
MF	MissForest
MI	Multiple Imputation
MIDA	Multiple Imputation with Data Augmentation
MICE	Multivariate Imputation by Chained Equation

ML	Maximum Likelihood
MMLE	Marginal Maximum Likelihood Estimation
MRD	Multiple Random Draws
MRF	Mean Response-Function Imputation
NLPCA	Nonlinear Principal Component Analysis
NP	Not Presented
NRM	Nominal Response Model
OECD	Organisation for Economic Co-operation and Development
OM	Overall Mean
OREALC	Regional Bureau of Education for Latin America and the Caribbean
PCM	Partial Credit Model
PERCE	First Regional Comparative and Explanatory Study
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
PMS	Person Mean Substitution
PPP	Preprimary Project
RDS	Random Drawn Substitution
RF	Response-Function Imputation
RM	Random Mean
RMSD	Root-Mean-Square Deviation
RSMCM	Restricted Samejima-Multiple Choice Model
SERCE	Second Regional Comparative and Explanatory Study
SI	Single Imputation
SMCM	Samejima-Multiple Choice Model
SRD	Single Random Draw
TIMSS	Trends in International Mathematics and Science Study
TW	Two-Way Imputation
TW-E	Two-Way Imputation with normally distributed error
UNESCO	United Nations Educational, Scientific and Cultural Organization

ACKNOWLEDGMENTS

I want to express my deepest gratitude to my Adviser, Dr. Rafael De Ayala, for his professional guidance and support from the very beginning of my days in this journey. The confidence he has put on me has been of invaluable relevance. He has been not only a mentor, but also a role model and an amazing professor. My gratitude to Dr. Charles Ansorge, who has become an important figure in my life and has contributed with my teaching experience in an innovative and challenging manner. I thank to all the Quantitative, Qualitative, and Psychometric Methods Faculty and staff for all the academic experience; especially to Dr. Wayne Babchuk for his support. Also thanks to Dr. Eric Buhs for the in-field opportunity experience. My sincere thanks to Patti Farritor, who enlightens my days with her beautiful smile and has been an amazing friend.

Thanks to my friends from everywhere, the new and the old ones. My gratitude also to the Fulbright Program for this learning experience in the U.S. My gratefulness to the LLECE team, especially to Mauricio Holz. Thanks to my family, the foundation of my world. To you, papito, for being my star and the inspiration of my life. Thanks for all the days of love and hard work you have put into my family. To you mamita, because every day you are there for me and for everybody. To Emilce, Dinie, Aldo, J. Antonio, J. Armando, Nasser, and Adecia.

CHAPTER I: INTRODUCTION

Large-scale achievement assessment originated in 1922, with the implementation of the Stanford Achievement Test (Kelley, Ruch, & Terman, 1922). Since then other large-scale tests have appeared and assessments have been extended to comparing performance across countries. International achievement assessments such as the Trends in International Mathematics and Science Study (TIMSS), the Progress in International Reading Literacy Study (PIRLS), and the Programme for International Student Assessment (PISA) are among the most well-known international large-scale achievement tests. These three surveys have been administered to elementary and high school students in different domains. The majority of the participant countries in these tests are European and Asian with only few being from Africa or Latin-America. More recently, the United Nations Educational, Scientific, and Cultural Organization (UNESCO, using its Spanish acronym) supports a relatively new effort to measure education quality in Latin-American countries at the elementary level in mathematics, reading and writing, and sciences.

All the above instruments measure cognitive skills across nations. Their main goal is to determine the performance of educational systems and the position of the countries in this matter. To assess the quality of education requires establishing comparability of scores between countries. Score comparability is connected to validity and test fairness. As such, lack of bias and equality in outcomes of testing are needed in order to guarantee quality of interpretation (Standards for Educational and Psychological Testing, 2014). One of the factors that contributes to bias is unplanned missing data or missingness

observed in the data collection stage. In large-scale assessment, and in any other survey, not all the examinees will provide answers to all the items. This may be because they do not know the answer or are unsure about their answer, they do not have time to answer all the items, they inadvertently skip one or more, or simply due to individual characteristics (e.g., risk-aversion, self-confidence, test-wiseness, test-taking behavior, etc.). Missingness reduces the sample size, and thus affects the representativeness of the population, the accuracy of parameter estimates (Mislevy & Wu, 1988, 1996), and the generalizability of inferences.

The literature shows that the proportion of the missing data, the nature of the missingness (i.e., missingness mechanisms) during data collection, and the way it is addressed in statistical analysis yield different results (e.g., Allison, 2006; Dong & Peng, 2013; Enders, 2010; Lord, 1973, 1980; Mislevy & Wu, 1988; Rubin, 1976; Little & Rubin, 1987; Schafer, 1997). For example, Mislevy and Wu (1988, 1996) assert that treating missing data as incorrect downwards the inferences about person ability. As will be explained soon, how missingness is handled can yield biased parameter estimates and distort both the final results and the quality of the inferences.

Missing data and Ignorability

Missingness mechanisms. Missing data can be either planned or unplanned. If the missing data are planned they are said to be missing by design. They are due to the researcher's decision and under the researcher's control. Data are missing due to characteristics or design of the instrument (e.g., adaptive tests), or are associated with the cost-effectiveness of the measurement (i.e., to save time and/or money). Consequently,

these missing responses can be ignored with negligible consequences on the inferential analysis. This type of the missingness can be modeled using either maximum likelihood or multiple imputation techniques (Enders, 2010). When the missing data are unplanned, they may or may not be ignored, depending on the cause of the missingness.

Rubin (1976) and Little and Rubin (1987) introduced a taxonomy based on the missingness nature (ignorable or nonignorable). According to them, that there are three different missingness mechanisms. They depend on the probability of response conditioned to the outcome, some covariate variables, or both. Data are missing completely at random (MCAR) if the nonresponses are independent of the variable being measured. When the nonresponse is conditional on any other variable (covariate) except the outcome, the data are missing at random (MAR). Both, MCAR, and MAR are considered ignorable missingness for "likelihood-base inference" (Little & Rubin, 1987, p.15). Therefore, it is possible to get unbiased parameter estimates working only with the part of the data that has no missingness (when data are MCAR), or considering the conditional distribution if needed (when data are MAR).

A last mechanism assumes that the data are missing not at random (MNAR). That is, the probability of response is conditional on both outcome and covariate variables. In this case, the missing data are systematic or non-random. MNAR is also known as nonignorable missingness, because the analysis of only the part of the data that are complete produces biased estimates (Little & Rubin, 1987; Rubin, 1976). For example, when dealing with the estimation of item parameters under IRT statistical programs generally assume that the missingness is ignorable. Violation of this assumption biases

item and person parameter estimates. Similarly, incorrectly treating missingness may overlook the differential item functioning (DIF) for different participating groups (Emenogu, Barnabas, Falenchuk, & Childs, 2010).

Unfortunately, there are not many techniques to test the nature of missingness. MCAR can be tested using Little's test for MCAR (Little, 1988). The plausibility of MAR could be checked with a *t*-test of the mean difference between the group of participants with and without nonresponses (Diggle, Liang, & Zeger, 1995; as cited in Dong & Peng, 2013; Tabachnick & Fidell 2012). Nevertheless, even when it may be evident that the missingness is conditional on some variables (i.e., MAR), it is practically impossible to verify that the missing data are not related also to the variable under measure (Dong & Peng, 2013).

Missing data handling methods. There are different ways to deal with missing data. They vary according to the missingness mechanisms and the researcher's willingness to work with complete or incomplete data. Appendix A shows the different techniques. In *complete data analysis* only the cases that do not have missing responses are used. This approach assumes that the missingness is ignorable (MCAR). The incomplete data analysis implies working with the whole collected data, including missing responses. In this case, the missing data are assumed to be random missingness (i.e., either MCAR or MAR) most of the times. Methods to treat nonresponses can be classified in deductive, deterministic, and stochastic. The method is deductive when the missing values are imputed using additional logic information. If a predicted or specific value is used, then the method is deterministic, and it is stochastic if randomness is

incorporated into the process.

In *listwise deletion* (LD), observations with missing responses are discarded. The analysis is then done only with the complete cases. The advantages of this approach is simplicity and comparability (Little & Robin, 1987). The disadvantages are the loss of information due to reduction of sample size, and biased sample estimates if the mechanisms is not MCAR. When the observations are eliminated according to the variables to be studied the analysis is said to be available-case or *pairwise deletion*. The advantage of this technique is that the sample loss is less than in listwise deletion. The drawback is the variability of the sample size from variable to variable, which depends on the missingness pattern (see Enders, 2010 for description about the patterns).

Single Imputation (SI) methods have shown to yield biased parameters estimates, even under MCAR (Enders, 2010). Appendix A lists the different SI techniques. The *unconditional mean imputation* (or mean substitution, or arithmetic mean imputation) consists of taking the mean of the available data and assigning that value to all the missing responses. This approach "systematically underestimates variances and covariances" (Little & Rubin, p.44). *Person mean imputation* (or averaging the available items or prorating a scale score) can be used when the researcher wants to work with scale scores instead of working with item responses. That is, the scale score is computed by taking the average (or the sum) of the items with responses. Although more studies need to be conducted to check on the disadvantages of person mean imputation, "it may produce biased parameter estimates [even] with MCAR data" (Enders, 2010, p.51). Other mean-based imputation techniques are explained in more detail in the next chapter.

Regression imputation (also called Buck's method or conditional mean imputation) assumes that the missing data are conditional on the observed variables. Therefore, the missing values for each variable are estimated by a linear regression where Y is the variable with missing values and X_i are other variables in the dataset. The disadvantage of this method is that the imputed value will be the same for all the cases that have the same predictor values. Due to that lack of variability, this method overestimates the correlation and the R^2 statistic (Enders, 1999, 2010). This approach also underestimates variances and covariances, although in a smaller degree than the mean-based imputation (Enders, 2010; Little & Robin, 1987).

Another alternative is the *stochastic regression imputation*. In this method, the regression equation includes a residual term. This term is randomly selected from a set of numbers normally distributed with mean zero and variance equal to the residual variance from the model (if the criterion is a continuous variable). The variability issue is taken into account with this additional element in the imputation process. This approach produces unbiased estimates even under MAR assumption. Nonetheless, it still underestimates sampling error, increasing type I error. Also, it becomes complex with multivariate missingness (Enders, 2010; Little & Robin, 1987).

Hot-deck imputation is a set of techniques that imputes the missing values with scores from similar respondents (Enders, 2010). There are several versions of this approach (e.g., random hot-deck, the deterministic hot-deck, hot-deck nearest neighbor, etc.). However, the idea behind them is the same. They replace the missing values with the observed values from respondents that share the same characteristics (matching

variables) in the dataset. This method preserves the data distribution and does not artificially make the data leptokurtic, because it maintains the data variability (Enders, 1999, 2010). It does not rely on model fitting and avoids cross-users inconsistencies between imputation and data analysis (Andridge & Little, 2010). It, however, underestimates the sampling error (Enders, 2010). Moreover, it is not convenient for estimating measuring association, because this imputation approach affects the correlation and regression coefficient estimates (Schafer & Graham, 2002). Other disadvantages are: (a) it does not have a theoretical foundation; (b) the likelihood of still having missing data at the end of the procedure is high because donors may not be found; and (c) when the missingness is present in more than one variable, then the order in which the variables are imputed can affect subsequent imputation (Enders, 2010).

Another matching-case method is the *cold-deck imputation*, which is similar to hot-deck imputation. In this case, the imputed data come from a different dataset against which missing values are matched. *Similar response pattern imputation* also shares some commonalities with the hot-deck imputation. It also uses matching variables between the complete and incomplete cases. In this method, the complete case that minimizes the standardized difference of the matching variables between the two sets (i.e., complete and incomplete) will donate its value to the incomplete one. If there is more than one donor, then the average of them is the imputed value. This approach works fine when data are MCAR (Brown, 1994; Enders, 2001; Enders & Bandalos, 2001; Gold & Bentler, 2000; as cited in Enders, 2010), but the bias could be substantial when data are MAR (Enders,

2010). The disadvantages for these two approaches are the same as for the hot-deck imputation.

Last observation and *worse observation carried forward* are used in longitudinal measures. In the former case, after participant's dropout, the last observed value in each variable is repeated for the rest of waves (i.e., points of measure in the time) in the study. In the latter case, the lowest registered value of each variable is repeated for the rest of the waves in the study. Different studies have shown that the parameter estimates are biased (under-or over-estimated) even with MCAR data (Cook, Zeng, & Yi, 2004; Liu & Gould, 2002; Mallinckrodt, Clark, & David, 2001; Molenberghs Thijs, Jansen, & Beunckens, Kenward, Mallinckrodt, et al., 2004; as cited in Enders, 2010) due to the distortion of mean and covariance structure (Carpenter, Bartlett & Kenward, n.d.).

Maximum likelihood (ML) estimation is a set of approaches recommended by the literature (e.g., Enders, 2010; Rubin, 2014; Schafer & Graham, 2002) because it produces unbiased parameter estimates with smaller standard errors, even under MAR condition (Enders, 1999, 2010). It is also superior to other techniques when data are MCAR. ML estimates the parameters through an iterative process in which several values are tried until the estimates that yield the highest log-likelihood value are found. With ML concrete imputation is not needed. Instead, missing data become part of the input in the log-likelihood estimation. The main assumption with ML approaches is the normality of the data. However, violation of this assumption slightly impacts the parameter estimates. Although it distorts the likelihood ratio test and biases the standard error (Enders, 2010).

An extensively used technique within the ML approach is the expectation-maximization (EM) algorithm proposed by Dempster, Laird, and Rubin (1977). This algorithm is a two-stage iterative process. In the first stage (the expectation or *E*-step) sufficient statistics of the unobserved data are estimated based on the observed data. In the second stage (the maximization or *M*-step), the unknown parameter are estimated by maximum likelihood by treating the estimated sufficient statistics as observed. This stage yield a set of estimated parameters that feed the next cycle (*E*-step) where again sufficient statistics are estimated. The difference of the estimate parameter values between each cycle is evaluated and the iteration stops when convergence is reached (i.e., the difference is smaller than a value set a priori) (Dempster, Laird, & Rubin 1977).

All the previous missingness treatment approaches are framed within the frequentist paradigm. *Multiple imputation* (MI) is framed within an alternative paradigm: the Bayesian estimation. MI also refers to a collection of techniques and is currently regarded as an efficient approach for treating missing data along with ML estimation. Both approaches can deal with almost all the different missing data patterns (Enders, 2010). MI makes the same assumptions as the ML: the MAR condition and the normality distribution of the data. However, MI differs from ML in that MI gives the complete-data condition to any dataset with missing responses (Little and Rubin, 1987). Therefore, MI allows standard analysis methods to be applied to the now complete dataset.

Compared to SI techniques, MI has shown to be better because SI does not consider sampling variability in the imputation process. That is, most of the SI techniques replace the missing responses with the same value and do not take into account the

uncertainty about what the true response could be (i.e., only a single plausible value is considered). In contrast, MI considers both sampling variability and uncertainty. MI provides several plausible values for each missing response, which implies various complete-data sets on which the analysis is conducted (Little & Rubin, 1987).

Enders (2010) describes MI as a three-phase process. During the first phase (i.e., imputation phase), several versions of complete-data are created, each of them with different missing response estimates. This phase heavily relies on Bayesian principles. The second phase (i.e., analysis phase) is the easiest part of the MI approach. Here, each dataset is analyzed with standard complete-data methods. Thus, this phase yields as many estimates of parameters and standard errors as the number of imputed datasets. In the last phase (i.e., pooling phase), all the parameter estimates and standard error estimates are averaged to generate only one set of results.

It is in the imputation phase where the different MI techniques differ. There are several imputation methods, two of which that are used most frequently. These are Marcov Chain Monte Carlo-based strategies for missing data imputation: the multiple imputation with data augmentation (MIDA), based on the joint modeling, and the fully conditional specification (FCS) (Enders, 2010; Lee & Carlin, 2010; van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006).

One difference between these two strategies is that MIDA specifies a parametric multivariate density which assumes a particular form of multivariate distribution. FCS does not assume such distribution, instead it specifies individual conditional density for each variable in the dataset. A second difference is that FCS does not use information

from the variable with missing values for the imputation, while MIDA does (Raghunathan, Lepkowksi, van Hoewyk, & Solenberger 2001; van Buuren, Boshuizen, & Knook, 1999; van Buuren et al., 2006).

Schafer (1997) developed models for continuous (multivariate normal imputation or data augmentation), categorical, and mixed format data using the MIDA algorithm. MIDA is a two-step iterative imputation process: the imputation (*I*-step) and the posterior (*P*-step) steps. In the *I*-step an estimate of the mean vector ($\widehat{\boldsymbol{\mu}}$) and covariance matrix ($\widehat{\boldsymbol{\Sigma}}$), based initially on the EM algorithm (Leite & Beretvas, 2010), is used to build a set of regression equations that predicts the nonresponses from the observed data. The type of data would determine the type of regression equation (e.g., logistic regression for categorical data) if specified.

These predicted responses are used to re-estimate the $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ in the *P*-step. Here, random residual errors drawn from a posterior distribution using MCMC are incorporated to the $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ elements. This modification is then carried to the *I*-step, where the new values of $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ are in turn used to generate a new set of regression equations. For each cycle a new complete dataset is generated (Enders, 2010). The EM algorithm is used only to estimate the initial parameters for the first imputation step algorithm (Leite & Beretvas, 2010).

FCS was independently developed by van Buuren et al. (1999) and Raghunathan et al. (2001) (Enders, 2010; Lee & Carlin, 2010), although its premise has been present since 1991 (van Buuren et al. 2006; van Buuren, 2007). FCS (also known as the chained equations, sequential regression imputation, regression switching or MICE[1]) uses the

Gibb sampler algorithm, which is a Marcov chain Montecarlo algorithm that imputes one variable at a time. FCS is a semi-parametric MI algorithm that specifies a conditional density $p(Y_j \mid Y_{-j}, \theta_j)$, in which the missing values in the variable Y_j are conditional on the other j-1 variables (Y_{-j}) and the model parameters (θ_j). This density is used for the imputation of y_i^{miss} given y_{-j} ($y_i^{miss} \mid y_{-j}$) with a regression model (e.g., linear or logistic regression) on the observed data (van Buuren et al., 2006).

The imputation consists of three steps. First, the posterior distribution of θ is estimated based on the observed data, $p(\theta \mid y^{obs})$. Second, a specific θ value, θ^*, is drawn from the posterior distribution. Third, a specific value, y^*, from the $p(y^{miss} \mid y^{obs}, \theta = \theta^*)$ is drawn; this represents the imputed value. Notice that unlike JM no information about y^{miss} is used to draw θ^*. The imputation normally starts with the variable with the lowest missingness level and progresses to the ones with higher missingness (Enders, 2010). The imputed value from one variable is used as a predictor in the next variable (Enders, 2010; Raghunathan et al., 2001; van Buuren et al., 2006). Once all the nonresponses in the dataset have been imputed (i.e., the first iteration is over), a Bayesian procedure is used to select a new set of regression parameters estimates. A subsequent iteration takes this set of parameters into account along with the imputed values from the previous iteration to generate new values. The cycle is repeated m times until the process generates a unique set of complete data (Enders, 2010).

FCS has advantages over MIDA (Enders, 2010; van Buuren et al., 2006). First, FCS algorithm seems to work better than MIDA with all types of data (categorical, continuous, and mixed data) under MCAR and MAR mechanisms. Second, the creation

of flexible multivariate models is easier than with MIDA. That is, a common distribution of variables within the dataset is not necessary because each variable is individually modeled according to its distribution. Third, FCS preserves original data features that sometimes are hard to keep when working with MIDA (van Buuren et al., 2006; van Buuren, 2010). Enders (2010) and van Buuren (2010) give examples, such as linking two variables to avoid logical inconsistencies or to accommodate designed survey patterns.

A fourth advantage is that that generalizations to missingness mechanisms that are different from MAR may be easier (van Buuren et al., 2006). FCS is an interesting approach because it does not require one to define the number of factors, or identify the items to the scale to which they belong. FCS does not require an assumption about the conditional independence among items nor to define the scale structure (van Buuren, 2010). Fifth, as with MIDA, FCS uses auxiliary variables for imputation and it is available in more software packages (e.g., *R*, SPSS, STATA, Mplus) than MIDA (Enders, 2010; Lee & Carlin, 2010; van Buuren et al., 2006; van Buuren, 2010).

FCS has also disadvantages. The first one is the lack of an underlying theoretical framework (van Buuren, 2007). Second, each conditional density has to be specified separately. Therefore, as the number of variables increase so does the modeling effort required. Third, "typical computational shortcut may not apply, and not much is known about quality of imputation because the implied joint distribution may not exist theoretically" (van Buuren et al. 2006, p. 1051). A fourth disadvantage is that FCS imputation does not pick up higher-order interactions, unless they are explicitly modeled in the imputation process (Vermunt, van Ginkel, van der Ark, & Sijtsma, 2008).

Fifth, convergence can only be guaranteed when compatibility of conditionals is met, which is hard to verify. According to van Buuren et al. (2006), "two conditional densities are compatible if a joint distribution exists that has the given densities as its conditional densities" (p. 1052). In other words, compatibility of conditionals refers to whether the model used to impute the nonresponses in one variable, conditional on the other variables in the dataset, is the correct "true" models (e.g., is it Y_2 conditional on Y_1, or on Y_1^2?). Simulated data analysis, however, indicates that FCS seems to be robust against incompatibility, converging normally with 5 to 20 iteration (van Buuren et al., 2006; van Buuren, 2007).

Missing data and item response theory (IRT)

Unplanned missing data in achievement assessment are generally classified into *not-reached*[2] and *omitted*. Nonresponses are not-reached if they occur due to insufficient time to complete the test. They are omitted if the participant accidentally skip the item or intentionally decided not to answer. It is expected that the not-reached responses would appear at the end of the test, assuming that the test was answered linearly, whereas the omitted responses are found "throughout the response vector, and not only at the end of it" (De Ayala, 2009, p.150).

Not-reached items can be ignored, especially for ability estimation purposes. Omitted responses, on the other hand, represent nonignorable missing data because they are related to the examinee's ability (Lord, 1973, 1980; Ludlow & O'Leary, 1999; Mislevy & Wu, 1988, 1996). That is, examinees with more knowledge about the construct under assessment tend to omit responses at a lower level than less proficient

examinees. These proficient examinees base their decision on their perception of correctness, because they have a better understanding of the measured construct. The omission is conditional to the person's proficiency (Stocking, Eignor, & Cook, 1988).

Ignoring omitted responses in IRT models may affect estimation accuracy. Examinees could improve their overall performance by only answering the items they are sure they will get right if they know their ability is estimated based on correct responses only (Lord, 1980; Mislevy & Wu, 1988). Also, ignoring omitted responses may violate the unidimensionality assumption of responses if these are found to be loading on a variable different than the one measured by the instrument (Ludlow & O'Leary, 1999).

The effectiveness of some of the approaches previously described were evaluated in different contexts, including in IRT models with findings that favor one approach over others (e.g., Huisman, 2000; Shin, 2009; van Buuren, 2010). There are, however, other deterministic and stochastic approaches that were especially developed for IRT analysis. For example, the two-stage approach (Ludlow & O'Leary, 1999) is used to handle missingness in some large-scale assessments. This method assumes that not-reached responses are ignorable for calibration phase, but not for the person ability estimation. Imputation methods such as midpoint and fractional imputation were explored and found to work well with IRT models (De Ayala, Plake, & Impara, 2001; De Ayala, 2003, 2006; Finch, 2008; Lord, 1973; Oshima, 1994). Other models incorporated the missingness as an indicator of a second dimension in the data analysis (Glas & Pimentel, 2008; Holman & Glas, 2005; O'Muircheartaigh & Moustaki, 1999; Pimentel; 2005).

Research problem and research questions

Most of the missingness treatment approaches used with IRT models were evaluated with simulated data that reproduced the missingness pattern and just few (e.g., Rose, von Davier, & Xu, 2010) used empirical data in their study. Simulated data provide a benchmark for comparison purposes because the parameters are first estimated on a complete dataset whose values are later removed to study the different missingness approaches. The missingness pattern in simulated datasets, however, may not necessarily be the same as in empirical data. Consequently, missing data approaches may perform differently with empirical data to the extent to which the missing data pattern differ from what has been done to date.

The purpose of this study is to compare the effectiveness of missingness mechanisms in the person ability estimation using IRT models on data from the Second Regional, Comparative, and Explanatory Study (SERCE, using its Spanish acronym). SERCE has been implemented in 2006 in the member countries of the Latin American Laboratory for Assessment of the Quality of Education (LLECE, using its Spanish acronym). The LLECE is a network of quality assessment systems focused on education evaluation among its Latin-American member countries.[3] It is coordinated by the Regional Bureau of Education for Latin America and the Caribbean (OREALC/UNESCO) located in Santiago, Chile. Thus, it is part of the United Nation efforts to improve the education quality. SERCE is one of the largest learning achievement study implemented in Latin America and the Caribbean. LLECE have

conducted other measures; PERCE (1997) and TERCE (2013). All these assessments (i.e., datasets and supporting materials) are available online.[4]

SERCE assessed elementary students in grades 3rd and 6th in sixteen Latin-American countries and the Mexican state of Nuevo Leon. It focuses on mathematics, reading and writing, and sciences. As other large-scale assessments, SERCE data contain both planned and unplanned missing responses. The unplanned missingness per person ranges from 2.9% to 5% per domain. Although these numbers are slightly lower than TIMSS or PIRLS[5], only half the data are complete.

Missingness in SERCE is handled differently than in most international assessments. In most international assessments, missing data are classified into not-reached and omitted and treated differently between the item calibration and person's ability estimation stages. Using marginal maximum likelihood estimation (MMLE) during the items calibration, the not-reached are ignored (i.e., left blank) or removed whereas the omitted are scored as incorrect (PISA Technical Report, 2012; TIMSS, PIRLS Technical Report, 2011).

Person ability is later calculated using the estimated item parameters. In this stage, both not-reached and omitted responses are treated as incorrect. PISA, TIMSS, and PIRLS generate persons' scores using the plausible values approach with examinees' background information. That is, there are at least five scores per examinee. On the other hand, missingness in SERCE was not classified into not-reached and omit and both item and person parameters were simultaneously estimated using joint maximum likelihood estimation (JMLE). In this process, missing responses were treated as incorrect (Trevino,

Bogoya, Glejberman, Castro et al., 2008); however, several authors have found that this approach yields the worst estimates (Custer, Sharairi, & Swift, 2012; De Ayala et al., 2001; De Ayala, 2006; Huisman, 2000, Rose et al., 2010), especially when used with not-reached responses (Ludlow & O'Leary, 1999; Oshima, 1994).

In order to explore the extent to which this and other approaches affects the quality of the person estimates the SERCE mathematics data are used. The missingness pattern of incomplete cases in this dataset is obtained and reproduced in the part of the data that has complete responses. This approach allows both to reproduce the empirical missing data pattern at the same time as to estimate the person parameters with a complete dataset that serves as benchmark for the effectiveness comparison.

This study contributes to the literature by examining the performance of traditional missing data handling approaches using large scale assessment. Other contributions are to provide validity evidence of the approach used in SERCE and to see the extent to which students' ability level could have been biased based by the treatment approach used. This also could have impacted the participant countries relative positions in the comparative ranking created based on the assessment.

Three different missingness approaches are compared: when the missing data are treated as incorrect, when midpoint imputation is used, and when multiple imputation with and without covariates is utilized. This study aims to answer the following research questions: (a) is there a difference in the person parameter estimation associated to the missing data approach utilized? (b) does the effectiveness of missing data approaches differ when Rasch or two-parameter IRT model are used?

In the following, Chapter 2 contains the literature review regarding missingness approaches in IRT models. Chapter 3 describes the SERCE data and presents the methodology for the missing data pattern replication and the analysis of both complete and incomplete data. Chapter 4 presents the results. Person ability is estimated using both Rasch and 2PL, while item parameters are treated as fixed. In the final chapter the findings, limitations, and future research are discussed.

CHAPTER II: LITERATURE REVIEW

The common characteristic among large-scale assessments are the format of the items and the way item responses are scored. Achievement assessments may consist of items whose format is either selected or constructed-response. With the selected item format the examinee chooses a response from the item's set of responses, whereas with the constructed-response the examinee generates the answer. In either case this response is scored either dichotomously or polytomously. In a dichotomous scoring the examinee either gets the item right (score=1) or wrong (score=0). With polytomous scoring the item response is scored into one of more than two response categories or assigned a rating. For example, with a polytomously scored item you will have a correct answer, one or more partially correct responses (i.e., partial credit), as well as an incorrect response. In any case, these items are said to be categorical because the responses are limited to the number of categories that are defined by the test developer.

Examinees may respond to all of some of the items. For example, the examinee may accidentally skip one or more items, may not be able to finish the exam due to insufficient time, or may purposely decide not to respond to an item. The first and third cases can be found throughout the test and are regarded as nonignorable missingness, meaning that the nonresponse is likely to be related to the examinee's proficiency. In the second situation, the nonresponses would appear at the end of the test, assuming that the test was answered linearly. They can be considered to be ignorable missingness especially for ability estimation purposes. The missingness mechanism defines the way nonresponses should be treated (Little & Robin, 1987; Mislevy & Wu, 1988).

Shculte Nordhold, and Hoof Van Huijsduijnen (1997) group the methods for item nonresponse imputation into three categories: (a) deductive, if the missing values are imputed using other known information, (b) deterministic, if a predicted or specific value is used, and (c) stochastic, if randomness is incorporated in the process. Deductive imputation is normally employed with numerical data. It basically relies on logical or mathematical relationships between the variables with and without missing values (Eurostat, 2014). de Waal, Pannekoek, and Scholtus (2011) present the deductive imputation method adapted for categorical data following Felligi and Holt's (1976) procedure (as cited in de Waal et al., p. 308). However, because deductive imputation requires dependency across items it cannot be applied to achievement data where items may be independent from each other.

Deterministic imputation encompasses some of the missing data handling procedures presented in Chapter 1, such as unconditional mean imputation, person mean imputation, and regression imputation. With these methods, the imputed value is basically "copied or transferred" from other observed cases. On the other hand, stochastic imputation includes procedures in which uncertainty is incorporated through a randomness variable (i.e., error). Stochastic regression imputation, maximum likelihood and multiple imputation fall in this last category. Both deterministic and stochastic imputation has been used with categorical variables such as that found in cognitive assessments. A review of these approaches is presented next. Note that, throughout this chapter, all the missing data approach's acronyms have been unified for comprehensive purposes.

Deterministic imputation for categorical variables

Large-scale assessment (e.g., PISA, 2009; TIMSS, 2011; PIRLS, 2011) data analysis normally entails two stages: the psychometric scaling or item calibration and the students' proficiency estimation. This proficiency or ability within a domain is considered to be a latent variable and thus assumed to be a latent continuum (De Ayala, 2009). As such, the student proficiency represents the person location on that continuum. The student proficiency is normally represented with theta (θ) and its estimation with theta hat ($\hat{\theta}$). Different IRT models are used according to the way the items are scored. In TIMSS and PIRLS (Technical Report, 2011), this stage is based on three IRT models: the three-parameter logistic (3PL) model for the multiple choice items, the two-parameter (2PL) model for the constructed-response items that are dichotomously scored, and the partial credit model (PCM) for the items that are scored polytomously. The model used in PISA is a generalized multidimensional Rasch model, where the different dimensions of the latent variable (θ_D) are conditional on the population's characteristics (PISA Technical Report, 2012). The PCM is used for items with multiple scores and the simple logistic model is used for dichotomously scored items (Organisation for Economic Co-operation and Development [OECD], 2012).

In PISA 2012, two item calibrations are done: the national and the international. For the national calibration, unweighted data are used and all the cases are included. The omitted and not-reached data are scored incorrect (i.e., coded with zero) for the item calibration. For the international calibration, a subsample of equal size from most of the OECD participant countries (i.e., 31 countries) is selected and cases with not-reached

responses are removed. New OECD country members are not included in this stage. The proficiency score is then estimated in a second stage. In this stage, the item parameters previously estimated are taken as fixed. PISA, TIMSS, and PIRLS generate the scores using the plausible values approach with conditioning variables or covariates, such as socio-economics status or gender.

The way in which missing responses are treated varies in each of these two stages. In the calibration stage, cases with not-reached items are sometimes removed (e.g., PISA) or generally ignored (e.g., TIMSS, PIRLS) given that responses missing at random do not carry information about examinee's ability and item parameter estimation beneficiates from ignoring them (not-administered or blank) (Lord, 1974, 1980; Ludlow & O'Leary, 1999; Mislevy & Wu, 1988, 1996; Oshima, 1994). That is, speed and ability are considered to be independent (Mislevy & Wu, 1988). In the ability estimation stage, not-reached items are treated as incorrect (coded with zero), as suggested by Ludlow and O'Leary (1999). Note that Lord (1980) suggested ignoring the not-reached items even for the person ability estimation, pointing out that ability does not depend on the items administered. That is, the "examinee's θ is the same for all the items in the unidimensional pool" (pp. 182, 226).

PISA, TIMSS, and PIRLS report that they treat omitted responses as incorrect in the two stages of the data analysis. This missing responses are assumed to be dependent upon the assessed ability (Lord, 1973, 1980; Mislevy & Wu, 1988, 1996). According to Ludlow and O'Leary (1999), both person ability and item parameters are better estimated when omitted responses are treated as incorrect in both-stages. Using the Rasch model

with empirical data (n=116 students, J=50 multiple choice items with m=4 categories and missingness per item ranging between 1.7% and 32.8%) ordered by item difficulty, they demonstrated that, unlike the other approaches they studied (i.e., ignoring omitted and not-reached or treating them as incorrect), the two-stage approach does not lead to an inflated or overestimated difficulty parameter of the items located towards the end of the test, nor rewards students with a higher ability location that have a lower response rate.

Oshima's work (1994) followed the same line as Ludlow and O'Leary (1999). She evaluated alternatives to treat not-reached items, assuming independence between speededness and ability (MAR) and using the Bayesian expected a posteriori approach (EAP) to estimate person location in a 3PL IRT model. In her study (n=1000 simulees, J=60 multiple choice items, and missingness level: 5%, 10%, and 15%), she defined two different types of not-reached items: (a) blank-not-reached, when the students do not answer the items, and (b) not-reached with random responses, that refers to the items for which examinees randomly chose an option. Moreover, she investigated how these not-reached items were treated. For example, the blank-not-reached items were coded as wrong, not considered in the calibration process, or imputed with fractional value, 1/m.

She found that item parameter estimates were affected by the not-reached items proportion, the treatment of the missing data responses, and the item difficulty order in the test (i.e., from easy to hard, and ordered randomly). If the items were ordered by difficulty, item parameters were recovered better when the blank-not-reached items were excluded from the calibration process, regardless of the proportion of not-reached responses. The worse recovery was obtained when the not-reached items were treated as

incorrect. Specifically, the discrimination and difficulty parameters were underestimated at the beginning of the test and overestimated towards the end. The discrimination parameter was more affected, especially when the items were not ordered by difficulty. The guessing parameter showed the opposite trend. Oshima (1994) found that ability estimation was robust when not-reached responses were treated as incorrect, despite the not-reached response types and missingness level. Assigning fractional scores (1/*m*) to the blank-not-reached items led to better recovery of person and item parameter.

Both Ludlow and O'Leary and Oshima assumed that not-reached responses and ability are not related. Other authors have studied the opposite situation. DeMars (2002) compared JMLE and MMLE performance in the item difficulty parameter estimation, under violation of speededness and ability independence condition (i.e., when not-reached responses are related to the examinee's ability). Factors where n=2000, J=60 multiple choice items, and missingness per item level ranging between 39% and 69%. She discovered that the 1PL model with JMLE is more accurate in the item difficulty parameter recovery than MMLE, regardless of the missingness level, when data are MNAR. MMLE underestimated the item difficulty parameter in this condition. When the not-reached responses are MAR, however, MMLE is also a valid option. DeMars argues that this is due to the fact that JMLE is based only on available data, whereas MMLE relies on a prior ability distribution that is constant for all the examinees, regardless of the items they answer. This homogeneous or unified ability distribution assumption is not tenable when not-reached responses are dependent on ability (i.e., data are MNAR).

Shin (2009) complemented DeMars' work by studying the JMLE effectiveness in the theta estimation when data are MNAR and non-equivalent groups are assessed. Unlike DeMars, Shin did not differentiate between omitted and not-reached responses meaning that missingness was not only located toward the end of the test. She used two coding schemes (incorrect and blank) for missing responses with both empirical and simulated dichotomous datasets. The Rasch model and JMLE were used. Factors were test length (J=50 for the empirical study and J=20 for the simulated study), sample sizes (n=2,941 for the empirical study and n=200, 500, 1000, and 3000 for the simulated one) and level of missingness (20.6%, 10.1%, and 7.6% for the empirical study, and 7%, 10% and 20% for the simulated).

In the empirical data, Shin (2009) found that ignoring missing data or treating them as incorrect made no difference in the theta estimation. In the simulated data analysis, however, she found the opposite. Missing data coding schemes mattered when the missingness was present in anchor items of the different forms to be equated, regardless of the level of missingness for high, medium and low ability, respectively. That is, ignoring missing responses (blank) yields better results than treating them as incorrect when the missingness is MNAR and mainly observed in the anchor items that are used in the equating process. This statement becomes stronger with larger sample sizes.

Custer, Sharairi, and Swift (2012) also examined the quality of item and person parameters recovery when not-reached responses were either dependent (MNAR) or independent (MAR) of ability using the Rasch model and JMLE. They used simulated

data with items ordered by difficulty (n=500, J=40 multiple choice items). They treated missing data in three different ways: (a) ignoring (blank) omitted and not-reached responses, (b) treating omitted as incorrect and ignoring not-reached, and (c) treating both as incorrect. For the two first missingness approaches, they found that the item difficulty was recovered with the same margin of error, regardless of the missingness level (0.81%-10%; 1.62%-20% for omitted and not-reached, respectively), and the not-reached responses mechanisms (i.e., MAR or MNAR). The accuracy of this recovery, however, was directly proportional to the level of missingness. Conversely, treating omitted and not-reached responses as incorrect had the worse item difficulty recovery accuracy.

Custer et al. (2012) encountered almost the same pattern for the person ability. Ignoring both omitted and not-reached responses led to better theta recovery regardless of missingness level and not-reached responses mechanisms. The worse recovery was found when omitted and not-reached were treated as incorrect. Also, they reported that the lowest level of root-mean-square deviation (RMSD) of theta ability was obtained when the not-reached responses were independent of ability, despite the missingness treatment. Considering the direction of the ability parameter bias, ignoring the two types of missing responses led to overestimation of the true parameters, whereas the other two approaches led to underestimation of the real theta values. Custer et al. (2012) demonstrated that these findings were tenable even when the Rasch model and JMLE were applied to data originally created under the 2PL condition.

The reason why Oshima (1994) and Custer et al. (2012) findings seem to contradict Ludlow and O'Leary's (1999) conclusions may be due to the difference in

methods. Ludlow and O'Leary used a two-stage parameter estimation each including a different treatment for not-reached responses, whereas Oshima simultaneously estimated both item and person parameters. This implies that the not-reached responses were treated in the same way throughout the data analysis. As a consequence, the θ estimates may be affected by the biased item parameter.

On the other hand, De Ayala, Plake and Impara (2001) and De Ayala (2006) found that person ability estimation is more accurate (less underestimated) when the omitted data are imputed with the midpoint value (i.e., 0.5), despite the item nature (i.e., dichotomous or polytomous) and the person ability estimation methods (i.e., ML estimation or the Bayesian EAP). Both studies assumed MNAR for the omitted responses. They utilized the 3PL IRT model for the dichotomous items, and the partial credit model (PCM) and the generalized partial credit model (GPCM) for the polytomous instrument. Theta was estimated using ML, EAP, and biweight estimation with dichotomous data, whereas EAP was used with the polytomous data. Factors were sample size (n=41000 simulees for both dichotomous and polytomous data), test length (J=39 multiple choice items and J=24 polytomous items), and person-level missingness (5.1%, 10.3%, 15.4%, and 20.5% for dichotomous data and 4.2%, 12.5%, and 20.8% for polytomous data).

Contrary to Ludlow and O'Leary's (1999) findings, these authors showed that treating omitted answers as incorrect resulted in worse theta estimates for both dichotomous and polytomous items. Their findings, however, supported Custer et al.'s (2012) work. They showed that ignoring omitted responses has better results in the

person ability estimation than treating them as incorrect, although this procedure may not recommended due to reduction of sample size. The level of omission played an important role in the accuracy of θ recovery in all the cases. The higher the omission level the worse the θ recovery. Finally, they found that ML estimation is more affected than EAP when items are binary. However, likelihood-based method performed well in the θ estimation with polytomous data (De Ayala, 2006).

In both studies, the authors mentioned that the caveats of this approach are the lack of the theoretical justification for this imputation value (De Ayala et al., 2001; De Ayala, 2006) and the introduction of additional measurement error to the extent that this answer (0.5) does not approximate the student's true response. Finally, when working with rating data (e.g., Likert response scale), missing responses[6] are best handled with the hot-deck approach in first place or with the midpoint imputation in second place (De Ayala, 2003). De Ayala (2003) demonstrated that with this type of data, neither likelihood-based models nor ignoring missing responses worked well. ML estimation and the rating scale model (RSM) were used in this research (n=41000 simulees with J=15 Likert scale items and person-level missingness of 7%, 13%, and 20%).

Huisman (2000) reviewed best practices with missing responses for non-cognitive categorical data. He explored the effectiveness of nine deterministic methods: random drawn substitution (RDS), incorrect answer substitution (IAS), person mean substitution (PMS), item mean substitution (IMS), corrected item mean substitution (CIM), item correlation substitution (ICS), and three variants of the hot-deck method: the hot-deck next case (HDNC), the hot-deck deterministic (HDD), and the hot-deck random (HDR)[7].

Two different levels of parameters were evaluated: person ability and scale quality. Person ability was estimated as the weighted sum of the item scores. The scale quality was measured with the Cronbach's *alpha* and the Loevinger's *H*-coefficient. He considered different factors, such as different sample size (*n*=100, 200, 400), test length (*J*=40, 36), and number of item categories (*m*=2, 3, 5, 6 options), missingness level (5%, 12%, and 20%), and missingness mechanisms (MCAR, MAR, and MNAR).

At the person ability level, Huisman found that: (a) the effectiveness of the missing data treatment approaches were negatively related to the level of missingness and its mechanisms and positively related to the test length; (b) there was no interaction between the missing treatment methods and the sample size, meaning that no approach behaved differently due to *n*; and (c) "imputation techniques that take into account the relationships between items perform better than those that do not" (p. 345), meaning that CIM was the best technique to estimate person ability given the factors under study (*n*, *J*, *K*, missingness level, and mechanisms); (d) there was an interaction between the missing treatment methods and the number of categories per item on the person ability recovery. That is, although CIM was the best method for all the different analyzed cases, other approaches were as good as CIM depending on *K*. ICS performed well for *m*=2 or 3, whereas PMS and HDD were good for *m*=5 or 6. The worse techniques were RDS and IAS. At the scale quality level, Huisman could not identify a best technique. The effect of missing data handling procedure on the scale quality indices depended on test length, the missingness level, missingness mechanism, and *K*. However, he found that Cronbach's *alpha* was more affected than *H*-coefficient.

All the points mentioned before can be summarized in two main outcomes. First, as argued by Huisman (2000), "using information from both persons and items results in better estimates of the missing values" (p. 349). That is, imputation corrected by person ability such as CIM yields less biased estimates. Consequently, the author said that using IRT-based models for both imputation and data analysis may improve the imputation effectiveness. Second, having a "good" recovery of person ability does not necessarily mean that the item/scale quality is also well recovered. Actually, they may be overestimated if IRT-based models are employed to impute missing values and to analyze the data.

The fact that several authors (Custer et al., 2012; De Ayala et al., 2001; De Ayala, 2006; Huisman, 2000) have independently demonstrated the poor performance of treating omitted responses as incorrect reinforces what Mislevy and Wu (1988) said:

> Supplying incorrect responses for omits leads to a "marginal conditional" MLE for θ under the assumption that responses to omitted items would surely have been incorrect. This may be reasonable for open-ended items, but it is not plausible for multiple-choice items for which even the least able examinees have nontrivial probabilities of success. In these cases, supplying incorrect responses for omits would bias estimates of θ downward. (p. 41).

Stochastic imputation for categorical variables

Multiple imputation and maximum likelihood. ML estimation and MI are referred as the "state of the art" (Schafer & Graham, 2002) or the "principled methods" (Dong & Peng, 2013; Carpenter, Bartlett & Kenward, n.d.) in the missing data literature

and have appealing characteristics: both approaches can deal with almost all the different missing data patterns (Enders, 2010), they both assume the normality distribution of the data, and are robust with data MAR.

Both approaches have advantages and disadvantages. ML-based models have the advantage of producing deterministic results from the data analysis. That is, the same outcome will be obtained every time the data analysis is done with the ML. Conversely, MI-based methods will return different results every time it is run (Allison, 2012). The reason why this happens is that randomness of draws is the main characteristic in MI models. Additionally, ML offers the advantage of likelihood ratio test for nested model comparisons.[8] Collins, Schafer, and Kan (2001) and Schafer and Graham (2002) found that ML "produces smaller standard error than MI;" whereas Graham, Olchowski, and Gilreath (2007) concluded that "ML-based methods have greater power than MI" (as cited in Dong & Peng, p.15). Enders (2010) stated that although ML generates biased parameter estimates under MNAR, "the bias tends to be isolated to a subset of the analysis model parameters," unlike other traditional procedures (p.87).

The advantage of MI over ML is the differentiation between the "imputation" phase and the "analysis" phase. In the imputation phase, MI replaces the missing value with imputed data, thereby generating complete data. In the analysis phase, complete-data approaches can be used on the imputed data set. That is, MI provides plausible values for the missing responses, whereas ML does not. However, certain available software such as SPSS and LISREL provide the option of imputing the missing values "with the raw data after the final EM cycle" (p. 113) providing the users with a complete-data version.

Enders (2010) cautions about using this option to generate complete-response data set within ML context. According to von Hippel (2004), this practice leads to biased parameters and attenuates data standard error (as cited in Enders, 2010).

ML methods include the missingness as an additional variable in the parameter estimation procedure. Therefore, no plausible values are created for missing responses. This is exactly the reason why ML has advocates. When working with ML estimation, the concern about the compatibility between the imputation phase and the data analysis phase is not an issue, because these phases are indistinguishable. When using MI, however, the difference between these two phases (i.e., variables and the underlying model used) needs to be considered. For example, if X_2 was not used for the imputation of Y missing values, a posterior study of the relationship between these two variables may show a weak association. Likewise, interaction among variables can be weak if the model that underlies the MI procedure did not include this feature (Collins, Schafer, & Kam, 2001; Dong & Peng, 2013; Schafer & Olsen, 1998).

Schafer and Olsen (1998) stated that although both approaches are equally efficient, "ML methods will be slightly more efficient (with large sample size) than MI because they do not rely on simulation" (p. 37). Yuan, Yang-Wallentin, and Bentler (2012) also found that ML-based methods tend to be more accurate and efficient than MI. Nevertheless, the fact that the two approaches tend to yield similar results is known. According to Collin et al. (2001) this depends on the congeniality among (a) the model that underlies the ML process, (b) the model that underlies the MI process and (c) the model used to analyze the imputed datasets. Both ML and MI have become accessible

options on a variety of statistical packages, such as M*plus*, SAS, S-PLUS, and R, among others (Schafer & Graham, 2002; Sinharay, Stern, & Russell, 2001). However, there seems to be a preference for MI over ML. For instance, there are twice as many MI publications as ML and, generally speaking, the MI approach has shown a steady increase in publications across the time.[9]

ML and MI with categorical data. Most of the references in the literature (e.g., Chan, Yi, & Cook, 2009; Chou, Bentler, & Satorra, 1991; Curran, West, & Finch, 1996; Enders, 2001, 2002; Enders & Bandalos, 2009; Finch, West, & MacKinnon, 1997; Savalei & Falk, 2014; Yuan, Bentler, & Zhang, 2005;) about the application of these methods to categorical missing data are related to the effectiveness of ML estimation in the structural equation model (SEM) context. This is probably because the available software programs with ML options are mainly for SEM purposes, such as M*plus* or LISREL (Enders, 2010).

According to Enders (2010), some studies (e.g., Chou, Benter, & Satorra, 1991; Curran, West, & Finch, 1996; Finch, West, & MacKinnon, 1997; Hu, Bentler, & Kano, 1992; Yuan, Bentler, & Zhang, 2005) have demonstrated that ML with nonnormal data can impact standard error and distort the likelihood ratio test, although it has little impact on the parameter estimates. However, he also stated that there are corrective procedures for these two issues that are currently incorporated in statistical programs, but that these procedures were designed for complete data.

Nevertheless, some researchers (Enders, 2001, 2002; Gold & Bentler, 2000; Graham, Hofer, & MacKinnon, 1996; Savalei, 2008; Savalei & Bentler, 2005, 2007;

Yuan, 2007; Yuan & Bentler, 2000; as cited in Enders, 2010) have explored nonnormal missing data and ML imputation. For example, Yuan and Bentler (2000) (as well as Yuan & Lu, 2008 and Yuan, 2009) investigated the performance of ML in parameter estimates (e.g., mean, covariance, factor loadings) under normal and nonnormal data distributions and two missingness mechanisms: MCAR and MAR.

The conclusions from these studies are: (a) if the population distribution is known, ML should be modeled taking into account the true distribution of the data. This, however, is not easy to do (unless the data are normally distributed) given that ML models available in software programs are generally based on normal distribution of the data. This adds computational burden to the actually complicated work of modeling with missingness (Yuan & Lu, 2008); (b) ML techniques (e.g., EM) produce accurate parameter estimates under MCAR or MAR when the data are normally distributed. This is actually the scenario for which ML methods were originally designed (Yuan & Bentler, 2000; Yuan, 2009); (c) the discrepancies (bias) they found in their analysis (i.e., between the parameters estimated for the complete normally distributed data and the data with MCAR or MAR) are not due to the use of ML techniques, but mainly due to the size of the sample (Yuan & Bentler, 2000). They, however, said that the variability in the parameter recovery across iterations reflected "that the estimates under MAR may not be as accurate as under MCAR" (Yuan & Bentler, 2000, p. 189), even for normally distributed data; (d) these normal-distribution-based ML methods can still produce efficient parameter estimates when the data are nonnormal and the missingness mechanism is either MCAR or MAR. In the case of MAR, this holds only if the observed

variables are linear combinations of independent random components (Yuan & Bentler, 2000; Yuan, 2009). ML, however, is not equally efficient for all nonnormal distributions, despite the missingness mechanism (Yuan & Bentler, 2000); (e) Yuan and Bentler studied the performance of three ML estimators (the minimum chi-square, the two-stage ML and the direct ML) with nonnormal missing data. They found that the minimum chi-square method (Ferguson, 1996) works better with large samples. The direct ML and two-stage ML methods should be used with medium sample size, although direct ML produces less consistent standard errors with non-normality (Yuan & Lu, 2008). Two-stage ML is recommended for SEM with missing data and unknown population distribution (Yuan & Lu, 2008); and (f) estimates with either contaminated data (i.e., with outliers) or under MNAR are highly biased and inaccurate, regardless of the sample size and the data distribution (Yuan & Lu, 2008).

Yuan, Yang-Wallentin, and Bentler (2012) compared MI and ML approaches for different levels of data missing at random (5%, 6%, 15%, 18%, 25%, and 30%), sample size, number of variables with missing responses, and underlying population distributions (normal, log-normal, and uniform). The authors found that in all the cases ML produced better and more efficient parameter estimates (i.e., variance-covariance matrix) than MI. This suggests that ML is more robust to departure from a normal distribution than MI. It is especially true when sample size is small. However, as sample size increases the estimation bias observed in MI decreases.

When comparing the performance between sandwich-type covariance matrix and the observed-information covariance matrix for estimating the sample standard deviation,

the sandwich-type-based covariance matrix is more precise for ML-based estimates. For MI, none of these two formulas are consistent. Furthermore, the mean parameter was equally and efficiently estimated by both methods. Yuan et al. (2012) also found that the biases are related to the missingness level for both approaches, regardless of the data distribution. Finally, they cautioned that the MI may outperform ML when underlying distribution is known or suspected, given that MI allows working with informative priors.

Bernaards and Sijtsma (1999, 2000) evaluated several imputation methods including EM-based approaches with categorical data. The authors' first study evaluated RDS, OM, CM, IMS, PMS, LD, and EM. The second study compared OM, PMS, CM, IMS, TW, CIM, and additional variants of these methods that incorporate residual variance (denoted as OM-E, PMS-E, CM-E, IMS-E, TW-E, and CIM-E).[10] Also, two EM-based algorithms were included in Bernaards and Sijtsma's second work: the EM-loading and the EM-covariances. The difference between these two methods is whether the first cycle of estimations is conducted using the whole dataset that contains missing values or only the portion with complete data. The EM-loading starts the iteration towards the convergence by replacing missing data with random values and the factor scores are estimated. Then the missing values are adjusted given the estimated factor scores from the previous round until convergence is reached. The EM-covariances method starts the parameters estimation using the data with no missing values only (Bernaards & Sijtsma, 2000).

The fixed factors in the first study were the number of latent traits (two), item categories (m=5), test length (J=20 items), and the correlation between the two latent

variables (r=0.24). The variable factors in the first study were sample size (n=100 and 500), missingness level (5%, 10%, and 20%), the missingness mechanisms (MCAR and MAR), and item mixture ratio (1:0, 3:1, and 1:1). The item mixture ratio represents the proportion that the item measures each of the two latent traits. This ratio has an effect on the scoring weight value, where 1:1 ratio has the same weight for all the items and thus is considered unidimensional (Bernaards & Sijtsma, 1999, 2000). Also, Bernaards and Sijtsma (1999) had the factor extraction methods (i.e., principal components and ML) as variable condition. The fixed factors in the second study were the same, except for the correlation between the two latent traits, which was set to vary (r=0, 0.24, and 0.5). Likewise, the variable factors were the same in both studies. The authors measured the performance of the imputation methods by comparing the factor loadings recovery in both studies (Bernaards & Sijtsma, 1999, 2000).[11]

Additionally, two special designs were studied in their first paper, both with data MAR (Bernaards & Sijtsma, 1999). All the imputation method except LD were compared. In the first design, the data had four dimensions or latent traits but only two factors were extracted (i.e., the variables were loaded on two factors only). The same test length and missingness level as the main study were kept. The sample sizes (n=50, 100, and 150) and item mixed ratios (3:1:1:0, 3:1:0:0, 1:3:0:3, and 1:3:0:0) were different. In the second special design, the data were bidimensional and the conditions were 5% and 20% of missing data, n=100, and the item mixed ratios were the same as in the main study (1:0, 3:1, and 1:1). In this case, the number of factors underlying the data were decided upon the eigenvalue > 1 criterion obtained from the data with imputed values. If

the eigenvalue was equal to or larger than one, the factor was retained. The number of latent traits detected with the eigenvalue criterion was the measure of good (if 4 factors are retained) or bad performance of the missing data approaches.

The authors found that the missingness level affected the performance of the missing data handling methods (Bernaards & Sijtsma, 1999, 2000). For example, doubling the missingness level (e.g., from 5% to 10%) at least doubles the bias in the factor loading recovery (Bernaards & Sijtsma, 1999). They also saw that the bias decreased as *n* increased (Bernaards & Sijtsma, 1999, 2000). The results showed that the bias was higher when the missingness progressed from MCAR to MNAR, regardless of the other analyzed conditions. The relative performance of the imputation methods, however, seemed to be independent of the missingness mechanism (Bernaards & Sijtsma, 1999, 2000). Furthermore, all the imputation approaches, except the EM-based models, improved their performance when the correlation among the latent traits increased. The best situation was when the data were unidimensional. The EM-loading and EM-covariances approaches, however, were independent of the level of association among latent traits (Bernaards & Sijtsma, 2000).

Among all the methods, EM was consistently the best method to handle missing data (Bernaards & Sijtsma, 1999, 2000). Both EM-loading and EM-covariance produced good factor loading values recovery when working with rating scale data. PMS was the second best method, performing even better with unidimensional data (i.e., high latent traits correlation value) (Bernaards & Sijtsma, 1999). In general, person mean techniques (PMS, TW, TW-E, CIM, and CIM-E) are good alternatives for factor loadings recovery if

the researcher prefers working with simpler approaches. Although CIM and TW tend to inflate the correlation between the latent traits. CIM-E and TW-E are hence better. These person mean techniques performed better than LD and imputation methods not based on the person mean (IMS, RM, CM, and OM) (Bernaards & Sijtsma, 2000). The worst methods were RDS, LD, IMS-E, CM-E, and OM-E. In terms of factor extraction, they found that both ML and principal components factor analysis equally estimated factor loading (i.e., they had similar level of bias) (Bernaards & Sijtsma, 1999, 2000).

Finally, in the first special design Bernaards and Sijtsma (1999) found that the effect of missingness level and sample size was the same as in the main study. EM did not perform well in this design. They found that when factor analysis extracts the wrong number of latent traits it distorts the performance of all the imputation methods. Thus, not knowing the number of dimensions underlying the data can have consequences on the quality of performance of missing data approaches. The second special design found that the eigenvalues were affected by the item mixed ratios, regardless of the missingness level and missing data approach, except EM. When the items were 1:0 four eigenvalues were larger than 1. For the ratio 3:1 two eigenvalues were larger than 1 and for 1:1 (unidimensional), it reduced to one. The eigenvalues were not significantly affected by missingness level and were similar for RDS, OM, CM, IMS, and PMS. Finally, datasets imputed with EM algorithm yielded eigenvalues that led to the identification of correct number of dimensions in all the item mixed ratios and missingness levels.

Among the MI algorithms, the most widely used is MIDA (Schafer, 1997). This approach has been extensively studied, especially in the educational and psychological

context. The other MI approach, Fully Conditional Specification or FCS (Raghunathan, Lepkowski, van Hoewyk, & Solenberger, 2001; van Buuren, Boshuizen, & Knook, 1999), is well known and more used in the medical field, but there are applications of this approach in IRT models.

Fully conditional specification. When comparing FCS and MIDA, there are some interesting findings. For instance, Lee and Carlin (2010) found no difference between the FCS and MIDA. They compared the performance of these two methods with data with J=5 mixed variables (one continuous and 4 categorical), n=1000, and data MAR. Up to 33% of missingness was present in one, three, or four variables and none of the variables was normally distributed. The continuous variable was skewed and the categorical variables had 2 or 5 categories. When imputing the skewed continuous variable the two imputation methods were compared under three conditions: when skewness was ignored, when the variable was log transformed, and when the variable was log transformed so that the skewness was zero. Variables with 5 categories were imputed with FCS ordinal logistic regression and using MIDA with rounding to the nearest value. The binary variable was imputed with FCS logistic regression and MIDA with either simple or adaptive rounding. The simple rounding was first suggested by Schafer (1997) and other authors after that (Schafer & Olsen, 1998; Allison, 2001). In simple rounding, the imputed values will be rounded to either zero or one. Imputed values equal to or higher than 0.5 will be rounded to 1, if less than 0.5 the imputed value is replaced with zero. In adaptive rounding, a normal approximation to the binomial distribution is used to decide the imputed value.

Lee and Carlin (2010) observed the parameter recovery (here, regression coefficients) after imputation with both methods. They found that the recovery was equally poor with both imputation methods when the missingness was only in the continuous variable and its skewness was not considered. Once skewness was incorporated (i.e., variable was transformed so skewness=0), both imputation approaches performed equally well. Likewise, FCS and MIDA were equally accurate when missingness was also present in the categorical variables with 5 categories. Finally, the two approaches yielded the same results when the binary variable had missing values, with the adaptive rounding imputation showing the best parameter recovery.

van Buuren (2007) also found that both FCS and MIDA with simple rounding recovered the parameters (here, regression and correlation coefficients) with almost the same accuracy. In his study (n= 3801, J=3 items, m=2 or 5 categories, and 58% of the sample with at least one missing response), he warned about the accuracy of the reference curves values estimation. Reference curves are standard curves computed from the responses of reference participants (e.g., the ratio weight/height for children to detect undernourishment). In this study the reference curve refers to breast development by age. When using MIDA with rounding procedure, the parameters were well recovered, but the reference curved showed underestimation of breast development at early ages and overestimation otherwise. On the other hand, FCS produced good reference curves for the different breast development stages. That is, MIDA did not preserve the original ratio between the dependent and independent variables along the continuum as good as FCS did. Consequently, van Buuren (2007) concluded that FCS is better than MIDA when

dealing with categorical variables, regardless of the number of categories, under ignorable missingness. The author, however, cautioned that MIDA performance could be affected by the fact that the imputed data were rounded to the nearest plausible value. This technique has proven to not be effective when working with categorical data, especially with binary variables (Ake, 2005; Allison, 2006; Horton, Lipsitz, & Parzen, 2003).

In another study, van Buuren (2010) compared the performance of FCS approach with two versions of TW (Bernaards & Sijtsma, 2000). Two datasets were simulated, both of the same length (10 items) and size (n=11000). The first one consisted binary items and the second dataset was comprised of items with five response categories. In both datasets, half of the items loaded on one dimension and the other half did on the second dimension, the two dimensions were correlated (r=.10). The author examined three different MCAR levels (44%, 58%, and 73%). One version of TW (TW1) imputed the missing values assuming that all the items loaded on the same construct, whereas the second version (TW2) correctly assumed the data were bi-dimensional. The performance of the imputation methods was compared using three indices: (a) the number of valid cases used in the data analysis (after the imputation). That is, the number of cases that were not removed by the software after being flagged as extreme values, (b) the Cronbach's alpha, and (c) the correlation between the two dimensions or scales measured with the instrument.

van Buuren (2010) found that the number of valid cases to be used in the data analysis was higher when FCS was used, regardless of the number of item categories.

That is, the TW approaches imputed values that were categorized as extreme during the model fit analysis (done with Rasch). Likewise, the Cronbach's alpha was best recovered, although slightly underestimated, by the FCS technique for both dichotomous and polytomous items. TW1 and TW2 greatly overestimated this coefficient regardless of the number of item categories. The same pattern was observed with the correlation coefficient between the two scales. The correlation between scales was inflated with TW1 and TW2 and it was around the real value (.10) with FCS.

Multiple imputation with data augmentation. Several authors have stated that MIDA is effective for handle missing categorical data (e.g., Ake, 2005; Allison, 2006; Bernaards, Belin, & Schafer, 2007; Horton, Lipsitz, & Parzen, 2003; Leite & Beretvas, 2010; Schafer, 1997; Schafer & Graham, 2002), even under clear violation of MIDA's multivariate normal distribution assumption (Allison, 2006; Schafer & Olsen, 1998). This approach also has been shown to produce acceptable results with ordered categorical data, especially with high number of item categories (Leite & Beretvas, 2010).

For instance, Leite and Beretvas (2010) studied the performance of MIDA with rating scale such as Likert-type items. They evaluated the correlation coefficient recovery under different missingness levels (10%, 30%, and 50%), mechanisms (MAR and MCAR), number of item categories (m=3, 5, and 7), inter-item correlation (r = .2 and r = .8), and data distribution (normal and non-normal) with n=400. Their results showed that: (a) the correlation coefficients using data with imputed values using MIDA were consistently underestimated; (b) MIDA was robust to violations of normality and continuity; and (c) MIDA's effectiveness was not affected by the inter-item correlation

level but by the missingness level and mechanisms. For example, MIDA was robust to MAR or MCAR with 10% of missing data, but when the missingness was 30%, MIDA only performed well when data were MCAR. MIDA produced unacceptable bias when 50% of data were missing. They concluded by saying that MIDA can be safely used with a low missingness level (i.e., less than or equal to 10%).

Regarding binary variables, there are contradictory recommendations concerning rounding the estimated missing value. Schafer (1997), Schafer and Olsen (1998), and Allison (2001) suggested rounding the imputed values to 1 when the estimated value is 0.5 or higher, and to zero otherwise. Horton, Lipsitz, and Parzen (2003), on the other hand, showed that the proportion of correct responses or the probability of success (p) with binary variables is better estimated when the imputed value is not rounded under MCAR condition. Ake (2005) also found that dichotomous variables with unrounded imputed values resulted in less bias (difference of the estimated p with respect to the real p), with up to 40% of missing data when MCAR or MAR was observed. He showed similar results for non-binary categorical data. Allison (2006) expanded Horton et al.'s (2003) findings by demonstrating that rounded imputed values led to the worst recovery method for both p and linear regression coefficients, despite the missingness mechanisms (MCAR or MAR).

Bernaards, Belin, and Schafer (2007) found that rounding methods also play a relevant role in parameter estimates bias. Using two different missingness levels (25% and 50%) and sample sizes (n=50, and 500), Bernaards et al. tested three different approaches (simple rounding, coin flipping rounding, adaptive rounding) of rounding

values imputed with MIDA. The data were MAR and contained both categorical (binary) and continuous variables. The rounding methods were only applied to binary variables. Values imputed for continuous variables were kept as generated by MIDA.

In simple rounding, the imputed values will be rounded to zero when they are if less than 0.5, and to 1 otherwise (Schaffer, 1997). Coin flipping rounding is based on a Bernoulli distribution where the imputed values between 0 and 1 were treated as the probability of drawing 1. In adaptive rounding, a normal approximation to the binomial distribution is used. Here, the threshold (*t*) values for the rounding decision were estimated with:

$$t = \bar{\omega} - \Phi^{-1}(\bar{\omega})\sqrt{\bar{\omega}(1-\bar{\omega})}, \quad (1)$$

where $\bar{\omega}$ "denotes the mean value on a single variable [i.e., an item] of available <observed> binary observations and imputed values produced by the multivariate normal imputation procedure" (p. 1372) and can range from 0 to 1; Φ^{-1} is the quantile function of a normal distribution, with $\Phi(Z)$ for $Z = (\bar{\omega} - p)/\sqrt{\bar{\omega}(1-\bar{\omega})}$, which has a normal distribution for a given population proportion (p) (Bernaards et al., 2007).

Several parameters were evaluated (Bernoulli proportion, odds ratios, continuous parameters, and logistic regression coefficients). For all of these, Bernaards et al. (2007) found that the parameters were recovered with little bias when the variables were continuous, while the parameter estimates showed higher bias in binary variables. The adaptive rounding generated parameter estimates only slightly better than the simple rounding regardless of sample size and missingness level. The worse performance was seen with the coin flipping method.

Other multiple imputation methods. There are other MI approaches for categorical data such as RandomForest (RF), MissForest (MF), log-linear multiple imputation (LLMI), and latent-class multiple imputation (LCMI). RF is an algorithm developed by Breiman and Cutler in 2001. It is currently available as stand-alone software (https://www.salford-systems.com/) and as R package. The goal in the RF is the imputation of continuous and categorical variables with classification trees. The classification tree analysis is one of the main techniques used in data-mining. It consists of defining the different outcomes (plausible values) that can be potentially obtained from the combination of different variables (decision tree). In the RF, several classification trees are constructed and the one with the higher chances is selected. Detailed information of how this method works can be found in Breiman's (2001) work and at his website (http://www.stat.berkeley.edu/~breiman/RandomForests/).

MF It is a non-parametric iterative approach that deals with mixed-type data (i.e., categorical and continuous). Stekhoven and Bühlmann (2012) proposed this approach, and it is based on RF algorithm. In the first stage, a RF estimation is computed only on the complete data. Then the missing values are predicted and they are carried again to the first stage as input for the next cycle. The process continues until convergence is reached. MF's main advantages are that it does not need tuning of parameters nor it requires previous data distribution assumption. Also, it can be used in data with complex interaction, non-linear relation or high dimensional datasets Stekhoven and Bühlmann (2012).

Stekhoven and Bühlmann (2012) tested this technique with mixed-type format. They compared it with different techniques for both continuous, categorical, and mixed data. For continuous data, MF was compared to the *k*-nearest neighbor imputation (KNNimpute), and the missingness pattern alternating lasso algorithm (MissPALasso). Whereas for categorical data, MF was compared to the MICE algorithm, and a dummy variable encoded KNNimpute. Different missingness levels (10%, 20% and 30%) were studied under MCAR condition. They found that MF performed better than KNNimpute with continuous variables. MF was better than MICE and the dummy variable encoded KNNimpute with categorical variables. With the mixed-type data, MF again did well. Unfortunately, no evaluation of this has been done with less strict missingness mechanisms.

Andreis and Ferrari (2012) examined the performance of four missing data handling methods: LD, MICE, forward imputation (FI), and MissForest. LD and MICE were previously described. FI is "based on an iterative algorithm which alternates nonlinear principal component analysis (NLPCA) on a subset of the data with no missing data and sequential imputations of missing values by the nearest neighbor method" (Ferrari, Annoni, Barbiero, & Manzi, 2011, p. 2412). FI is effective in factor loading estimation and score recovery in multidimensional categorical analysis. Andreis and Ferrari compared the performance of the missingness methods in the estimation of item parameters using multidimensional IRT, specifically the M2PL, with the imputed data. The dataset consisted of N=113 examinees and 10 dichotomous items with different missingness levels (5%, 10%, and 30%) and mechanisms (MAR, MCAR, and MNAR) in

four of them. The first problem the author faced was that FI and MICE do not fill in all the nonresponses. Therefore, they had to stochastically impute values that the methods did not fill in.

Andreis and Ferrari found that missingness level and mechanism had an effect in the missing data handling techniques. Among the methods, there was not a technique that yielded good estimates. The item difficulty was recovered better by FI and MF, regardless of the missingness mechanisms or level, whereas MF and MICE recovered the item discrimination parameter best. Nonetheless, they found that in the majority of the cases, the δ_j was overestimated, except when 30% of the data were MNAR. Conversely, the α_j was underestimated. The LD outperformed the other methods when recovering α_j and the missingness was high.

LLMI was proposed by Schafer (1997) and has been shown to perform well. Also, the author offers a free stand-alone software called CAT that imputes values with the LLMI approach. Research has shown that LLMI "yields unbiased statistical inference, and it is robust against departures from the assumed imputation model" (Vermunt, van Ginkel, van der Ark, & Sijtsma, 2008, p. 371). However, LLMI's major drawback is that it works only when one has a small number of categorical variables. As such, it is impractical when using empirical data (Finch, 2008; Gebregziabher & DeSantis, 2010; Schafer, 1997; Vermunt, van Ginkel, van der Ark, & Sijtsma, 2008).

LCMI was suggested by Vermunt, van Ginkel, van der Ark, and Sijtsma (2008). LCMI is an unrestricted latent model that incorporates the missingness through a binary variable. According to the authors, the advantages of this approach are: (a) the imputation

with LCMI can be done separately for each variable with missing values. Thus, the size of the datasets is not an issue as it is with LLMI; (b) LLMI "respects the categorical nature of the variables" (p. 390); (c) its flexibility, because it is able to detect and conserve complex dependencies between the variables present in the imputation model; and (d) it "is easy to apply and [it is] neutral in the sense that no detailed a priori content knowledge is needed to build an imputation model" (p. 390). The LCMI model assumes that the joint probability density of the person's observed responses on *J* categorical variables is:

$$P(y_{i,obs};\theta) = \sum_{k=1}^{K} P(x_i = k;\theta_x) \prod_{j=1}^{J} \left[P(y_{ij}|x_i = k;\theta_{y_j})\right]^{r_{ij}}, \quad (2)$$

where $P(y_{i,obs};\theta)$ is the joint probability density of y_i, the vector of observed responses of person *i* on *J* categorical variables; y_{ij} is the answer of the person *i* on the item *j*; x_i is a particular latent class; *K* is the total number of latent classes with index *k*; $\underline{\boldsymbol{\theta}} = (\theta_x, \theta_y)$ is the vector with unknown parameters, the subscripts indicate to which set of multinomial probabilities the unknown parameters belong; and r_{ij} is the missingness indicator for the person *i* on item *j*. If the person did not answer the item, then r_{ij}=0, 1 otherwise. Likewise, the conditional distribution of the missing responses is:

$$P(y_{i,miss}|y_{i,obs};\theta) = \sum_{k=1}^{K} P(x_i = k|y_{i,obs};\theta) \prod_{j=1}^{J} \left[P(y_{ij}|x_i = k;\theta_{y_j})\right]^{1-r_{ij}}, \quad (3)$$

The extent to which the LCMI imputation model approximates the distribution of y_i depends on *K*, with a larger *K* providing a better approximation than a smaller *K*. Three model-fit statistics are used to determine the appropriate *K*: BIC, AIC, and AIC3.[12] The *K* with the lowest model-fit statistic value should be selected.

Vermunt et al. (2008) used a nonparametric bootstrap for the imputation of values under this model. In their study they compared LCMI, ML estimation with missing data, and LLMI using simulated data with 70% of values MAR. This dataset was comprised of six dichotomous variables and n=10000. Additionally, they examined parameter recovery using LCMI with empirical data (n=4292) which had 81.5% of missingness and 79 categorical variables with different number of categories (between 2 and 17). In both analyses, the parameters to evaluate were regression coefficients. Vermunt et al. found that imputation with larger latent class number yielded better results than a smaller number. They also said that the actual value of K is not relevant, as long as it large enough. They suggested to use AIC and AIC3 over BIC to have an idea of the K value. They also found that LCMI recovered parameter estimates with the same accuracy as LLMI and ML with missing data. Moreover, LCMI worked well under MAR regardless of the number item categories.

MI and ML with auxiliary variables. Additional research has found that both ML and MI perform better in handling missing data with the support of auxiliary variables or covariates than when they do not use covariates (Meng, 1994; Rubin, 1996; Schafer, 1997; Schafer & Olsen, 1998; Collins et al., 2001; Schafer & Graham, 2002; Graham, 2003; and Enders, 2010). Auxiliary variables improve the missing data handling procedure. With simulated data, Collins et al. (2001) showed that values imputed with auxiliary variable were less biased than when they were not used. As a consequence, imputation with auxiliary variables reduces the standard error and thus increases statistical power. Auxiliary variables are especially useful with MI, which can handle a

higher number of auxiliary variables than ML. Also, MI with auxiliary variable is available in a larger number of statistical programs than ML (Collins et al., 2001).

There are three model strategies to incorporate auxiliary variables into ML-based analyses: the extra dependent variable, the saturate correlates model, and the two-stage approach (Enders, 2010). The first two are correlation-based models proposed by Graham (2003), but the saturate correlates model is easier and more efficient than the extra dependent variable model. The third strategy estimates a mean vector and covariance matrix that incorporates as many auxiliary variables as desired (stage 1) and uses this information as input for the following analysis (stage 2). Its major disadvantages are (a) the need to specify the sample size a priori and this can "bias the standard errors from the analysis stage (Enders & Peugh, 2004)" (Enders, 2010, p. 134), and (b) the lack of friendly programs to implement some features that affect parameter estimates precision (Enders, 2010).

Collins et al. (2001) mentioned that auxiliary variables are useful because they may be related to the cause of missingness or at least correlated with the variables that have missing values. This increases the likelihood of adjusting the missingness mechanisms from MNAR to MAR (Sinharay et al., 2001) or to mitigate the bias under MNAR condition (Collins et al., 2001; Enders, 2010). Schafer and Olsen (1998) also suggested including variables in the MI process when they seem to be strong predictors of missingness, even if they are not needed for later substantive analyses. In this regard, Collins et al. suggested using auxiliary variables when the correlation between them and the variable with missing data are 0.9 and the missingness exceeds 25%. She found that

when missingness is less than 25% and the correlation is .4 omitting auxiliary variables has negligible effects in the results. Enders (2010) recommended using auxiliary variables when the correlation between them and the variable with missing data are at least 0.4, despite the missingness mechanisms (MCAR, MAR, or MNAR).

On the other hand, there is no agreement on the consequences of including too many auxiliary variables. Sinharay et al. (2001) found a negative effect on the accuracy of the parameter estimates (i.e., correlation coefficients) when the number of covariates increases in the imputation model. Likewise, multicollinearity problems and variance inflation (VIF) due to large numbers of covariates may be also present (Wayman & Swaim, 2002; as cited in Leite & Beretvas, 2010; Yuan & Lu, 2008). Conversely, Collins et al. (2001) advised that researchers should care more about the implications of omitting auxiliary variables than including irrelevant ones. That is, including too many of them does not harm the results, although there is little benefit in using a large number auxiliary variables (Enders, 2010).

Supporting this argument, Schafer and Olsen (1998) stressed the importance of including variables in the MI process on which later investigation will be carried out in order to preserve the association between them. Also, the model used for the imputation should not be too different from the analysis model in order to avoid altering (strengthening or weakening) potential relationship among variables (Meng, 1994; Schafer & Olsen, 1998). This includes interactions among variables or any other complex relationship they may have (Enders, 2010). However, Sijtsma and van der Ark (2003)

said that using the same model, say IRT, for both imputation and analysis produces a dataset biased in favor of the hypothesis that is modeled.

One explanation for the lack of agreement in the number of auxiliary variables and its consequences can be found in Thoemmes and Rose's (2014) work. They showed that the inclusion of auxiliary variables not necessarily mitigates bias, but it can also enhance it. They found that "true" the relationship between the auxiliary variable, the missingness variable (which is a binary variable that shows whether the value is observed or missing), and the outcome variable (which has missing values and is the one for which the imputation is aimed) plays a role in the quality of the imputation using either ML or MI techniques. They classified auxiliary variables into bias-induced and bias-reduced variables based on that relationship. They used direct acyclic graphs and simulated data (*n*=500, 30% missingness, multivariate normally distributed variables, and six different values of explained variance associated with the auxiliary variable, 0%, 5%, 15%, 20, 25%, 35%, and 45%, under MCAR, MAR, and MNAR conditions) to prove their hypotheses. They evaluated the outcome variable mean and variance recovery in this study.

They found that when data are MCAR and MAR: (a) auxiliary variable induces bias in the estimates when the auxiliary variable is "truly" not related to the outcome variable or the missingness, even though when the correlation between the auxiliary variable and the other two variables seems to be important. The bias worsens when the missingness mechanism detaches from MCAR; (b) if the auxiliary variable is "truly" related to the other variables (missingness and outcome variables); and when data are

MNAR (c) “bias can be increased in the presence of MNAR, even if an auxiliary variable is added that is directly related to missingness and outcomes” (p. 28); (d) the sign of the coefficients between the auxiliary variable and the measured variable determines whether the inclusion of auxiliary variables increases or reduces bias. Thoemmes and Rose recognized that identifying whether the variable is bias-reduced or bias-induced in the imputation context is hard. They, however, suggest to avoid the inclusion of “as many as possible” criterion when selecting the auxiliary variables. Rather, they suggested to do a careful consideration of the variables before including them in the imputation process.

Additional concerns about the use of auxiliary variables are the quality of the collected information (normally done with a background questionnaire) and the missingness that may also be present in these variables. Generally, the auxiliary variables are categorical data that classify the participants into categories, such as socio-economic status, gender, religion, etc. When the missing values of these auxiliary variables are imputed, the risk of misallocating examinees into the categories is high. This increases the chances of biased parameter estimates for the groups. Rutkowsky and Rutkowsky (2010) address these issues for international large-scale assessments from the perspective of plausible values. Plausible values are also a MI procedure in which several achievement scores are assigned to each participant student in order to estimate the performance of the population they belong to. This approach relies on the students’ performance on the test and their background information (or auxiliary variables) such as gender and socio-economic status.

The authors affirmed that background information is not as accurate as it could be, because inconsistencies in participants' answers are normally present. This could be due to the fact they do not know the right response (e.g., how much students know about their parents' level of education), or simply because the question is not clear to them. The authors demonstrated this inaccuracy issue with two indicators: (a) the (low) correlation between children and parents' responses to items that are common to the two questionnaires, and the (b) (low) scale reliability for countries ordered by level of income. The worse results for both indices were present in the lowest-achieving participants and middle- to low-income countries.

Rutkowsky and Rutkowsky (2010) highlighted that missingness in the background data are a problem for the country performance estimation in large-scale assessments. Enders (2008), however, said that the decision of working with incomplete auxiliary variables depends on both (a) their level of missingness and (b) the extent to which the auxiliary variables are correlated with the missingness in the variable for which the imputation is done (i.e., the manifest variable). In his full information maximum likelihood-based structural equation study, he saw that when auxiliary and manifest variables are highly related, the auxiliary variable "works well" even when it has 50% of missingness, regardless of the auxiliary variable missingness mechanism (MAR or MNAR).

Enders also found a positive relationship between biased parameters and the proportion of missingness that is simultaneously present in both the manifest and the auxiliary variables. The bias is higher when the missingness is simultaneously present in

both variables, but missing values in either one or the other did not produce bias. Also, he found that the bias was extreme when 15% of the cases had matching missingness. However, bias was minimal when only 8% of the cases had this pattern. In his book, Enders (2010) suggests not including auxiliary variables that share more than 10% of missing cases with the variables for which the imputation will be done.

Missing data in IRT context. Some missing data studies explicitly incorporate the missingness treatment in IRT models. They either generated the imputed values with IRT models (e.g., Huisman & Molenaar, 2001; Sijtsma & van der Ark, 2003) or suggested adjusted IRT models that incorporate the missingness as latent variable (e.g., Glas & Pimentel, 2008; Holman & Glas, 2005; Pimentel, 2005), as a manifest variable using an indicator (Rose et al., 2010), or as grouping factor using missingness levels (Abad, Olea, & Ponsoda, 2009) in the data analysis.

Imputation with IRT models. The response-function (RF) imputation and mean response-function (MRF) imputation are nonparametric estimations based on the item response function (IRF) of a subsample of the data with no missing values. These approaches "do not impose restriction on the shape of the IRF and not explicitly on the dimensionality of measurement" (p. 514, 515) to avoid bias towards a particular IRT model. Imputation with RF and MRF include random draws from the Bernoulli distribution for binary data and from the multinomial distribution when the items are polytomous. The difference between these two approaches is that the first one uses proportion correct as part of the process, whereas MRF uses the mean of the regressions for all the items on the test (Sijtsma & van der Ark, 2003).

Sijtsma and van der Ark (2003)[13] evaluated these two methods along with other two approaches that are also based on random draws from the Bernoulli distribution: PMS and TW imputation. Factors studied were missingness level (1%, 5%, and 10%) and mechanisms (MCAR and MNAR), sample size (n = 10, 20, 50, 100, 200, 500, 1000, and 2000) and test length (J = 10 and 20). The recovery of four variables was evaluated: Cronbach's alpha, the Mokken's H scalability coefficient, R_{1c} and $Q_{2.}$[14] The first two coefficients are scale quality measurements, while the last two are Rasch coefficients of goodness of fit. They used R_{1c} and Q_2 to compare whether the data were unidimensional.

Using dichotomous data, Sijtsma and van der Ark found that RF best recovered all four coefficients previously mentioned, regardless of the missingness level or mechanism. However, RF was more accurate in recovering R_{1c} and Q_2 than estimating Cronbach's alpha and the Mokken's H scalability coefficient. The performance of the other three approaches (MRF, PMS, and TW) was conditional on the missingness level. The second best performer was the TW imputation although it often overestimated the scale quality measurements. Like RF, TW was found to work well with nonignorable missingness. The authors also pointed out that RF can be unstable when the subsample of data with complete information used for RF is small. In this case, TW imputation should be preferred. Although they did not clarify whether the data were unidimensional, they found that PMS and TW imputation tend to reject the unidimensionality assumption, contrary to MRF. Finally, the authors mentioned that RF and MRF may work best with unidimensional data, although they did not explore this idea.

van Ginkel, van der Ark, and Sijtsma (2007) evaluated the performance of six different missing data approaches (RDS, TW, TW-E, CIM-E, RF, MIDA) by comparing the discrepancy in the estimation of three statistics (Cronbach's alpha, Loevinger's scalability *H*-coefficient, and the item cluster solution from Mokken's scale analysis) using ANOVA in two studies. The discrepancy was defined as the difference between the coefficient obtained from the imputed dataset and the coefficient estimated from the complete dataset. For the first study, the variable factors were the latent-variable ratio, represented by different item ratios (1:1 and 3:1), sample size (n=200 and 1000), missingness level (5% and 15%), and missingness mechanisms (MAR, MCAR, or MNAR), and whether auxiliary variables were used in the imputation. Fixed factors were bidimensional data, J=20, polytomous (m=5) items and correlation between the latent variables of 0.24. For the second study, n=1000, J=20, missingness was 5% and MAR, auxiliary variables were used in all the imputations, and the data were bidimensional. The factors that varied in this second analysis were latent traits correlation (r=0.0, 0.24, and 0.5), latent-variable ratio was of (1:0, 1:1, and 3:1), and the number of m (2 and 5).

van Ginkel et al. (2007) observed that missingness mechanism had a little effect in the bias observed in the recovered parameters. More relevant to the parameter recovery were the levels of missingness and sample size. That is, they had an important role in the performance of imputation approaches. Moreover, the effectiveness of the imputation approaches varied depending on the variable used for the comparison (Cronbach's alpha, Loevinger's coefficient, or the item cluster solution). For example, TW-E and CIM-E were least affected by changes in the missingness level when recovering Cronbach's

alpha and Loevinger's coefficient. For the item cluster solution recovery, TW-E, MIDA, and RF were more stable across missingness level and sample size changes.

van Ginkel et al. found that Cronbach's alpha recovery was affected, although slightly, by the level of correlation of the latent variables and the dimensionality. They saw that Cronbach's alpha was better recovered TW-E, CIMS-E, and RF when the correlation between latent variables was high or the data were unidimensional. The item cluster solution coefficient were better recovered with MIDA under the same circumstances. Loevinger's coefficient was best recovered by TW-E, CIMS-E and RF, but this statistic was neither affected by the dimensionality of the data nor the correlation between the latent traits. Finally, the number of item categories did not affect the recovery of the Cronbach's coefficient. Discrepancy in the Loevinger's and the cluster solution coefficients, however, was higher when m=5 than when m=2. The imputation approaches that performed the best were TW-E, CIMS-E and RF for Loevinger's coefficient and RF and MIDA for the cluster solution coefficient.

Overall, van Ginkel et al. found that a combination of a small missingness level and a large sample size seemed to improve the performance of missing data approaches. Also, TW-E, CIM-E and the RF were consistently the approaches that yielded the smallest discrepancies in recovering two of the evaluated statistics regardless of the factors under study in this research. There were situations where RF outperformed the other two approaches in the case of the cluster solution coefficient. MIDA did not do better than the three aforesaid methods, but its performance was not the worst. RDS was consistently the worst approach.

Huisman and Molenaar (2001) worked with IRT models in both data imputation and data analysis stages; this was an extension of Huisman (2000). The factors they worked were sample size (N=20, 400, and 800), number of item categories (m=2, 3, and 4), test length (J=5 and 10), missingness level (5%, 12%, and 20%), and missingness mechanisms (MCAR and two levels of MNAR). The levels of analysis were (a) person's sum score, which was computed as the weighted or unweighted sum of correct responses, and (b) scale quality indices: Cronbach's alpha and the Loevinger's H-coefficient.

They evaluated six different missing data treatment approaches. Two of them, hot-deck nearest neighbor (HDNN) and CIM, were based on a technique they called adjustment cells. The idea behind adjustment cells is that respondents are grouped according to certain covariates or auxiliary variables and the imputation is done for each group individually. The third imputation technique was based on the non-parametric Mokken scaling-based model (MOK). In the MOK imputation, the dichotomous items are ordered by difficulty (i.e., proportion of correct response). Then, each person's imputed response for the *jth* item is determined by the number of correct answers the individual has before and after the item with nonresponse (see p. 227 for the imputation criteria).

The remaining three approaches are based on the 1PL IRT model: expected value (EV) rounded to the nearest integer, one single random draw (SRD) from the estimated distribution of responses, and multiple random draws (MRD) from the estimated distribution of responses (Huisman & Molenaar, 2001). Person ability is estimated using the observed responses. Later, the distribution of responses in every cell (i.e., estimated

probability) is estimated. A response then is imputed either by rounding the estimated probability to the nearest integer (EV), or drawing one or multiple responses based on the estimated probability (SRD or MRD, respectively).

At the person quality level, Huisman and Molenaar found that the performance of the imputation techniques was conditional on the test length (positive relationship), the number of item categories (positive relationship), as well as the missingness level (negative relationship) and mechanism (worse with MNAR than with MCAR). Sample size did not affect the person's sum score. Also, test length was more important than the number of item categories. That is, the sum score was better estimated for a long test with a small number of item categories than for a short test with a large number of item categories. Additionally, overestimation of sum score values was present for data MNAR, but not for data MCAR. The authors also found that person's sum score was best estimated when the missing data were imputed with MRD, regardless of the factors. The second best method was EV, although only slightly better than CIM and MOK. HDNN did not perform well at all. Huisman and Molenaar also highlighted that MOK and the three IRT-based imputation models recovered sum score better when data were dichotomous and the missingness level was low and nonignorable.

At the scale quality level, both scales were overestimated with all the approaches but HDNN. Loevinger's *H* was more affected than Cronbach's alpha. This could be due to the fact that Cronbach's alpha depends on the test length and number of categories. These scale indices cannot be used as evidence of the goodness of fit between model and data, especially when IRT-based models are used for both stages imputation and analysis

of imputed data (Huisman & Molenaar, 2001).

In a second section of the study, Huisman and Molenaar compared individual's performance estimate using person ability instead of sum of score using three different missing data treatment (MRD, HDNN, and CIM) and the analysis of incomplete data. The factors were missingness mechanism (MCAR and MNAR), J=5, m=2 and 3, N=400, and 12% of missing data. They found that person ability was best estimated with no imputation of nonresponses in three of the four conditions. CIM performed best only with data MCAR and m=3. Huisman and Molenaar found that MRD was more affected than the other approaches across the conditions. They argued that this is due to "a systematic change in the ability estimates by [running the model] twice, for imputation and estimation consecutively" (p. 241).

Finch (2008) studied seven approaches (4 deterministic and 3 stochastic) for handling missingness under different sample sizes (N=500, 1000), missingness level (5%, 15%, and 30%), and mechanisms (MAR, MNAR) within the IRT context and using dichotomous data. The seven approaches were CIM, not-presented (NP), incorrect (IAS), fractionally correct (FR), RF, MIDA, and EM algorithm. In the NP, the cases are completely removed from the data. The FR is the imputation with the reciprocal of the total number of item categories, 1/m. Item parameter recovery for the 3PL model was assessed. The number of items correctly answered was used as covariate in the MIDA process under the MAR condition.

Overall, Finch found that scoring missing data as incorrect was the worst method to address missingness for dichotomous responses with EM being the second worse and

MIDA only slightly better. Also, when data were MNAR none of the approaches performed well. MIDA, FR and NP had almost the same performance with data MNAR. As with previous research (e.g., Huisman, 2000; Huisman & Molenaar, 2001), sample size was not relevant in the item parameter recovery. Moreover, the standard errors of item parameter estimates were smaller under MNAR condition than under MAR.

There were some particularities with the item parameters' recovery. In the case of the item discrimination, the level of missingness affects the estimation bias level. MIDA yielded the best α_j estimation. Item difficulty was also best recovered with MIDA. Under MNAR, however, the δ_j was underestimated regardless of missingness level and missing data approaches, indicating that the items appeared to be easier than they really were. Treating nonresponse as incorrect led to the opposite conclusion; that is, the items appeared to be harder than they were. The missingness level had no effect on the estimated δ_j. Finally, in recovering the item pseudo-guessing parameter (χ_j) none of the methods performed well regardless of the missingness mechanism, although IAS was the worst. Under MNAR, a fourth index was estimated: the proportion of correct responses. The value of p was higher than its true value for all the approaches except IAS. The smallest discrepancy was given by MIDA and the largest discrepancy by EM.

Finally, three remarks are relevant to this study. First, when generating the missing data, simulees were more likely to be assigned missing values if the original answer was incorrect. Second, the imputed values obtained via MI were rounded to the nearest 0 or 1, although other researchers have shown that this rounding produces biased estimates (Ake, 2005; Allison, 2006; Horton et al., 2003). Third, Finch actually generated

complete data with the EM algorithm even though this is regarded as bad practice and is not recommended (Enders, 2010). Finch acknowledged that the last two points may have had an effect on the performance level of MIDA and EM.

More recently, Wolkowitz and Skorupski (2013) explored the idea of polytomous IRT model used in the imputation phase to study item statistics' robustness. They approached nonresponses imputation using the multiple-choice model (MCM, Thissen & Steinberg, 1984). The MCM is an (3*K*-1) parameter model where *K* is the number of item categories. The MCM model assumes that students who do not know the answer to an item are attracted to the options at different rates. Thus, they do not choose the answer totally at random. Therefore the option a student chooses carries information about his or her ability. The MCM has parameters for all the item's options, including a guessing parameter.

Multiple imputation with MCM (MCM-MI) uses the probability of an examinee selecting a specific option to impute his or her actual response option (e.g., a, b, c, or d for $K = 4$). The authors evaluated traditional item difficulty (i.e., proportion of correct responses) and the item-total correlation between MCM-MI and LD. To do this, they used simulated data with N=20,000, J=44 multiple choice items with m=5, and 16.5% of data missing according to one of the three missingness mechanisms (MCAR, MAR, MNAR). Results showed that for MAR and MCAR both LD and MCM-MI performed similarly, with negligible difference. However, when the data were MNAR, MCM-MI generated more accurate item statistics than LD. In the case of *p*, LD was 7.5 times more biased than MCM-MI. The inter-item correlation was 1.3 times more biased with LD

than MCM-MI (Wolkowitz & Skorupski, 2013). Unfortunately, the accuracy of item and person parameters recovery was not studied.

Missingness as latent variable. Another approach for handling missing data is assuming they are related to some sort of latent variable. O'Muircheartaigh and Moustaki (1999) was the first study doing this. In their work, they treated missing data with ML methods using response function. Their approach, called symmetric pattern model, does not categorize variables as dependent or independent. Therefore, the missing responses are not predicted by other variables. The models are pattern-based because both the item responses and nonresponses patterns are relevant. In the symmetric pattern model, two dimensions are thought to be present: the attitude (θ) and the response propensity (ξ).

The instrument is supposed to directly measure the theta and the response propensity represents the examinee's disposition to respond to the items in the instrument. They modeled the response function for an item (x_j) considering three different missingness mechanisms: (a) MCAR, where the probability of a missing response is constant across individuals; (b) MAR, where the probability of a missing response depends on ξ; and (c) MNAR, where the probability of a missing response depends on both θ and ξ. That is, with MNAR the likelihood of answering an item depends on the examinee's position on both attitude and response propensity, whereas the response itself depends only on the examinee's θ.

These probabilities are based on three coefficients, r_{j0}, r_{j1}, and r_{j2}, that define the effect of θ and ξ on the examinee's response function (O'Muircheartaigh & Moustaki, 1999). Assuming the following response function:

$$logit\{\pi_j(\boldsymbol{z})\} = \delta_j + \sum_{i=1}^{q} \alpha_{ji} z_i, \quad (4)$$

where $\pi_j = P(x_j = 1|\boldsymbol{z})$; $\mathbf{z}$ represents the q latent variables (θ and ξ); δ_j are the difficulty parameters; and α_{ji} are the discrimination parameters. When q=1 (i.e., one latent variable), and the item discriminations are equal to 1 ($\alpha_{ji} = \alpha = 1$) the model becomes the unidimensional Rasch model. The response function unfolds into two layers, where missing response is denoted by 9:

$P(x_j \neq 9|\theta, \xi) = \pi_{\xi j}(\theta, \xi)$ for each response item

$P(x_j = 1|\theta, \xi, x_j \neq 9) = \pi_{\theta j}(\theta)$ for each attitude binary item

The model for the attitude binary item is $logit\{\pi_{\theta j}(\theta)\} = \delta_j + \alpha_{j1}\theta$. Despite the missingness mechanism the response item section of the response function depends on the assumptions about the missing data mechanism:

For MCAR, $logit\{\pi_{\xi j}(\theta, \xi)\} = r_{j0}$

For MAR, $logit\{\pi_{\xi j}(\theta, \xi)\} = r_{j0} + r_{j2}\xi$

For MNAR, $logit\{\pi_{\xi j}(\theta, \xi)\} = r_{j0} + r_{j1}\theta + r_{j2}\xi$

The coefficients (r_{j0}, r_{j1}, and r_{j2}) can be zero, positive, or negative. If both r_{i1} and r_{i2} are equal to zero, the response is MCAR. If r_{i1} = 0, then the response is missing at random. Values of r_{i1} that are further away from zero mean that (a) an individual with a high position on the θ scale is more likely to respond to the item, and (b) more information about θ can be inferred from the not-answered item. The sign of r_{i1}, depends on both how the attitude is being measured (i.e., direct or reversed wording) and the question's sensitivity. For example, in a rating scale measure a high positive value of r_{i1} would mean that the attitude is measured directly. If the question is sensitive, the

participants who agree with the item's statement will have more willingness to answer that question than the ones who disagree. Thus, the missing response would probably be "imputed" as disagree.

Finally, O'Muircheartaigh and Moustaki (1999) said that this approach works with binary, continuous, and mixed (binary and continuous) variables with missing data. However, they found that the uncertainty of what the imputed value must be is greater with binary variables. This uncertainty is reduced with mixed variables and, thereby, the "strength of the predicted scope of the model with respect to the missing value" increases (p. 187). The authors pointed out that this approach is compatible with attitude scaling given that each item provides information about the other items in the scale.

Based on this antecedent, Glas and Pimentel (2008), Holman and Glas (2005), and Pimentel (2005) modeled the missingness mechanisms in the IRT "language." That is, they assumed a second latent trait (or dimension) that determined the probability of answering an item (i.e., response propensity). The relationship between the response propensity (ξ) and the latent construct (θ) that is of interest to measure determines the missingness mechanism. Thus, they should both be included in the IRT data analysis.

Holman and Glas (2005) studied item parameter recovery in the IRT context when the missingness mechanism is taken into account and when it is not. Their proposal starts from the IRT framework for missingness. They say that the missing data suitable for IRT analysis can be grouped in four different categories. The first category, MCAR, is when the data were set to be missing by design in the study. The missingness obtained in the adaptive test, two-stage, and multi-stage testing is the second category. In this case,

the data are missing according to the person's response and it said to MAR. The third category refers to the missingness caused by the "don't know" or "not applicable." This missingness is not related to the ability being measured and thus is another type of MAR. Finally, the last category also has an unknown cause, but in this case the nonresponse is dependent on the person ability (MNAR).

According to these categories, Holman and Glas presented four different IRT-based approaches that model the missingness mechanism, three of them being different cases of MNAR and one of MAR. Basically, Holman and Glas (2005) assume that there are two latent traits that drive the person's decision about whether to answer an item. The first one is the response propensity which represents individual characteristics (e.g., personality trait, omission propensity, etc.) that affect the person's propensity to answer an item. This latent trait is not measured by the instrument. The second latent trait is the person ability that is measured with the instrument. The four approaches are based on the dependence of both observed response and nonresponse on these two latent traits, and the relationship between them, $\rho(\theta, \xi)$.

In model 1 (G_1), the probability of a particular observation for the person i on the item j, x_{ij}, depends on θ; the probability of a particular nonresponse (d_{ij}) depends on ξ; and there is no relationship between θ and ξ, $\rho(\theta, \xi) = 0.0$. In this case, the data are ignorable (missing at random). The nonignorable missingness is modeled in three different situations. In all the situations it is assumed that θ and ξ have a common distribution. In model 2 (G_2), the probability of x_{ij} depends on θ; the probability of d_{ij} depends on ξ; and there is a relationship between θ and ξ, $\rho(\theta, \xi) \neq 0.0$. In model 3 ($G_3$),

the probability of x_{ij} depends on θ, but the probability of d_{ij} depends on both θ and ξ, and $\rho(\theta, \xi) \neq 0.0$. Finally, in model 4 ($G_4$), the probability of x_{ij} and d_{ij} depends on both θ and ξ, and $\rho(\theta, \xi) \neq 0.0$.

Although these "missingness models" can be combined with different data analysis models, the authors only worked with them in conjunction with IRT analysis models (GPCM, PCM, 2PL and Rasch) using MML. Factors for the study with simulated dichotomous data and 50% of the data missing were sample sizes (N=500, 1000, and 2000), test length (J=10, 20, and 30), and levels of $\rho(\theta, \xi)$ (0.0, 0.1, 0.2… 0.9). The closer $\rho(\theta, \xi)$ is to 1 (normally positive) the more likely the missingness can be considered nonignorable. Holman and Glas showed that the higher the $\rho(\theta, \xi)$ the greater the bias of item parameter estimates if the missingness is treated as ignorable when analyzing the data. That is, ignoring nonignorable missing data yielded biased estimators. The bias in the item parameter estimates recovery was higher for shorter tests and small sample. However, these biases can be reduced by incorporating or modeling the missingness through the above described models.

Additionally, Holman and Glas (2005) tested the proposed models with empirical data (32 five-point rating scale items) in two ways. First, they modeled the missingness with the Rasch model and the observed data with the PCM. Second, the missingness was modeled with the 2PL and the observed with GPCM. In both ways, they used the four missingness models and an additional model in which there was only one latent trait (i.e., $\rho(\theta, \xi) = 1$) that determined the probabilities of both x_{ij} and d_{ij} (named G_0). They found that (a) the GPCM fitted the observed data better than the PCM and that G_3 (with GPCM)

and G_4 (with GPCM and PCM) were more efficient at modeling the missingness process than G_0, G_1 and G_2; (b) the estimated $\rho(\theta, \xi)$ was larger than the data-based $\rho(\theta, \xi)$ coefficient; and (c) the model has a better fit when the missingness is modeled (G_1 to G_4) than when it is not (G_0).

Pimentel (2005) wrote a couple of chapters (2 and 3) for his doctoral dissertation related to missingness modeling with IRT. Chapter 2 refers to non-speeded tests and Chapter 3 combined this phenomenon with not-reached responses. Pimentel's chapter 2 (2005) simulated diverse conditions, some of which were studied by Holman and Glas (2005), such as MML estimation, N=500, J=10 items, dichotomous and polytomous (with m=3), MAR and MNAR mechanisms, 25% and 50% of missingness level, and different $\rho(\theta, \xi)$=0.0, 0.4, and 0.8. However, Pimentel included a couple of additional conditions such as the bi-dimensionality (ξ_1, ξ_2) of the missingness data process, whether these dimensions were considered in the model, and the extent to which these dimensions were related, $\rho(\xi_1, \xi_2)$ = 0.0, 0.4, and 0.8. Also, he analyzed the effect of incorporating covariates in both models (i.e., the missingness model and person ability model), assuming a linear association between the covariates and the latent variables (θ, ξ_1, ξ_2). Pimentel used the multidimensional 2PL and multidimensional PCM for the observed data modeling for dichotomous and polytomous items, respectively. Given that the missing data are based on a binary matrix that records the response or nonresponse of the examinee i on the item j as 1 or 0, respectively (d_{ij} = 1 or d_{ij} = 0), the missing data process was modeled with the multidimensional 1PL in both item formats.

Pimentel's (2005) conclusions aligned with those presented by Holman and Glas

(2005): (a) ignoring nonignorable missingness increased the error in the estimation of item parameter; (b) modeling the missingness, even partially, contributed to an improvement in estimation; (c) the higher the non-ignorability condition of the missing data (i.e., larger $\rho(\theta, \xi)$), the greater the bias in the estimates if the missingness was not modeled; (d) the use of covariates (for both θ and ξ) improved estimation, regardless of the missingness level and mechanism, item format, or latent correlation, $\rho(\theta, \xi)$; (e) the use of covariates contributed to the efficiency even under MAR condition; (f) the use of covariates only (i.e., not modeling the missingness or ignoring the $\rho(\theta, \xi)$ value) reduced the bias in estimation (i.e., nonignorable missingness can be ignored if covariates are included in the parameter estimation, but a combination of both missingness modeling and covariates improved the estimation even more); and (g) there is not clear effect of $\rho(\xi_1, \xi_2)$ values.

Chapter 3 of Pimentel's dissertation was published as Glas and Pimentel (2008). This paper also studied the efficiency of IRT models that incorporated missingness mechanisms in the data analysis. Not-reached responses were the type of missing data they dealt with in this study. They assumed that speediness and ability were related and thus the missing data were not ignorable (MNAR). Glas and Pimentel evaluated: (a) the position effect of dichotomous items in the test in their parameter estimation quality; (b) the effect of sample size, test length, and non-ignorability intensity on item parameter bias with dichotomous items; and (c) whether this approach applied to polytomous items. In all cases MML estimation was utilized. The models utilized in this study were the sequential or step model (Tutz, 1990, 1997; Verhelst et al. 1997), both uni- and bi-

dimensional 2PL, and GPCM according to the item response format. Finally, the missingness models were based on the same approach as Holman and Glas (2005).

To study condition (a) and assuming $\rho(\theta, \xi)$=0.8 with 50% of MNAR, they had two scenarios: one in which the nonignorable missingness was explicitly modeled as if it were a second dimension on which the items load (scenario 1), and another in which it was ignored; that is, MNAR was assumed to be MAR (scenario 2). In scenario 1, the bi-dimensional 2PL was used in the estimation, whereas the parameters in the scenario 2 were estimated with the unidimensional 2PL. To study condition (b), the simulated data had N=500 or 1000, m= 10 or 40, and MNAR= 25% or 50%. The non-ignorability intensity was $\rho(\theta, \xi)$ =0.0, 0.2… 0.8. When $\rho(\theta, \xi)$=0.0, the data were MAR. Like in the previous condition, the data were modeled assuming that nonignorable missing data were MAR or MNAR. The step model was used for the missing data and the 2PL for the observed responses. For condition (c), the dataset was comprised of polytomously scored items with 4 response categories. Factors were N=500 or 1000, m=10 or 40, MNAR= 25%, and $\rho(\theta, \xi)$=0.2… 0.8. Again, the data were modeled assuming that nonignorable missing data were MAR or MNAR. The step model was used for the missing data and the GPCM for the observed responses.

Overall, the results showed that the item's position in the test affected the quality of its parameter estimates whether missingness was modeled or not (scenarios 1 or 2). That is, the parameters of items located towards the end of the test showed higher standard error, on average. Also, as found with previous research (Holman & Glas, 2005; Pimentel, 2005), correctly modeling the missingness mechanism improves item

parameter estimates, regardless of their format (dichotomous or polytomous), sample size, test length, or non-ignorability intensity. However, non-ignorability intensity and the proportion of not-reached responses positively impacted the bias level if the missingness was not explicitly modeled. That is, as $\rho(\theta, \xi)$ increased, the bias in the item parameter estimation increased if data were treated as MAR. The sample size counterbalanced this effect such that the bias was less when *N* was higher, regardless of the non-ignorability intensity and the proportion of not-reached response.

When nonignorable data were ignored, the parameters of items located in the second half of the test were consistently underestimated with item difficulty more severely affected. Glas and Pimentel (2008) argued that the high bias in the parameter estimates for these items showed that models that assumed MAR when the data were in fact MNAR do not accommodate differences in the variability of the proficiency level. When the not-reached responses are nonignorable, the participants that answer the last items are supposed to be more proficient than the ones that did not. Hence, this difference should be taken into account when estimating parameters. This is exactly what does not happen when nonignorable missingness is modeled with MAR assumption. An additional comment by Glass and Pimentel was that this approach may be seen as a test of goodness of fit. The main limitation of this study is that the authors ignored the omitted responses, when theory normally assumes they are also MNAR.

Likewise, a study conducted by Rose, von Davier, and Xu (2010) evaluated the performance of different missing data techniques within the large-scale achievement assessment context. Using both, simulated and empirical data (PISA, 2006) they

compared six IRT models. Some of the models (models 2 to 4) used deterministic treatments for missing data. Whereas others (models 5 to 7) approached missing responses through what the authors called the latent response propensity. The idea behind the stochastic approaches considered by Rose et al. were the same as Holman and Glas (2005), Pimentel (2005), and Glas and Pimentel (2008).

The first model (*M1*) is a unidimensional IRT that uses the simulated cases without missingness. *M1* served as benchmark for evaluating the other models. In the second model (*M2*), the missing data were ignored in person and item estimation, whereas in the third model (*M3*) the missing data were scored as incorrect in the analysis. Model 4 (*M4*) was based on the two-stage approach that was originally suggested by Ludlow and O'Leary (1999) and that is normally applied by agencies that do large-scale assessment (PISA, TIMSS, and PIRLS). In this model, the missing data were ignored for the item calibration, but treated as incorrect for ability estimation. However, in PISA, TIMSS and PIRLS only not-reached responses are treated as ignored in the first stage and as incorrect in the second stage. Omitted data are always treated as incorrect in the two stages. This study did not distinguish between these two types of missingness.

The fifth model (*M5*), called the latent regression model, had missingness modeled via the observed response rate of person *i* ($\bar{d}_i$) and was used as a θ predictor. The sixth model (*M6*), called the between-item multidimensional IRT model, incorporated two dimensions, θ and ξ, to model the missingness. With this model the probability of correctly answering the item was weighted by the response propensity for the item. The last model (*M7*) was the within-item multidimensional IRT model, in which

the missingness was incorporated into the item by adding an additional item discrimination parameter.

For the simulated data, the conditions were 1000 cases with 26 dichotomous items, two levels of missingness (x_{miss}=30% and 49.81%), and $\rho(\theta, \xi)$=0.622 and 0.80. It was assumed that the missingness mechanism was nonignorable in all the cases. The analysis of the simulated data were done using the Rasch model. EAP was used to estimate the θ estimates. With the simulated data *M2* to *M7* were compared to *M1*. Rose et al. (2010) evaluated the recovery of both item and person parameter estimates using bias, standard error, and mean square error values. Additionally, the authors reviewed the correlation between the true and the estimated ability, $r(\theta, \hat{\theta})$, and two reliability coefficients: $r(\theta, \hat{\theta})^2$ and the EAP reliability.

The results showed that all the models, except *M3*, returned good item parameter estimates, albeit overestimated. There was no difference in the performance of the stochastic models. *M2* was the most accurate, when x_{miss} = 30% and $\rho(\theta, \xi)$ =0.622, but it did not do well when x_{miss} = 49.81% and $\rho(\theta, \xi)$ =0.80. The item parameter conclusions for *M2* also apply to *M4*, because *M4* is a two-stage approach and it is the same as *M2* in the first stage. *M3*, which assumed the missing data to be incorrect was the worst. It exhibited around 15 and 120 times more bias than the other models when x_{miss} =30% and $\rho(\theta, \xi)$ = 0.622, and when x_{miss} =49.81% and $\rho(\theta, \xi)$ =0.80, respectively.

Using EAP, Rose et al. compared the models at the person ability level. Here again, the models' accuracy was the same for the stochastic models. *M2* and *M3* were more accurate recovering person ability than recovering item parameters. An interesting

result was the performance of *M4* in its second stage (θ estimation). Even when the item parameter bias was the smallest with *M4*, the $\hat{\theta}$ based on the item parameter estimated (stage 1) greatly underestimated the true ability. This conclusion was the same for x_{miss}=30% and 49.81%.

The values of $r(\theta, \hat{\theta})$ for the different models were all above 0.78. The highest correlation was for the stochastic models, which had the same value despite the missingness level. When x_{miss} =30% and $\rho(\theta, \xi)$ =0.622, the lowest correlation was reported for *M3* and *M4*, and when x_{miss} =49.81% and $\rho(\theta, \xi)$ =0.80, *M2* had the lowest correlation. By comparing the values of the two reliability coefficients, $r(\theta, \hat{\theta})^2$ and the EAP reliability, it was possible to detect if the models were overestimating ($r(\theta, \hat{\theta})^2$ < EAP reliability) or underestimating ($r(\theta, \hat{\theta})^2$ > EAP reliability) the reliability. From this analysis, the authors found that ignoring the missing data (i.e., using *M2*) underestimated the reliability, *M3* and *M4* overestimated the model-based reliability with *M3* being worst. The stochastic models did not under- or over-estimate the reliability when x_{miss} =30% and $\rho(\theta, \xi)$ =0.622. With higher level of missingness and latent traits correlation (49.81% and 0.80, respectively), only *M5* produced a negligible underestimation of the reliability coefficient (Rose et al., 2010).

In summary, treating the missing data as incorrect yielded either overestimated item parameter if all the parameters were simultaneously estimated (*M3*), or underestimated ability if this parameter was obtained in a second stage (*M4*). This means that the estimated reliability coefficients will be spuriously high for *M3* and *M4*. Finally, *M2* was robust when x_{miss} = 30% and $\rho(\theta, \xi)$ = 0.622, although it underestimated the

reliability of the ability estimates. The stochastic models performed the best in both item and person parameters estimation and yielded the highest correlation and most precise reliability coefficient. None of the three stochastic model could be chosen as the best (Rose et al., 2010).

For the empirical data, only models *M2* to *M6* were reviewed. Given the complexity of PISA 2006, these models were adjusted to include: (a) a latent ability that was multidimensional (θ_i) to incorporate math, reading, science, and ξ; (b) for *M5*, the observed response rate was stratified into groups according to certain ranges (low, medium and high); (c) the country was included as a variable in the models (multi-group models); (d) the item parameters were constrained to be the same across countries; and (e) means, variance, and covariances were allowed to vary across countries (Rose et al., 2010).

The models were evaluated with the PISA 2006 data using variants based on the 2PL model. In this analysis, the conclusions were almost the same as with the simulated data. Specifically, *M2* and *M5* had the best and similar performance at the item parameter level (discrimination and difficulty), with *M6* being the second best. For the person ability estimation, the accuracy of the models was measured with the conditional expectation of the latent ability, conditional on the country (g), $\mathcal{E}(\theta_k \mid g)$. *M3* and *M4* underestimated the $\mathcal{E}(\theta_k \mid g)$, whereas *M2, M5, M6* performed almost equally. However, there was no evidence that the stochastic models outperformed *M2* (Rose et al., 2010).

Overall, Rose et al.'s conclusions are: (a) the stronger the relationship between the latent ability and the latent propensity response, the less ignorable the missingness is and

the more biased the estimates are; (b) stochastic models "adjust the EAP ability estimates selectively, due to the pattern of missing data, and correct for the unfair benefit of the systematically skipped items" (p. 43); (c) the higher the missingness level, the less efficient the unidimensional models are, although they perform well even with 30% of missingness; (d) treating missing data as incorrect is both unfair and inefficient. This approach distorts parameter estimates, underestimates reliability, and "tends to penalize respondents who actually might have solved the items" (p. 42); (e) stochastic models, although not outperforming the simpler ones in this study in terms of parameter estimates, offer the chance of determining the relationship between proficiency and non-response (via the reliability of ability); and (f) *M5* could be a good stochastic model given that there are greater computational costs with *M6* and *M7* than with *M5*, and the results are pretty much the same. A limitation of this study is that they did not differentiate between omitted and not-reached responses in neither the simulated data nor PISA, 2006 data. The authors assumed that both omitted and not-reached responses were equivalent.

Abad, Olea and Ponsoda (2009) evaluated the quality of parameter recovery for multiple-choice item responses with omitted data and polytomous IRT models, like the MCM, the Samejima-MCM (SMCM), and the nominal response model (NRM). Polytomous IRT models' primary advantage is the additional information obtained from the different item categories (Drasgow, Levine, Tsien, Williams, & Mead, 1995; as cited in Abad et al., 2009). Nevertheless, they have some disadvantages such as the lack of guessing parameters in some models (e.g., NRM), lack of unique parameters estimates, a relatively large number of item parameters to be estimated (as is the case for MCM and

SMCM), the requirement of a large sample size, and the fact that these models do not address omitted response issues (Abad et al., 2009).

Abad et al. suggested a restricted-SMCM (RSMCM) that addresses all these limitations of the three previous models. The key characteristic of this approach is that examinees are split into *G* groups with homogeneous propensity omission, that can be calculated as O_i=omits/ (omits+incorrect). Therefore, RSMCM assumes that there are different omitting propensity groups according to the omission level the examinees have. However, splitting the sample into groups could also be considered as the model's drawback. In large-scale assessment, there may be a large number of sub-groups that can complicate the data analysis. In the RSMCM, the probability of omission is the same within the group and the omitting propensity for an examinee is constant across the items. Finally, RSMCM also assumes that the response propensity is conditional on both the θ and the omitting propensity of the group the examinee is a member of (Abad et al., 2009).

The authors studied the performance of the four models using both empirical and simulated data. For the empirical data analysis (*N*=3,224 examinees, *J*=20 items, *m*=4), four groups were formed based on the propensity omission of examinees. Item parameters were estimated using MML. For NRM, SMCM, and MCM the omitted responses were coded as an additional response category. Two different levels of analysis were used to compare the four models' performance: the item-fit and model-fit analysis. The item-fit analysis revealed that 14 out the 20 items fit the model at some level (moderated or full) when the analysis was done with RSMCM. When working with the other three models almost all the items (19 out of 20) failed to fit the model when the

parameters were forced to be equal (Abad et al., 2009).

In terms of model-fit, the RSMCM showed the smallest AIC value, meaning that it fit the data better than the other three models. Person and item parameters for the different groups were estimated with only the 14 items that fit the model for all the four approaches in the study. Again, Abad et al. found that RSMCM had the best model-fit with this subset of items. Also, the mean difference analysis among groups showed that these groups had different ability levels. Therefore, RSMCM was shown to most accurately depict this situation. The authors also found that "lower ability examinees' omission rates cannot be estimated so reliably in the MCM and SMCM as in the RSMCM" (Abad et al., 2009, p. 210).

For the simulated data analysis, two factors were manipulated: sample size (N=1000, 2000, and 3152) and test length (J= 14 and 28). The items also had four response categories and four different groups were generated according to their omission propensity. Two sets of data were generated, the first one with the RSMCM and the second set with the other three models (NRM, MCM, and SMCM). Person ability was estimated using MAP. Both person and item parameters recovery were compared using different indices. Abad et al. (2009) found that sample size and test length affected item parameters recovery. The larger the sample and the longer test, the lower the difference between the estimated and the true item parameters. Additionally, the models were not equally effective when recovering item parameters. Contrary to what was expected, the NRM was found to be the best in item parameter recovery, but did not do so well with the theta estimation. The worst item parameter recovery was obtained with the MCM. On the

other hand, theta recovery was the best with RSMCM, when the data were generated with this model. When the data were generated with the other three models, RSMCM performed as well as the NRM, MCM, and SMCM in the θ recovery. Thus, RSMCM can be a good option when examinee's guessing strategies were not properly considered in the model (Abad et al., 2009).

CHAPTER III: METHODS

The 6th grader' mathematics dataset from the SERCE are used in this study. The domain was chosen because sixth grade students had a lower performance in this area than in others measured by SERCE. In 2006, the SERCE was administered to third and sixth grade students from sixteen countries and the Mexican state of Nuevo Leon. Around 196,000 students from the participant countries were evaluated in mathematics and reading and writing. A third area, sciences, was also administered to sixth graders. However, participation in this domain was not mandatory. Thus, only nine countries and Nuevo Leon have been assessed in science. Additionally, information regarding students, classrooms, and schools characteristics was collected. The SERCE is representative at the national, urban, and rural levels.

Overall, SERCE developed 45 instruments to accomplish its purpose. Of them, 34 are achievement tests, and 11 are background questionnaires. The instruments were paper-based and written in two languages: Spanish and Portuguese. All the instruments were previously piloted before their final use. The psychometric analyses were done using the Rasch model, Classical Test Theory and included differential item functioning. The items retained after the pilot testing were distributed in six blocks per domain.

These blocks were combined to form six booklets, such that each of them would contain two blocks. Test administration followed a balanced incomplete blocks design and was limited to 60 minutes for reading and science, 45 minutes for writing, and 70 minutes for mathematics. An additional 10 minutes was offered to students if they requested extra time to finish the test (Valdez et al., 2008). All the parameters were

estimated using the whole data (i.e., calibration was not done per country). Students' scores are based only on the booklets they completed. The scores were estimated with Winsteps, which uses JMLE.

This study uses the data of 6th grade students' performance in mathematics because sixth grade students had a lower performance in this area than in others measured by SERCE. Also, because 6th grade tests contained more items than did the 3rd grade instruments. This test has a mixed-format design and 96 items. Therefore, each booklet has 29 multiple-choice (MC) items and 3 constructed-response (CR) items. Both MC and CR answers range from 1 to 4; MC items are dichotomously scored, whereas CR items are polytomously scored. CR items were rated under a nested design in which raters scored across items, but not across students.

Missingness level

SERCE study identifies five different sources of missingness at the item level in the achievement tests: (a) missing by design (called not-administered in the literature); (b) not legible due to printing issues; (c) not-administered (SERCE assigned this category to the items that had some issues in the printed booklet); (d) invalid; and (e) not-answered. The first three categories are assumed to be missing completely at random and therefore ignorable. The item response is classified as invalid if the student selected more options than expected (i.e., multi-marks).

SERCE does not differentiate between not-reached and omitted responses; thus, category (e) includes both. Not-reached responses are generally considered ignorable, whereas omitted responses are intentionally not answered by the participant (Lord, 1973,

1980; Mislevy & Wu, 1988, 1996). This distinction is pertinent to the missingness mechanism that is observed in each case (MAR and MNAR, respectively), and consequently in the treatment approach to be employed. Omitted responses are the intermediate missing values observed in the person's response pattern and not-reached responses are the consecutive blank items clustered at the end of the test. There are no valid responses after not-reached items.

Two issues emerge when distinguishing between these two levels of missingness. First, there is no reference (or rule of thumb) about the proportion of the items in a test that can be considered not-reached. When considering the definition and following the common practices of coding this type of cases,[15] not-reached items can be found as early as in the second item in the response pattern in the SERCE study. Note that the concept of not-reached responses is linked to the time the students have to take the test. Thus, it is logic to think that only the last items are under this risk.

This situation leads to the second issue; that is, the number of "non-answered" items that were actually not reached because of lack of time. This is related to the dependency between speededness and ability. If they are related, then the assumption of random missingness cannot be tenable (i.e., the higher the number of blank responses categorized as not-reached, the higher the likelihood of them being conditional on the participant's proficiency). After all, low ability students tend to spend more time solving each item, given their comprehension limitation (Mislevy & Wu, 1988; Glas and Pimentel, 2008). There is some empirical evidence about the relationship between not-reached responses and ability. Specifically, van den Wollenberg (1979) found a

significant correlation between percent-correct scores for the reached items and the total number of reached items (cited in Mislevy & Wu, 1988). For this reason, no distinction between omitted and not-reached responses are made in this study.

The proportion of not-answered (i.e., omitted and not-reached) in 6th grade mathematics dataset can be measured in different ways. If considered as a matrix, 1.7% of the cells contain missing responses, whereas the missingness per item (or missingness by participant) ranges between .8% and 32.4%. Finally, the proportion of students with at least one nonresponse in mathematics is 48%. That is, almost half of the participant students from 6th grade did not answer at least one item in this domain.

Data generation for the missing analysis

For the purpose of this study, SERCE mathematics 6th grade MC and CR items data are dichotomously scored as correct (code=1) or incorrect (code=0), partial credits are considered incorrect as well as invalid responses. Not-answered items were `considered omitted (coded=2) and other sources of missingness (i to iii from missing data section) were considered ignorable. As said above, each of the six booklets has 32 items. Cases with more than 10 missing responses were removed from the dataset (2.6% loss). Data calibration was conducted using the Rasch and 2PL models. There were items for that did not perform well (i.e., the item parameter estimates had extremely high values) and thus were removed as presented in the Appendix B. This resulted in a varying number of items per booklet (between 13 and 31, see Table 2 in the result section) and per model (57 and 53 items for the Rasch and 2PL models, respectively). Consequently, the results for the two IRT models cannot be compared. The next step was to extract

missingness pattern from the dataset to be used with each IRT model, following a set of steps:

Step 1. The dataset was split into two files per booklet: the complete-data (C_b where b=booklet=1, 2, …, 6) and incomplete-data (I_{Ab}) files. The complete-data file contains the participants who answered all of the items in a booklet, whereas the incomplete-data file contains participants who have between 1 and 10 non-responses. The two datasets were compared using an algorithm,[16] which is based on Needleman and Wunsch algorithm used in genetics. The algorithm reported the level of match between all possible pair of cases in both files. The algorithm compared a case with a complete set of answers to the cases with omitted responses on an item-by-item basis. Differences in the responses across the two files could be due to different observed answers (e.g., one person answered correctly and the other incorrectly) or due to non-observed response (e.g., one person provided an answer and the other did not). The algorithm did not differentiate amongst these two differences, so the output was later analyzed and cases with differences due to different observed answer were discarded.

The complete-data and incomplete-data files were matched by booklet because the items were not the same across the booklets. The percent match was calculated based on the observed responses the incomplete cases have. For example, if the case x_{1i} with incomplete response pattern has 28 observed responses (i.e., 4 missing values) and 25 matched responses (i.e., 3 responses are different) with case x_{1c} which has complete response pattern, the percent match for these two pairs of cases is 89% (25/28). Only cases whose percent match was 100% were retained (i.e., the case in the example was

discarded). This yielded 6 complete-data files, one per booklet (C_b). They were used along with I_{Ab} to replicate the missingness pattern.

Step 2. Using the six C_b from step 1 a new set of incomplete-data files (I_{Bb}) was generated. This was done by removing specific responses from the complete-data file following the missingness pattern observed in their matched incomplete-data file (I_{Ab}). For example, if the case x_{2i} has its missing values in the 3rd, 7th, 23rd, and 31st items then its matched case, x_{5c}, had the same item responses removed from its pattern response. This procedure produced two datasets per booklet: the complete-data file (C_b) and its modification that contained omitted responses (I_{Bb}).

Step 3. The complete-data booklet files (C_b) were concatenated into one file (C) containing all the booklets. The same was done with the incomplete-data files (I_{Bb}) to create a single incomplete-data file (I_B). Both the complete-data file (C) and the incomplete-data file (I_B) each contained 17,126 cases. Dataset C was used to estimate the ability and item parameters. The item parameters were kept fixed for ability estimation in all conditions. Dataset I_B was used with the different missingness approaches to estimate the ability parameters and evaluate their performance against the ability parameters estimated from the complete-data file. The percentage of missingness per item in the I_B dataset used for the Rasch and 2PL models ranges between .6% and 67.6%. The correlation between the missingness level and the proportion of number correct in the C dataset was -.123 (p=.000) for both the Rasch and 2PL models. The correlation of missingness level and proportion of number correct in the I_B dataset reduced to -.025 for the Rasch dataset and became positive (.030) for the 2PL dataset.

IRT models

Item calibration. The complete-data file *C* (step 3) was calibrated using BILOG-MG (MMLE, Zimowski, Muraki, Mislevy, & Bock, 2003) for each of the two IRT models: the Rasch and the 2PL. These item parameters were taken as fixed in the person ability estimation stage.

Person ability estimation. EAP estimation was used for theta estimation in both the complete-data file and the dataset with imputed values. Thetas estimated with the complete-data file were the benchmark for the missingness approaches comparisons. For the theta estimation, an algorithm that uses EAP and that allowed the use of decimal numbers in response vectors was used.[17]

Missingness approaches

The midpoint imputation. In this approach, the missing values were replaced by .5 for the person ability estimation. Although this approach does not have a theoretical framework and may introduce additional measurement error, it was found to outperform other missingness techniques (De Ayala et al., 2001; De Ayala, 2003; De Ayala, 2006).

Treat as incorrect. In this approach, missing values were treated as incorrect for the person ability estimation. SERCE used this approach for both the item and person parameters estimation, using JMLE. In this research, however, only person estimation was evaluated.

MI with and without auxiliary variables. Multiple imputation was done using M*plus* 7.1 (Muthén & Muthén, 1998-2012). This software package used three different models for imputation: the variance-covariance model or MIDA (default), the sequential

regression model or FCS, and the regression model.[18] Continuous variables are standardized with mean zero and variance one (i.e., they are treated as normally distributed). After the imputation is done, the variables are again transformed to their original scale (Asparouhov & Muthén, 2010). Imputation of ordered categorical variable can be done in M*plus* by specifying which variables are categorical (Asparouhov & Muthén, 2010). Probit link is used to determine the category that will replace the missing value (Enders, 2015). These values are later rounded to the closest plausible category. As described in the previous chapter and, as literature points out, rounding imputed values does not seem to perform well.

M*plus* treats categorical variables as normally distributed if the specification is not included. In this case, the values do not necessarily match the categories and there may be negative values among the imputed ones (i.e., implausible values are possible). For this study, imputation of responses is done using MIDA and treating the variables as normally distributed. Subsequently, a probit function was used to transform the imputed responses into values that are within the range of plausible values (i.e., between 0 and 1).

The imputation of item scores was done per block, instead of by booklet in order to keep the original structure of the data (i.e., the planned missing values). A dichotomous vector indicating the block position (1 or 2) in the booklet was used in the imputation process. Five datasets per block were obtained (m=5) using this approach. These blocks were later combined into booklets for the ability estimation.

Auxiliary variables. MIDA imputation was done with and without auxiliary variables. The two auxiliary variables were ISEC (socioeconomic status) and ICEH

(household cultural and educational condition). These two indices were scores generated by LLECE using principal component analysis (Trevino et al., 2008). ISEC is comprised of parents' level of education, first language spoken at home, household physical characteristics, basic services, appliances, and number of books in the house. ICEH is based on reading habits, early childhood education attained, parental involvement in school-related activities, and perception of school quality (see SERCE technical report for more information about this). These auxiliary variables have 2% of concurrent missing values. That is, 342 cases have missing values on both auxiliary variables. These missing values were not imputed. There were 5 datasets imputed with auxiliary variables and 5 datasets imputed without them. The correlation coefficients between these auxiliary variables and the number correct per participant from the complete-response dataset ranged between .24 and .26.

These correlation coefficients were larger than the correlation values observed in the original dataset, which were less than .21 in both indices. The variability of these indices, however, did not differ much between the original dataset and the one used in this study. Additionally, the point-biserial correlation between the items from the complete response dataset and these two auxiliary variables ranged between -.04 and .26. Finally, the correlation between the proportion of omit and the auxiliary variables in the dataset used in this study was not significant for neither ISEC (r=-.068, p>.05) or ICEH (r=-.071, p>.05). The correlation values observed in the original dataset were between -.050 and 0 in both indices.

Evaluation criteria

Five criteria were used to evaluate the performance of the missingness approaches as described below. The estimation of these values was done using *R* (2014). All these indices were also calculated for the theta estimated using the complete response dataset (i.e., where none of the cases have missing responses). The outputs obtained from the complete response dataset are used as benchmark to evaluate performance of the missing data handling approaches.

Additionally, a variance partitioning analysis was done with the theta estimated using multiple imputations in order to decide how to compare the performance of this approach with the other missingness approaches. Little and Rubin (1987) explain that the variance of an estimate (e.g., estimated theta) across all *m* datasets generated using the multiple imputation approach is comprised of two components: the average within-imputation variance (i.e., within each dataset) and the between-imputation variance (i.e., between datasets).

In this study, the purpose of such an analysis is to determine the proportion of variance in the estimated theta that is accounted by for the between datasets variability. If this value is low, then the theta estimates are not that different from one another and thus an average across datasets can be used for the comparison with other missing data approaches. In contrast, if the differences between datasets are large, then the mean across datasets cannot be used for comparison purposes and the method may not be consistent through imputation iteration.

Signed difference. This refers to the difference between the estimated theta from

the incomplete-data file after applying the missingness approach and the theta from the complete-data file. That is,

$$diff_i = \hat{\theta}_{I,i} - \hat{\theta}_{C,i} \,, \tag{5}$$

where $\hat{\theta}_{I,i}$ is the theta for the person *i* estimated from the incomplete-data file after applying the approach *m* (midpoint imputation, SERCE imputation, or MI); $\hat{\theta}_{C,i}$ is the theta for the person *i* estimated from the complete-data file.

Root-mean-square deviation (*RMSD*). It is the difference between the theta estimated using the complete response dataset and the one using the dataset with missing responses after they were handled using each of the approaches. The square root puts the difference in the same scale as the parameter which makes easier the interpretation. This is an average estimated across all the cases within the dataset to summarize the difference between theta estimates for the whole approach.

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_{I,i} - \hat{\theta}_{C,i})^2} \,, \tag{6}$$

where *n* is the number of cases in the dataset (i.e., 17126) and all other terms are defined above.

Coverage. This is also a summary index, it is based on the maximum likelihood confidence limit estimator (Birnbaum, 1968 as cited in De Ayala, 2009). It refers to the percentage of cases whose theta estimated from the complete-data file is within the 95% confidence band of the theta estimated from the incomplete-data file after applying the specific missingness approach. That is,

$$coverage = \frac{1}{n} freq\left[\hat{\theta}_{I,i} - z_{(1-\alpha/2)}(SE_{I,i}) < \hat{\theta}_{C,i} < \hat{\theta}_{I,i} + z_{(1-\alpha/2)}(SE_{I,i})\right], \tag{7}$$

where $SE_{I,i}$ is the standard error of the estimated theta for person *i* from the incomplete-data file, after applying the specific missingness approach, and all other terms are defined above.

Average length of confidence interval. A procedure that has a similar or higher rate of coverage than another but yields substantially shorter intervals should be preferred over the other, because this translates into greater accuracy and higher power (Collins et al., 2001, p.340). This summary index is estimated by:

$$length = \frac{1}{n}\left[\hat{\theta}_{I,i} + z_{(1-\alpha/2)}(SE_{I,i}) - \hat{\theta}_{I,i} - z_{(1-\alpha/2)}(SE_{I,i})\right]. \quad (8)$$

Average standard error. This is defined as the average standard error of the estimated theta across cases within each dataset. The smaller the standard error the narrower the confidence interval and the higher the precision of the estimation (Thoemmes & Rose, 2014).

Between and within imputation variability. Little and Rubin (1987) described two variance components when dealing with multiple imputation: within and between variance as defined below. These values are also reported as part of the comparison between the MIDA imputation with and without auxiliary variables. These were estimated for all four evaluation criteria.

$$\bar{\gamma} = \frac{1}{m}\sum_{j=1}^{m}\bar{\gamma}_j, \quad (9)$$

$$B = \frac{1}{m-1}\sum_{j=1}^{m}(\bar{\gamma}_j - \bar{\gamma})^2, \quad (10)$$

$$W = \frac{1}{m}\sum_{j=1}^{m}S(\gamma_j)^2, \quad (11)$$

$$T = \left(1 + \frac{1}{m}\right)B + W, \quad (12)$$

where $\bar{\gamma}_j$ is the average criterion value (standardized bias, *RMSD*, coverage, average standard error) for imputed dataset *j*; *m* is the number of datasets imputed (*m*=5 in all the cases), $\bar{\gamma}$ is the average criterion value across all the imputed datasets; $S(\gamma_j)^2$ is the variance of the criterion for the imputed dataset *j*; *B* and *W* are the between and within variance, respectively; *T* is the total variance of the criterion; and $\left(1+\frac{1}{m}\right)$ is an adjustment for finite *m* (Little & Rubin, 1987). A large criterion variance indicates large variability in the theta estimation that is primarily due to the use of auxiliary variables in the imputation of the item responses.

CHAPTER IV: RESULTS

This section presents the data's descriptive statistics and the performance comparison of the different approaches used to handle missing values. The two IRT models used in this study are not comparable because the number of items differed for each model (Table 2). Therefore, performance of the missingness approaches is presented in two separate sections, one for each IRT model.

Table 1.
Distribution of participants and items per booklet and IRT model

	Rasch		2PL	
Booklet	Participants	Item	Participants	Items
1	108	31	108	30
2	4,626	16	4,626	15
3	4,277	10	4,277	8
4	2,099	20	2,099	22
5	4,317	15	4,317	13
6	1,699	22	1,699	18
	17,126	57	17,126	53

As previously mentioned, item parameters were estimated using the complete-response dataset and kept fixed for the ability estimation under the different conditions. These parameters are presented in Appendix C. For the Rasch model, the difficulty parameter ranges between -3.807 and 3.965. In the 2PL model, the difficulty parameter ranges between -3.105 and 3.461, whereas the discrimination parameter ranges between .335 and 3.375. Fourteen of these items have discrimination values that are less than one.

After imposing the missingness pattern on the dataset, the average percentage of missingness by item ranges between 0.6% and 64.7%. The item missingness level and the item difficulty is significantly positively associated for both the Rasch (r=.647, p=.000)

and 2PL (r=.540, p=.000) models. That is, the more difficult the item was, the more missing values were observed. The item discrimination in the 2PL model was non-significantly correlated with the item level of missingness (r=.200, p=.150).

Rasch model

The students' ability estimated using Rasch model on the complete-response dataset ranged between -2.99 and 2.45, with 71% percent of the students having a negative estimated theta. That is, most of the examinees showed low level of ability in this test (Figure 1). The mean of $\hat{\theta}$ was -.444 logits (SD=.821) and the average standard error of the estimated theta was .557 logits (SD=.056).

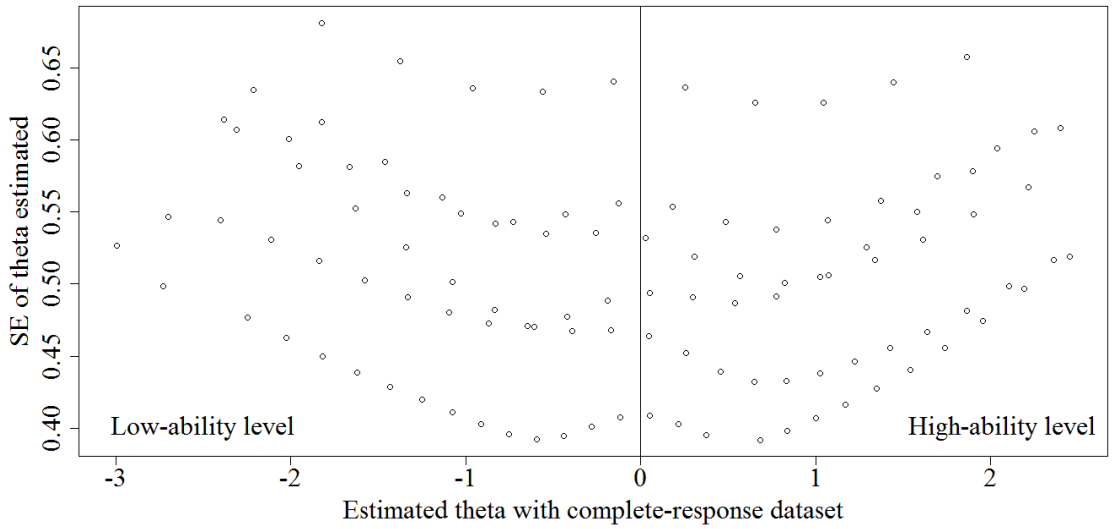

Figure 1. Distribution of estimated thetas and their standard errors using the complete-response dataset, Rasch model.

Additionally, there was a significant difference in the estimated standard error of $\hat{\theta}$ for those with high ability level (larger than zero) and for those with low level of ability

(less than zero), $F(1,17124)=91.46$, $p=.001$, $\eta^2=.005$, the difference was not meaningful. The standard deviation of the estimated error for the former was .057 whereas for the latter was .056. The average length of the confidence interval was 2.23 logits ($SD=.224$).

On the other hand, the correlation of both auxiliary variables with the $\hat{\theta}$s estimated from the complete-response dataset was significant. For the ISEC index, the correlation was .285 ($p=.000$) and for the ICEH it was .258 ($p=.000$). Also, the correlation between the $\hat{\theta}$s estimated from the complete-response dataset and the level of missingness ($M=.176$, $SD=.12$) in the corresponding incomplete-response dataset was significant, but relatively low ($r=-.036$, $p=.000$). This low correlation shows that the missing data cannot be linearly associated to the examinee's ability (Figure 2). In other words, the missingness mechanism might not be MNAR, but MAR or MCAR.

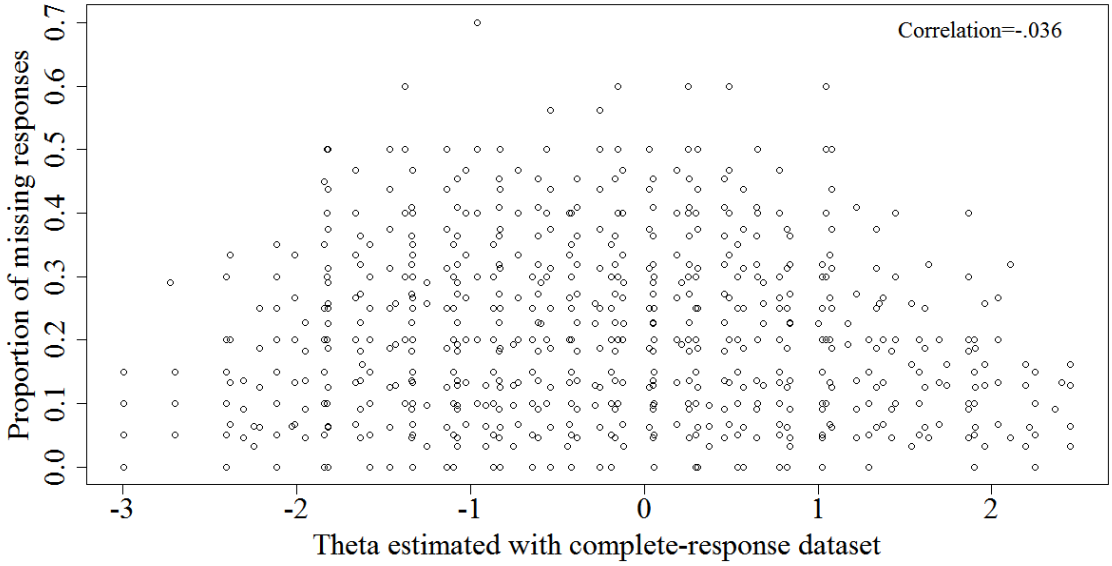

Figure 2. Correlation between ability estimated using the complete-response dataset and the proportion of missingness per examinee, Rasch model.

Between and within imputation variability. There were five datasets imputed with auxiliary variables, and another 5 datasets imputed without auxiliary variables. The between and within variance were estimated across each of these two sets of datasets for each of the criteria used in the comparison of missing data handling approaches. Overall, there was almost no variability in the indices estimated between datasets. Most of the total variance was within the data set (more than 99%).

Therefore, working with the mean of the imputed files will facilitate the comparison among conditions without compromising the results and the conclusions about the quality of the multiple imputation approach. That is, for each of the following indices the mean across all five imputed datasets without auxiliary variables and the mean of the five imputed datasets with auxiliary variables were reported.

Coverage. The coverage index was high (more than .995) for all the missing data handling approaches. That is, most of the estimated thetas from the complete-response dataset were within the 95% confidence interval formed by the estimated thetas from the datasets with missing values treated by the different missingness approaches.

Average length of confidence interval. The average lengths of the confidence interval (CI) for the estimated thetas were the same when the missing responses were imputed using multiple imputation with (*M*=2.208, *SD*=.224), without auxiliary variables (*M*=2.208, *SD*=.224), and the midpoint imputation (*M*=2.209, *SD*=.224). When the missing values were treated as incorrect the average length slightly increased by about 1.3% to *M*=2.237 logits (*SD*=.222) in comparison to the other three approaches.

On the other hand, the middle-point and the MIDA with and without auxiliary variables showed smaller average CI length, compared with the average CI length estimated with the complete response datasets. When the missing was treated as incorrect, the average length of the CI was slightly larger than the observed for the complete response dataset (see Figure 3).

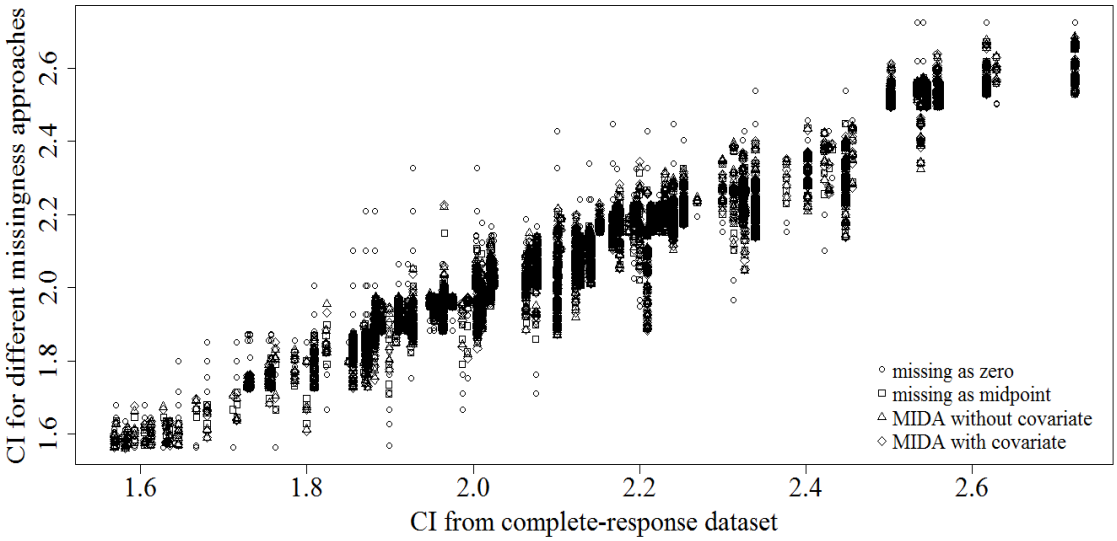

Figure 3. Confidence interval from the complete response dataset versus the CI estimated under the different missingness handling approaches, Rasch model.

Signed difference. The difference between the theta estimated with the complete-response dataset and when missing responses were treated as incorrect ranged between -2 and 0, which means that this approach tended to underestimate the thetas. However, only one third of the thetas were underestimated, the other two thirds of the thetas estimated using this approach were the same as the estimated with complete-response dataset (i.e., the difference was zero). The reason for the high proportion of estimation with no error is

that these cases were imputed with a response that matched the complete-response dataset. In other words, imputing the missing response with zero was correct for 67% of the cases. On average, the thetas estimated using this approach were underestimated by .132 logits (*SD*=.217).

When only the cases with underestimation are considered, the average underestimation was .398 (*SD*=.192). This means that when the original answer was 1, imputing the response with zero underestimated the examinee's ability level. The difference was inversely related to the theta values from the complete-response dataset (*r*=-.364, *p*=.000), which means that underestimation was larger at higher levels of ability (top of Figure 4). The accuracy in the estimation of ability using this approach was inversely related to the number of missing responses the student had. That is, the differences related to the theta estimates were larger when there were more missing responses (*r*=.376, *p*=000). Somewhat related to this is the fact that the correlation between level of missingness and estimated theta using this approach was greatly inflated from -.036 to -.144 (bottom of Figure 4).

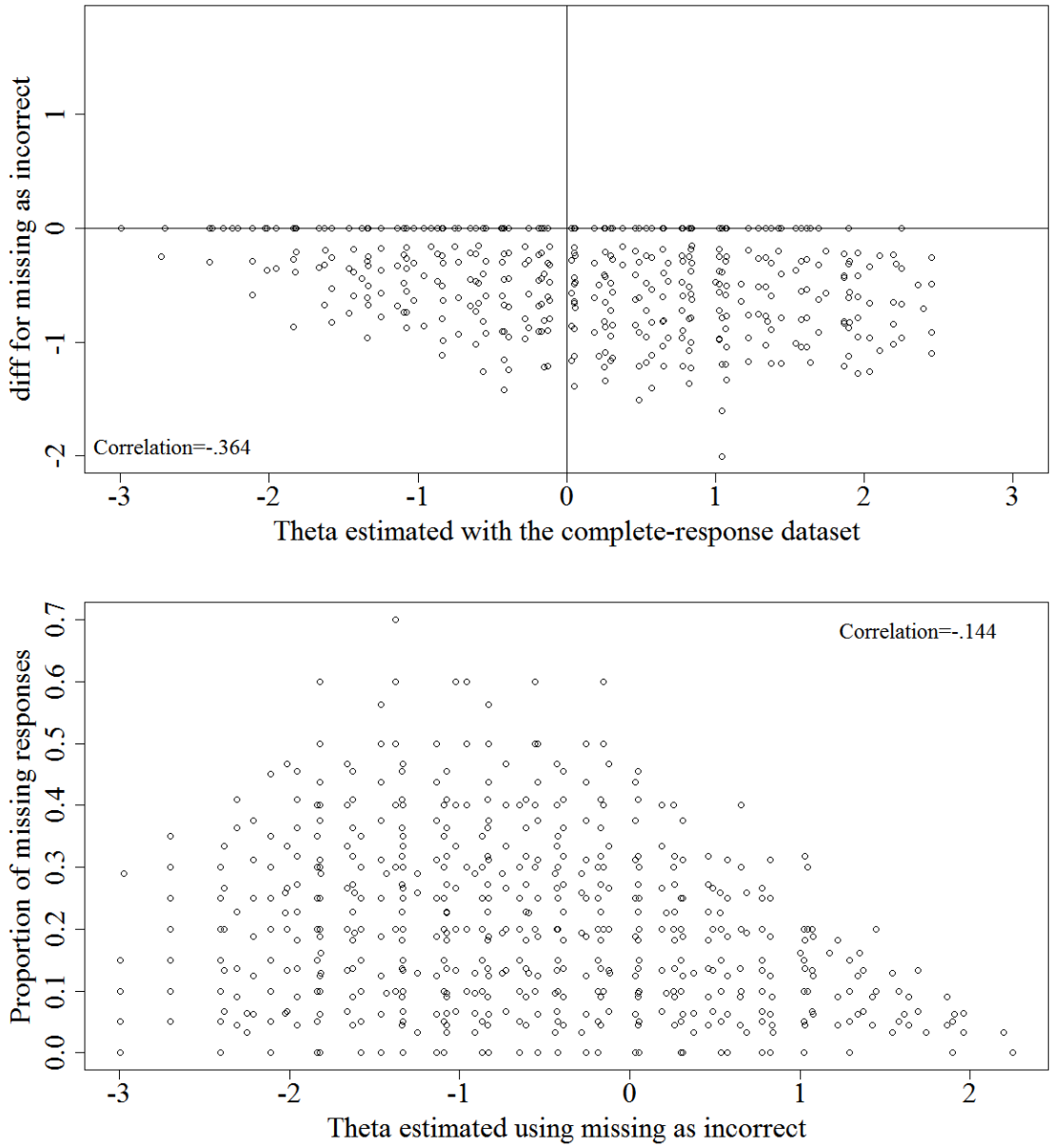

Figure 4. Difference between the theta estimated when missing is treated as incorrect and the theta estimated with the complete-response dataset (top), and correlation of ability estimated using missing as incorrect approach and the proportion of missingness per examinee (bottom), Rasch model.

The signed difference when missing are imputed with midpoint ranged between -.788 and 1.279, which means that this approach both underestimated and overestimated the ability estimators. Most of the times (78%) the thetas were overestimated, and only 8% of the cases had no difference with the thetas estimated from the complete-response dataset. The ability was on average overestimated by .201 (*SD*=.235). This difference was inversely related to the theta values from the complete-response dataset (*r*=-.423, *p*=.000), which means that there was overestimation at low levels of ability and underestimation at high levels of ability (top of Figure 5).

Also, the overestimation was higher for those cases that had incorrect responses in the complete-response dataset than for those cases that had correct responses in the same dataset (*M*=.293 vs *M*=.014). In other words, midpoint imputation did the opposite that imputed as incorrect did. As above, the accuracy in the estimation of ability using this approach was also inversely related to the number of missing responses the student had (*r*=.651, *p*=.000). This correlation was stronger than when missing were treated as incorrect. Additionally, the correlation between the estimated theta using this approach and the level of missingness increased and became positive (*r*=.123, *p*=.000) (bottom of Figure 5).

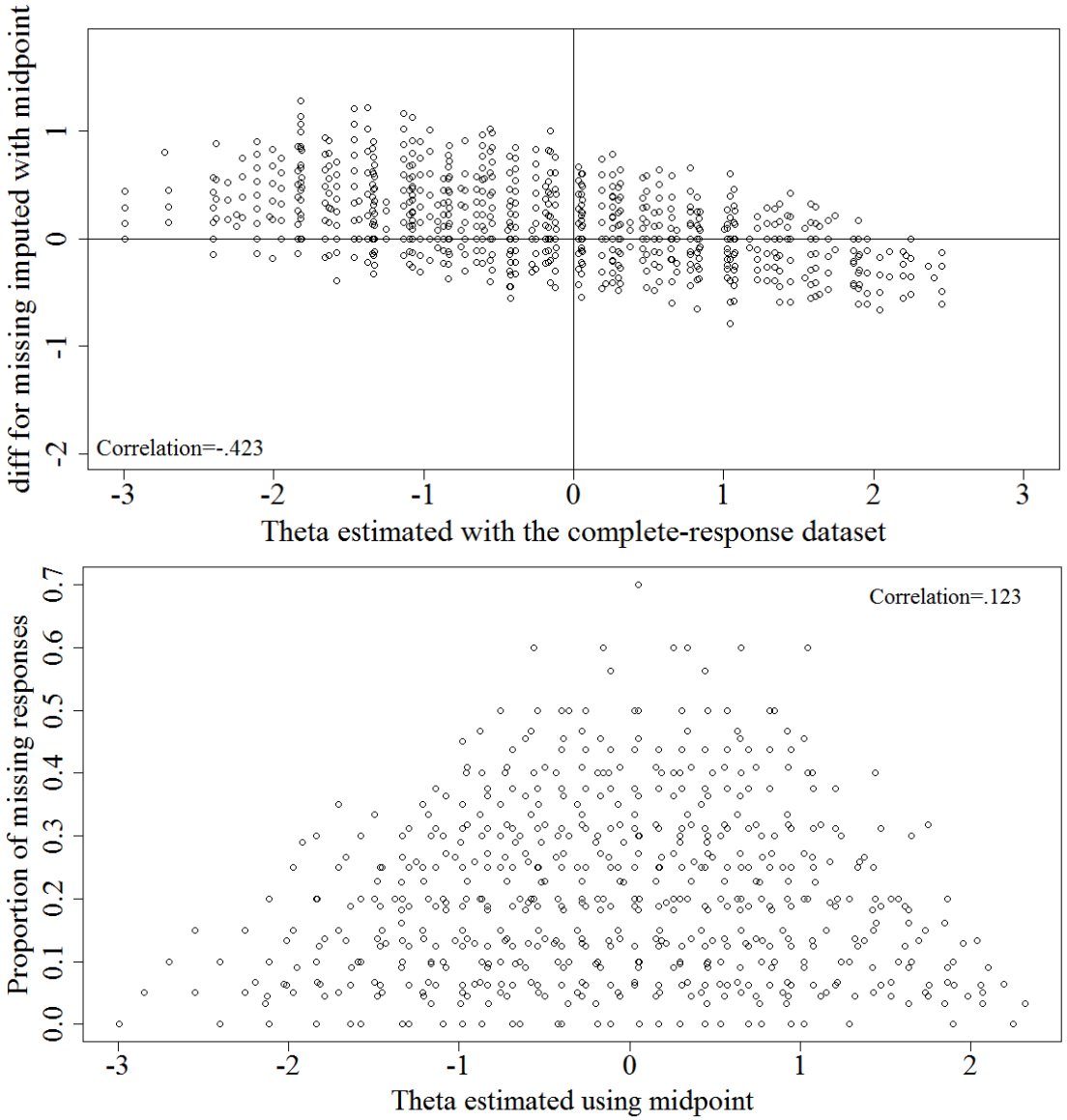

Figure 5. Difference between the theta estimated when missing is imputed with midpoint and the theta estimated with the complete-response dataset and correlation of ability estimated using midpoint approach and the proportion of missingness per examinee (bottom), Rasch model (top).

The signed difference when missing responses were imputed using multiple imputation without auxiliary variables ranged between -.588 and 1.328 logits. That is, the differences were larger when overestimation occurred than when underestimation was observed. Moreover, overestimation of thetas was more frequent (80%) than underestimation (13%), and only 7% of the cases had no difference with the thetas estimated from the complete-response dataset.

The ability was on average overestimated by .229 logits (*SD*=.235). This difference was inversely related to the theta values from the complete-response dataset (*r*=-.390, *p*=.000). In other words, the $\hat{\theta}$s were largely overestimated at their low values and modestly underestimated at their high values (Figure 6). The overestimation was higher for those cases that had incorrect responses in the complete-response dataset than for those cases that had correct responses in the same dataset (*M*=.310 vs *M*=.063).

Again, the accuracy in the estimation of ability using this approach was directly related to the number of missing responses the student had (*r*=.684, *p*=.000). This correlation is stronger than when missing responses were treated as incorrect. Additionally, the correlation between the missingness level and the theta estimated using this approach was also significantly positive (*r*=.144, *p*=.000) (Figure 7).

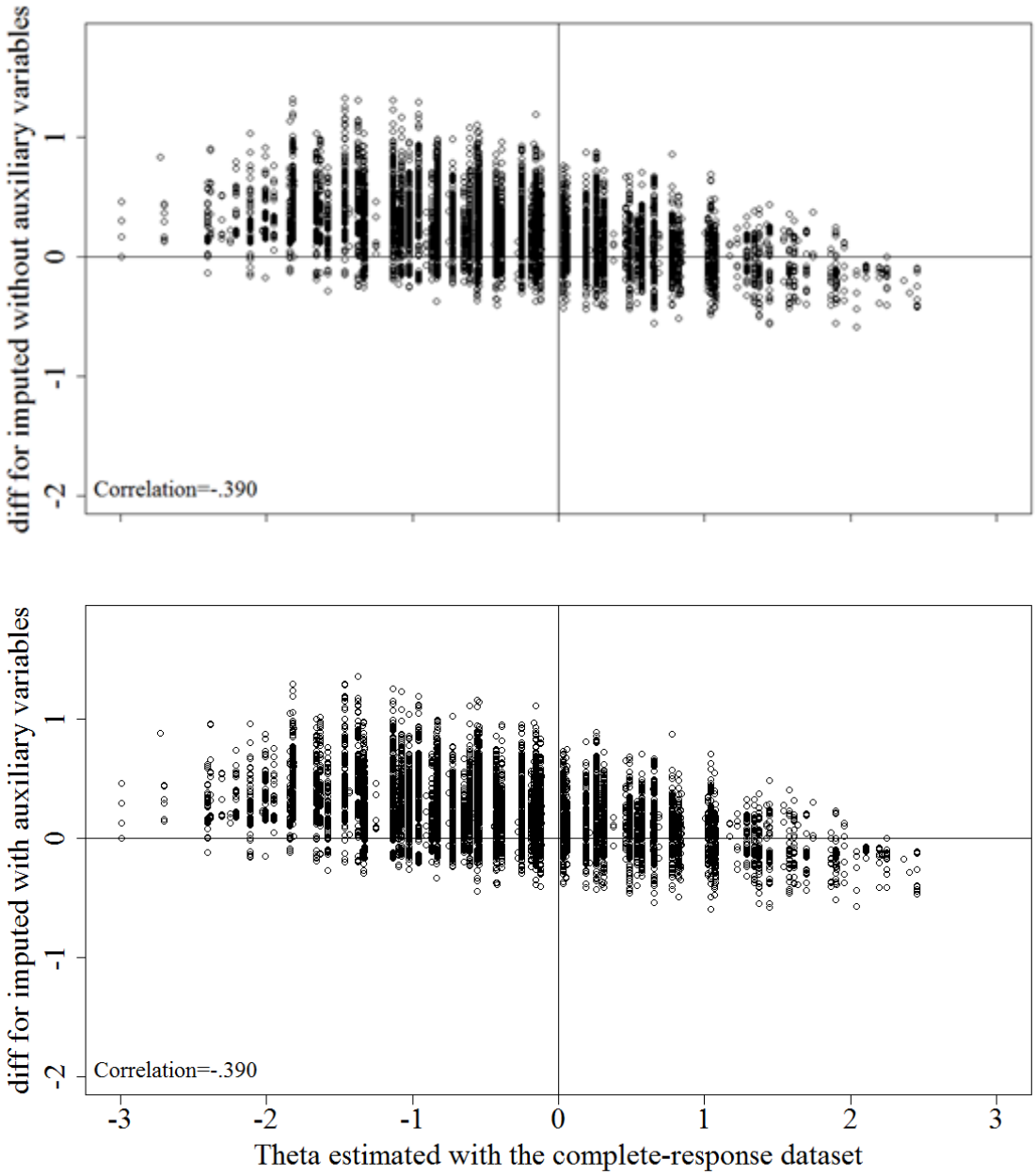

Figure 6. Difference between the theta estimated using multiple imputation without (top) and with (bottom) auxiliary variables and the theta estimated with the complete-response dataset, Rasch model.

When the imputation used auxiliary variables, the signed difference range did not change much from when there were no auxiliary variables in the imputation process. It actually increased slightly from -.596 to 1.354 compared to the previous approach. That is, differences were larger when overestimation occurred rather than when underestimation happened. The proportion of cases with overestimation and underestimation were the same as when no auxiliary variables were used. The ability was overestimated by .228 logits (*SD*=.235) on average.

The relationship between the signed differences and the ability estimated from the complete-response dataset was the same as the previous approach (*r*=-.390, *p*=.000). Again, the $\hat{\theta}$s were largely overestimated at their low values and modestly underestimated at their high values. The overestimation was higher for those cases that had incorrect responses in the complete-response dataset than for those cases that had correct responses in the same dataset (*M*=.309 vs *M*=.062). The correlation between the absolute difference between the $\hat{\theta}$s estimated with the complete-response data and when missing values were treated using this approach, was large as well (*r*=.685, *p*<.000). As with the previous approach, the correlation between missingness and the thetas estimated with this approach was positive (*r*=.144, *p*=.000) (Figure 7).

RMSD. The *RMSD* (calculated across the continuum) was the lowest when the missing values were treated as incorrect (*RMSD*=.254). This index was larger for thetas with positive estimates than for negative estimates (.362 vs .195). The second lowest *RMSD* value was observed when the missing responses were imputed with midpoint (*RMSD*=.309).

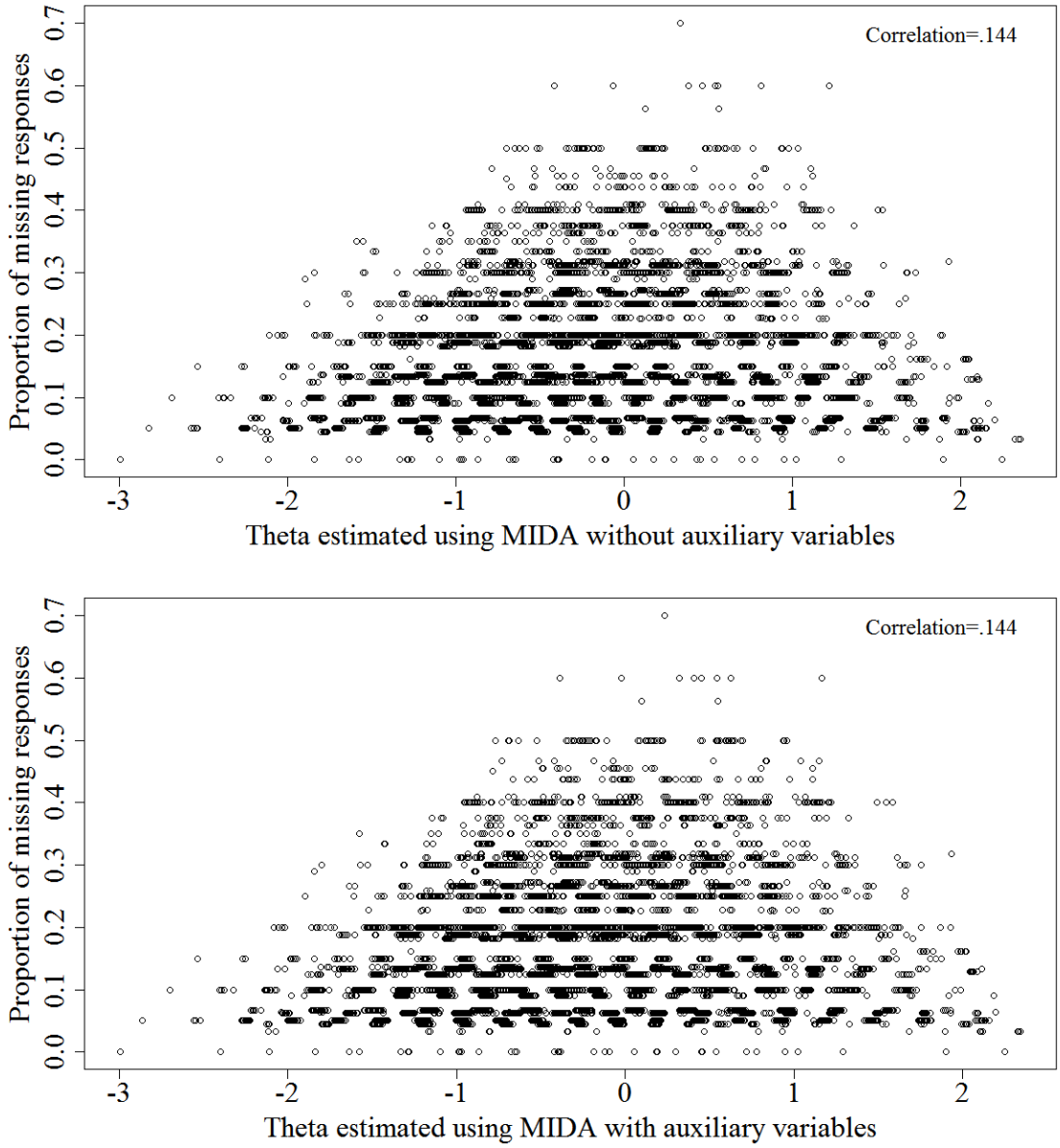

Figure 7. Correlation of proportion of missingness per examinee and ability estimated using MIDA without auxiliary variables (top) and with auxiliary variables (bottom), Rasch model.

This time, the *RMSD* was larger for ability estimates below zero than for positive $\hat{\theta}$s (.337 vs .225). The *RMSD* was equally large when multiple imputation was employed

using auxiliary variables (*RMSD*=.329) than when they were not utilized (*RMSD*=.330). The *RMSD* for these two approaches was also larger at the lower level of the ability estimates. For all the approaches, the *RMSD* was larger for those cases that had incorrect answers in the complete-response dataset.

Average standard error. The average standard error (*SE*) of the estimated thetas was very similar for all four conditions. When compared with the average SE from the complete-response dataset, the missing as incorrect condition yielded a slightly larger average *SE* (*M*=.559, *SD*=.056) than the other approaches. Moreover, in the cases where the thetas were estimated without error, the estimated *SEs* were also exactly the same as the observed when complete-response dataset was used. On the other hand, the average *SE* when missing values were imputed with midpoint (*M*=.552, *SD*=.056) or with multiple imputation with auxiliary variables (*M*=.552, *SD*=.056) or without them (*M*=.552, *SD*=.056) were smaller than the SE from the complete-response dataset.

The correlation between the SEs estimated when missing responses were treated as incorrect and the *SE*s obtained with complete-response dataset is slightly larger than the relationship of the other conditions and the latter (Figure 8). Additionally, there is a significant correlation (*r*=.378, *p*=.000) between the level of non-response per student and their theta *SE*. That is, the *SE* of the estimated theta is larger when the examinees have a larger number of non-responded items. This correlation was almost the same when the missing values were imputed with midpoint (*r*=.309, *p*=.000), and when they were imputed multiple times with or without covariates (*r*=.306, *p*=.000).

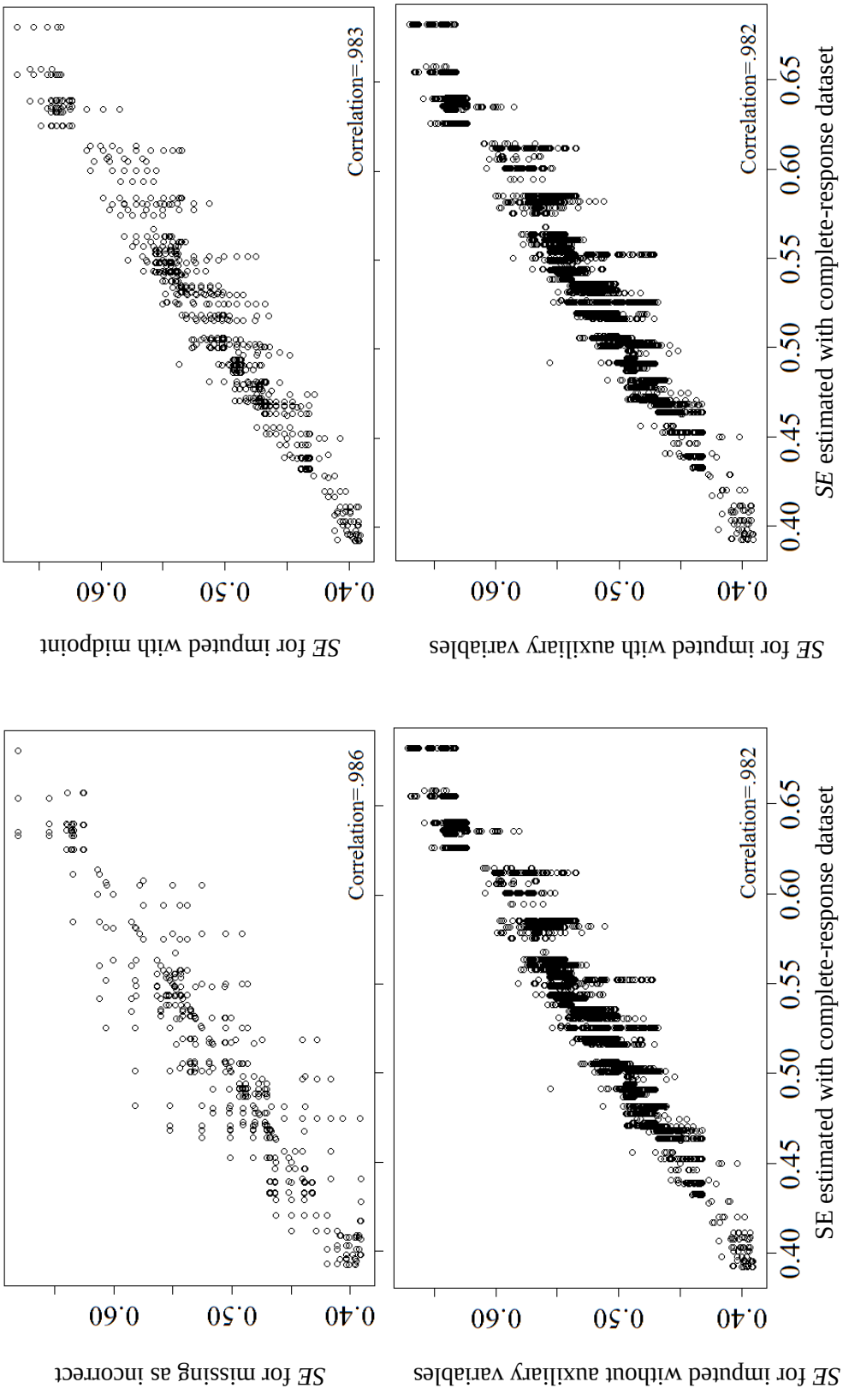

Figure 8. *SE* of estimated thetas under different conditions and SE of estimated theta using complete-response dataset, Rasch model.

Table 2.

Indices and coefficients estimated for comparison of missingness approaches using Rasch model

	Complete-response	Missing as incorrect	Imputed with midpoint	MI with auxiliary variables	MI w/o auxiliary variables
Proportion of coverage		.996	.999	.997	.996
Average confidence interval length	2.229	2.237	2.209	2.208	2.208
Confidence interval length, standard deviation	.224	.222	.224	.224	.224
Signed difference, mean		-.132	.201	.229	.228
Signed difference, standard deviation		.217	.235	.238	.238
Proportion of thetas correctly estimated		.669	.082	.073	.072
Proportion of thetas overestimated		.000	.778	.798	.798
Proportion of thetas underestimated		.331	.140	.129	.130
RMSD		.254	.309	.330	.329
Average standard error of estimated theta	.557	.559	.552	.552	.552
Standard error of estimated theta, standard deviation	.056	.056	.056	.056	.056
Theta, mean	-.444	-.575	-.243	-.215	-.216
Theta, standard deviation	.821	.769	.752	.762	.762
Correlation between confidence interval	1.000	.986	.983	.982	.982
Correlation between theta's absolute differences and number of missing responses		.376	.651	.684	.685
Correlation between estimated thetas	1.000	.965	.959	.958	.958
Correlation between *SE*	1.000	.986	.983	.982	.982
Correlation between *SE* and number of missing responses		297	.215	.211	.211

2PL IRT model

The students' ability estimated using 2PL model on the complete-response dataset ranged between -2.559 and 3.023, with 55% percent of the students having a negative estimated theta. The mean estimated theta was -.0019 logits (*SD*=.860). The average standard error of the estimated theta was .519 logits (*SD*=.078) and the average length of the confidence interval was 2.074 logits (*SD*=.312). The theta *SE* was significantly larger, $F(1,17124)=1662$, $p=.000$, $\eta^2=.088$, for the low ability estimates (*M*=.540, *SD*=.085) than for the high-level performers (*M*=.493, *SD*=.059). However, the sample size for the later was smaller than for the former.

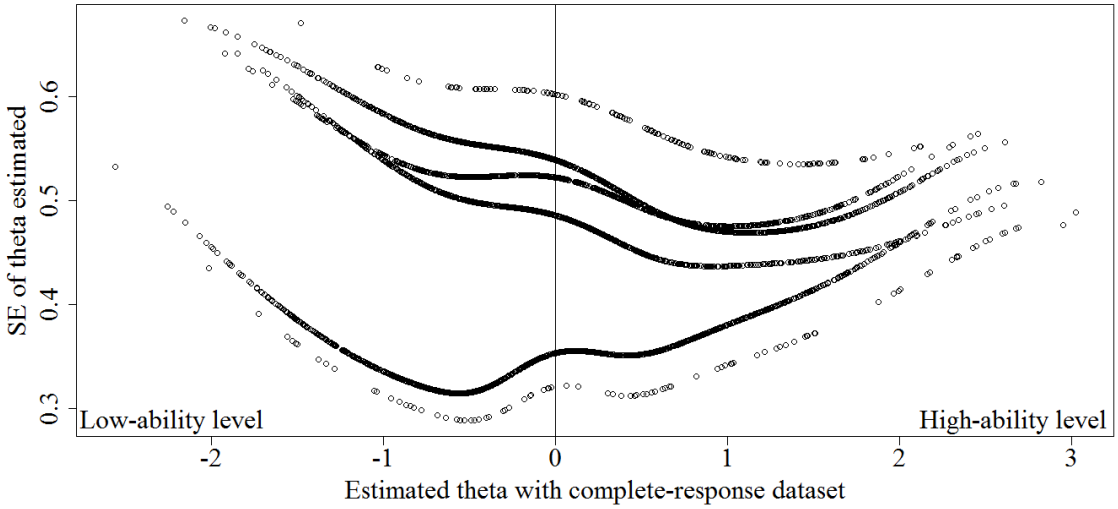

Figure 9. Estimated thetas and their SE, 2PL model using the complete-response dataset

The correlation between the theta estimated from the complete-response dataset and the level of missingness (*M*=.176, *SD*=.120) in the corresponding incomplete-response dataset was significant (Figure 10), but relatively low ($r=-.081$, $p=.000$). Additionally, the correlation of both auxiliary variables with the $\hat{\theta}$s estimated from the

complete-response dataset was significant. For the ISEC index, the correlation was .290 (p=.000) and for the ICEH it was .260 (p=.000).

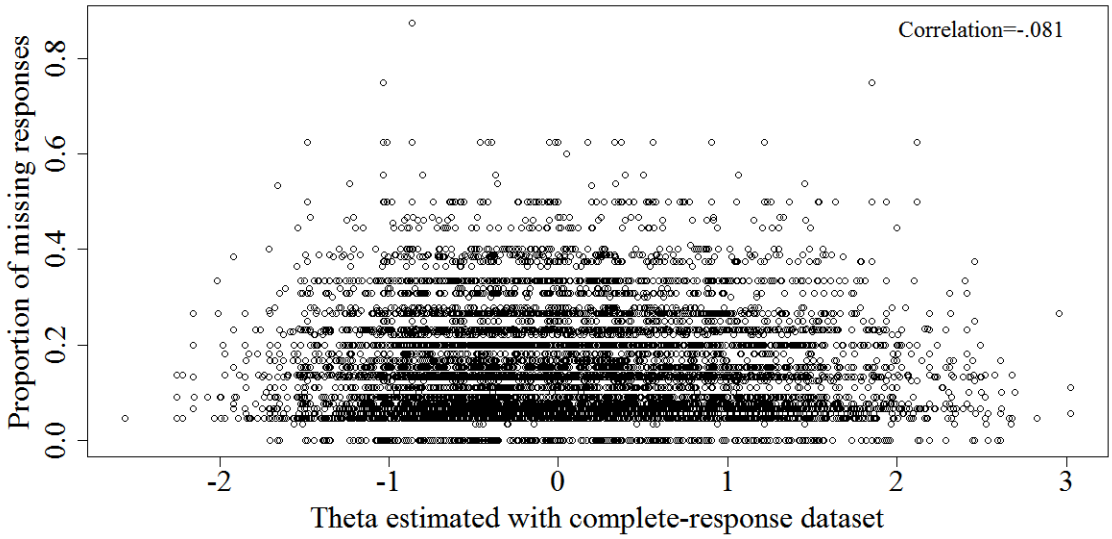

Figure 10. Correlation between ability estimated using the complete-response dataset and the proportion of missingness per examinee, 2PL model.

Between and within imputation variability. There were five datasets imputed with auxiliary variables, and another 5 datasets imputed without auxiliary variables. The between and within variance was estimated across each of these two sets of datasets for each of the criterion to be used in the comparison of missing data handling approaches. Overall, there was almost no variability in the indices estimated between datasets. Most of the total variance is within the data set (more than 99%). Therefore, working with the mean of the imputed files will facilitate the comparison among conditions without compromising the results and the conclusions. That is, for each of the following indices, the mean across all five imputed datasets without auxiliary variables and the mean of the

five imputed datasets with auxiliary variables are reported.

Coverage. As was the case with the Rasch model, the coverage index was high (more than .950) for all the missing data handling approaches. That is, most of the estimated thetas from the complete-response dataset were within the 95% confidence interval formed by the estimated thetas from the datasets with missing values treated by the different missingness approaches.

Average length of confidence interval. The average lengths of the confidence interval (CI) for the estimated theta were the same when the missing responses were imputed using multiple imputation, with (*M*=2.016, *SD*=.281), without auxiliary variables (*M*=2.016, *SD*=.281), and the midpoint imputation (*M*=2.018, *SD*=.283). These three approaches had very similar performance in terms of the CI. Compared with the CI estimated using the complete-response dataset, these showed smaller average CI length. Their correlation with the CI of the complete-response dataset is high (Figure 11).

When the missing values were treated as incorrect, the average length slightly increased (*M*=2.089, *SD*=.321) in comparison to the other three approaches. This change represented a 4% increase with respect to the average lengths from the other missing data handling approaches. The average length of the CI was slightly larger than the observed for the complete response dataset. Likewise, the dispersion was larger for this approach; this impacted the correlation with the CI of the complete-response dataset.

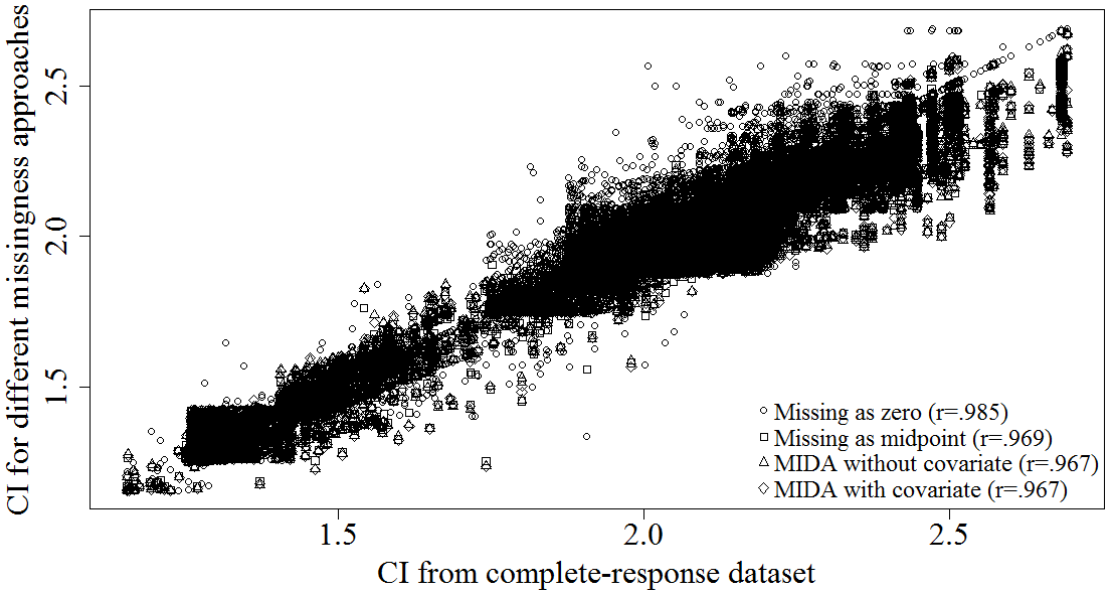

Figure 11. Confidence interval from the complete response dataset versus the CI estimated under the different missingness handling approaches, 2PL model.

Signed difference. The difference between the theta estimated with the complete-response dataset and when missing responses were treated as incorrect ranged between -2.424 and 0.000 logits. Around 68% of the thetas were accurately estimated (i.e., the difference was zero) and the rest (31%) were underestimated. No overestimation occurred (top of Figure 12). On average, the thetas estimated using this approach were underestimated by .123 (*SD*=.245). The differences were inversely related to the theta values from the complete-response dataset (*r*=-.384, *p*=.000), which means that the ability was more overestimated at lower levels of theta and less overestimated at high levels of ability.

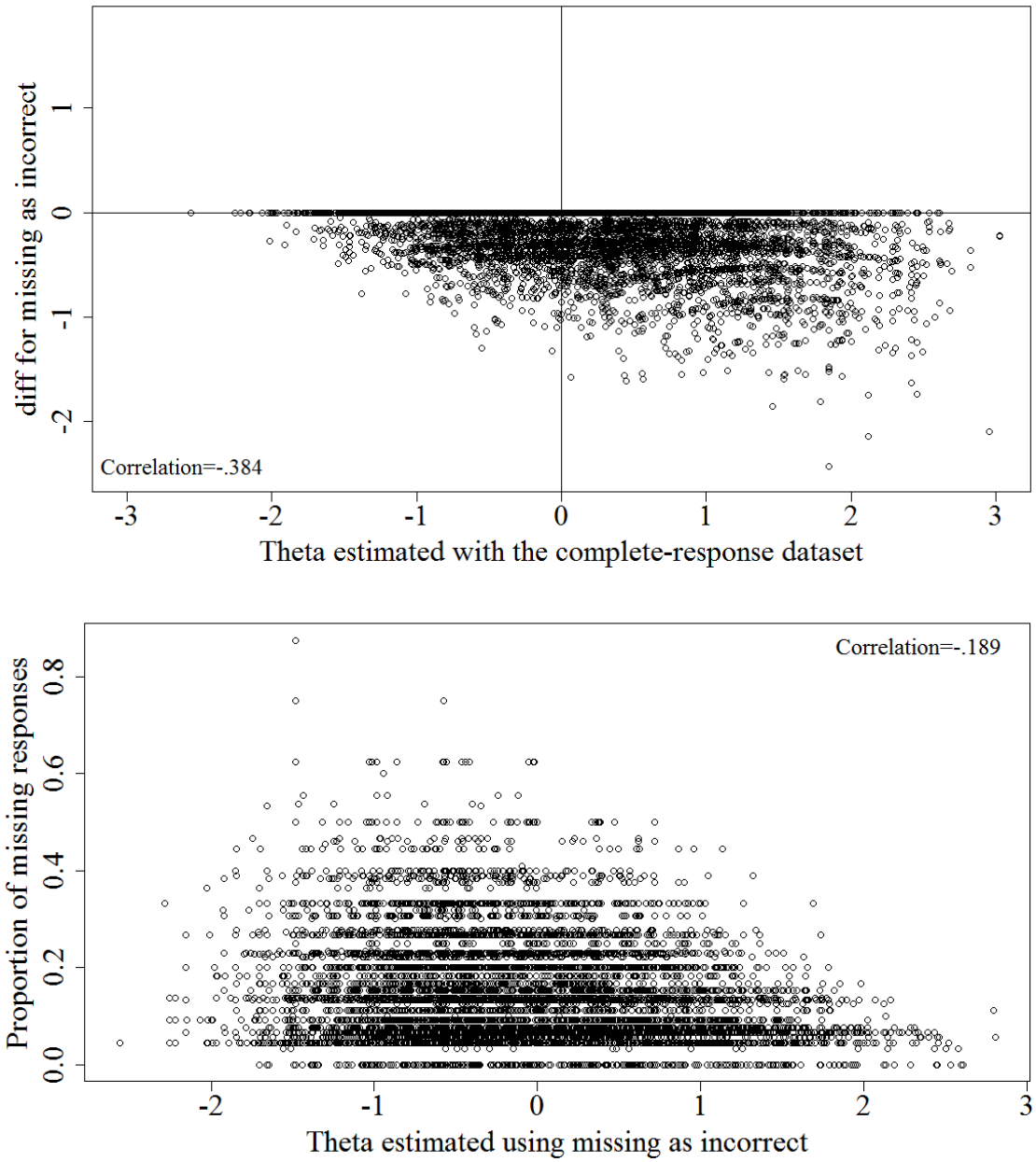

Figure 12. Difference between the theta estimated when missing is treated as incorrect and the theta estimated with the complete-response dataset (top), and correlation of ability estimated using missing as incorrect approach and the proportion of missingness per examinee (bottom), 2PL model.

In fact, the correlation between the theta from the complete-response dataset and the estimated with this approach was lower for $\hat{\theta}<0$ (r=.938, p=.000) than for $\hat{\theta}>0$ (r=.850, p=.000). On the other hand, the correlation of the absolute difference between the $\hat{\theta}$s estimated with the complete-response data and when missing values were treated as incorrect, and the missingness level was positive. In other words, there was more error when the missingness level was higher (r=.332, p=.000). The correlation between the ability level and the missingness level increased from -.081 to -.189 (p=.000) when this approach was used (bottom of Figure 12).

The signed difference when missing values were imputed with midpoint ranged between -1.225 and 1.627. Most of the times (81%) the thetas were overestimated, and only 3% of the cases had no difference with the thetas estimated from the complete-response dataset. The ability was on average overestimated by .297 (SD=.340). Like in the previous case, the difference was inversely related to the theta values from the complete-response dataset (r=-.483, p=.000). This means that the theta was more overestimated at lower levels of ability than it was at higher levels of ability (top of Figure 13). This is confirmed by the low correlation between the theta from the complete-response dataset and the estimated with this approach for $\hat{\theta}<0$ (r=.773, p=.000) compared with $\hat{\theta}>0$ (r=.861, p=.000).

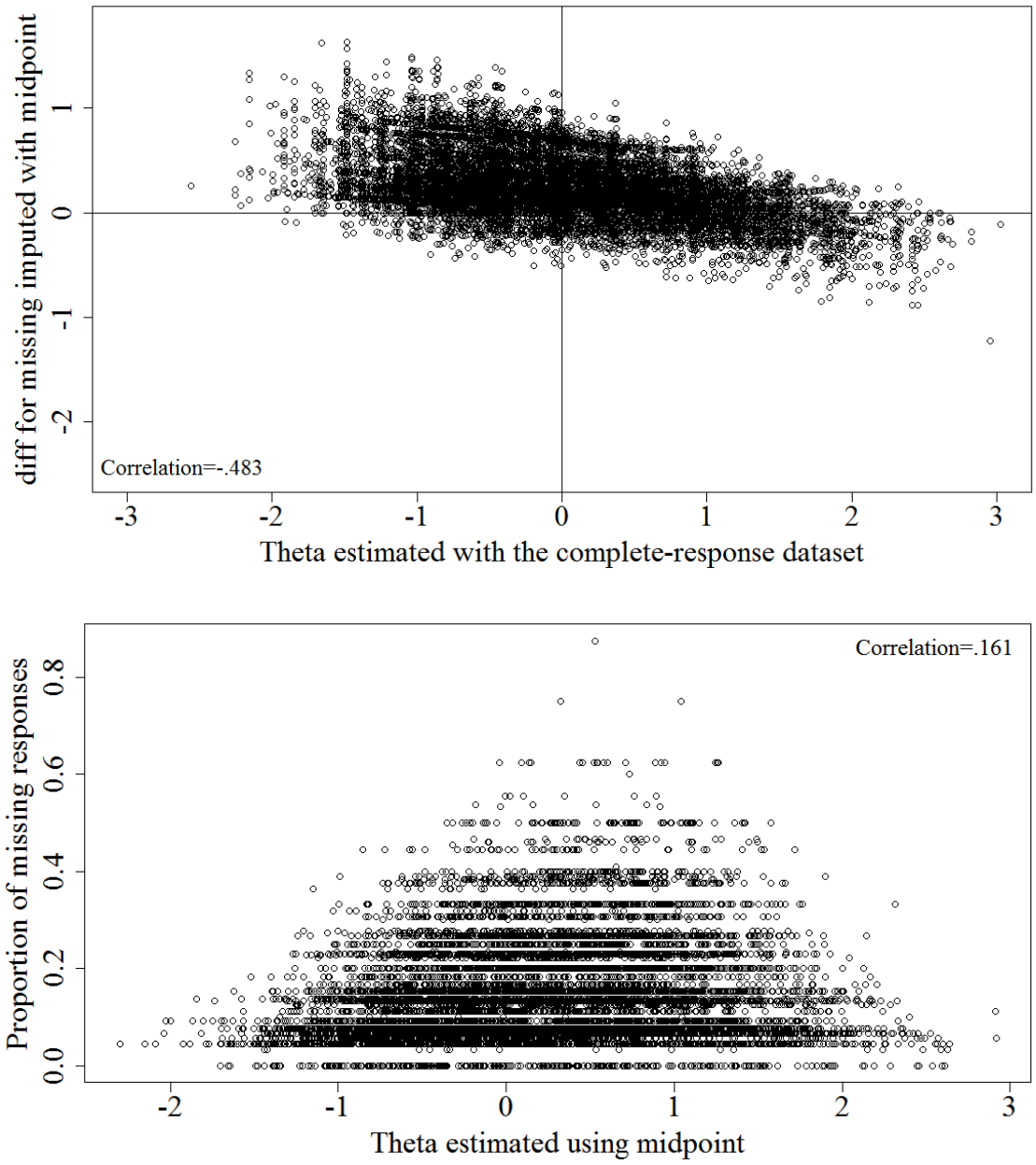

Figure 13. Difference between the theta estimated when missing was imputed with midpoint and the theta estimated with the complete-response dataset (top), and correlation of ability estimated using midpoint approach and the proportion of missingness per examinee (bottom), 2PL model.

On the other hand, the correlation of the absolute difference between the $\hat{\theta}$s estimated with the complete-response data and when missing values were imputed with .5, and the missingness level was also positive (r=.676, p=.000). That is, the differences were larger for those examinees that had a higher number of missing responses. The correlation between the ability level and the missingness level increased and became positive (r=.161, p=.000) when this approach was used (bottom of Figure 13).

The signed difference when missing values were imputed using multiple imputation without auxiliary variables ranged between -.864 and 1.819. Overestimation of thetas was more frequent (80%) and with larger difference values than when underestimation occurred (14%). The other 6% of the cases showed zero difference. The rest of the cases had no difference with the thetas estimated from the complete-response dataset. The ability was, on average, overestimated by .325 logits (SD=.343). The difference was inversely related to the theta values from the complete-response dataset (r=-.459, p=.000) as can be seen in Figure 14, graph on the top.

As the previous approach, the correlation between the theta estimates (from complete-response dataset and the imputed with this technique) was lower for the low ability examinees ($\hat{\theta}$<0) than for the high ones ($\hat{\theta}$>0). Also, there was a .692 correlation (p=.000) between the absolute differences in the theta estimated and the level of missingness per participant. Finally, the correlation between the missingness level and the theta estimated using this approach was significantly positive (r=.179, p=.000) (top of Figure 15).

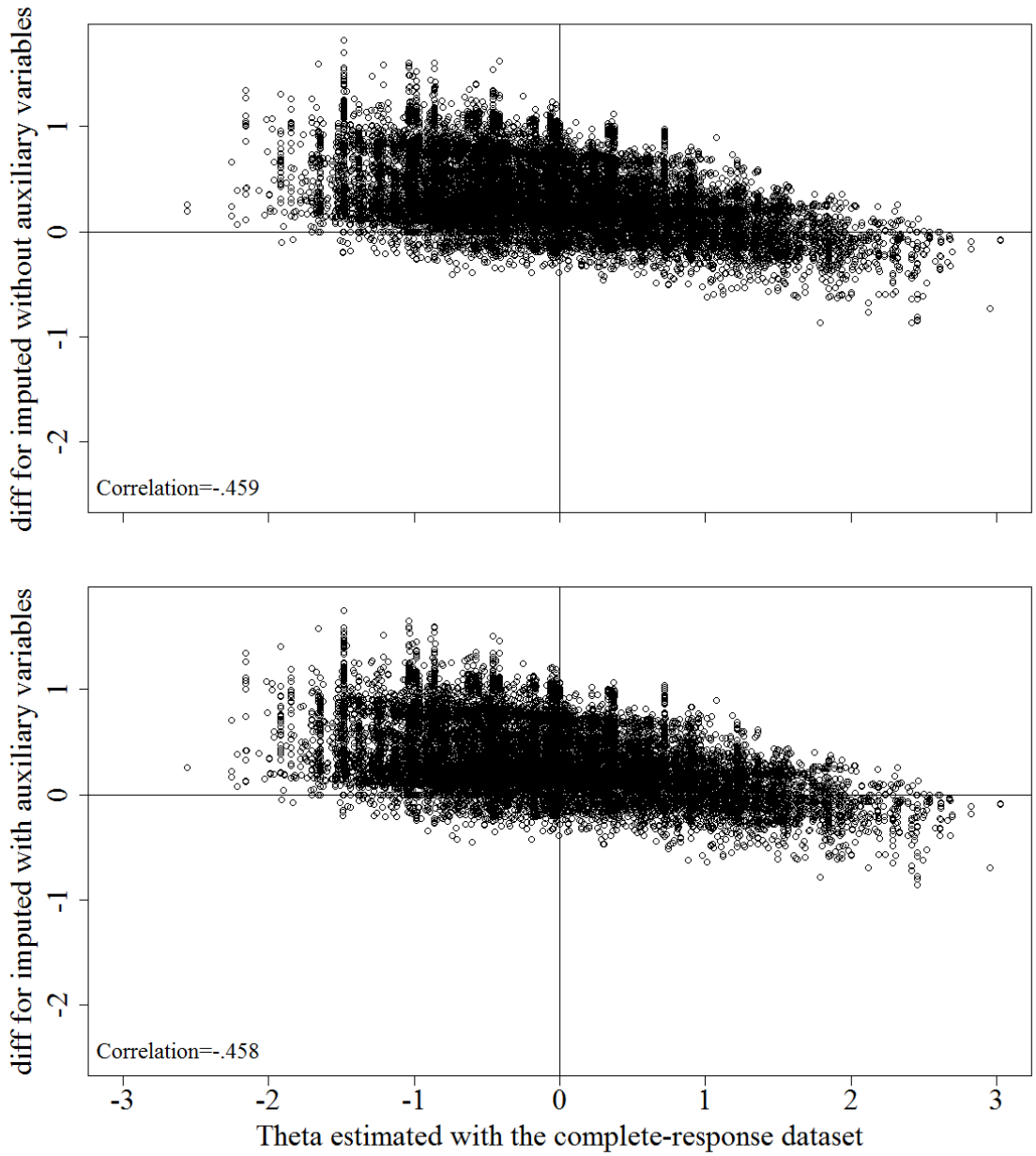

Figure 14. Difference between the theta estimated using multiple imputation without (top) and with (bottom) auxiliary variables and the theta estimated with the complete-response dataset, 2PL model.

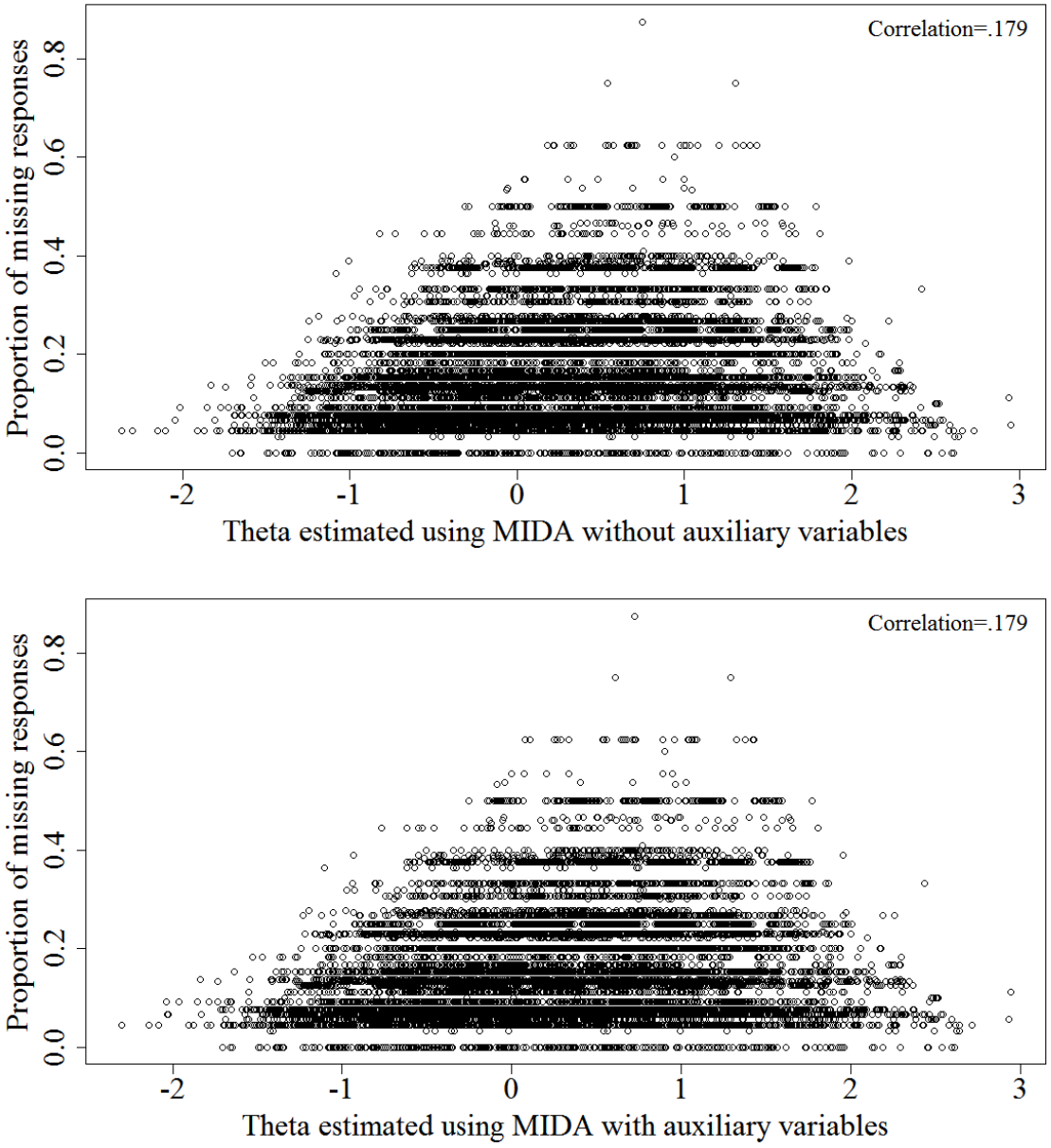

Figure 15. Correlation of proportion of missingness per examinee and ability estimated using MIDA without auxiliary variables (top) and with auxiliary variables (bottom), 2PL model.

Almost similar values were found when the imputation used auxiliary variables. The range of the signed difference was between -.859 and 1.748 logits. The proportion of cases with overestimation was 80%. Fourteen percent of the thetas were underestimated and the rest showed no difference at all. The differences between the benchmark values and the estimated thetas were larger when overestimation occurred than when underestimation was observed. Ability was overestimated by .324 logits (*SD*=.344), on average. The correlation between the differences and the ability estimated from the complete-response dataset was the same as with the previous approach (*r*=-.458, *p*=.000) (bottom of Figure 14) as it was the low correlation between examinees with low ability level. As before, larger levels of missingness per participant were associated with larger differences in the theta estimated (*r*=.694, *p*=.000), and the correlation between the missingness level and the theta estimated using this approach was also positive (*r*=.179, *p*=.000) as observed in Figure 15 (bottom).

RMSD. The *RMSD* was at its lowest when missing responses were treated as incorrect (*RMSD*=.279). This is because most of the thetas were accurately estimated, as mentioned before. This coefficient was larger for examinees with high ability estimate ($\hat{\theta}$>0) than for those with low ability estimation (.368 vs. .173). When the missing responses were imputed with midpoint, the *RMSD*=.452. Unlike the previous approach, the *RMSD* was larger for lower ability examinees and shorter for the ones that have higher theta values (.537 vs .319). The *RMSD* was equally large when multiple imputation without auxiliary variables (*RMSD*=.472) were employed than when these variables were used (*RMSD*=.473). The same pattern of missing imputed with midpoint

was observed with these two approaches (i.e., larger *RMSD* values for when the $\hat{\theta}$s were below zero). In other words, the error in the estimation was larger for low ability level when midpoint or MIDA (with and without auxiliary variables) were used, whereas the error was larger for high ability estimates when missing was treated as incorrect.

Average standard error. The average *SE* of the estimated thetas was the same when missing responses were imputed with midpoint (*M*=.504, *SD*=.071), as when they were imputed with MIDA without auxiliary variables *SE* (*M*=.504, *SD*=.070), and as when auxiliary variables were used for the multiple imputation (*M*=.504, *SD*=.070). When missing responses were treated as incorrect, the *SE* increased by 3.6% (*M*=.522, *SD*=.080). Compared with the average *SE* from the complete-response dataset, the missing as incorrect condition was closer to the average *SE* than the other three conditions. Moreover, for cases where the θs were estimated without error, the *SE* was also exactly the same as the observed when complete-response dataset was used.

The correlation between the *SE* from the complete-response data set and the one obtained under the different missing data approaches was relatively high (Figure 16). There was more disagreement or dispersion when the *SE* from the complete-response data set was large. Additionally, there is a significant correlation (*r*=.353, *p*=.000) between the level of non-response per student and their theta *SE* when missing was treated as incorrect. In other words, the SE of the estimated theta was larger when the examinees had larger number of non-responded items. This correlation was smaller when the missing values were imputed with midpoint (*r*=.219, *p*=.000), and when they were imputed multiple times with or without covariates (*r*=.212, *p*=.000).

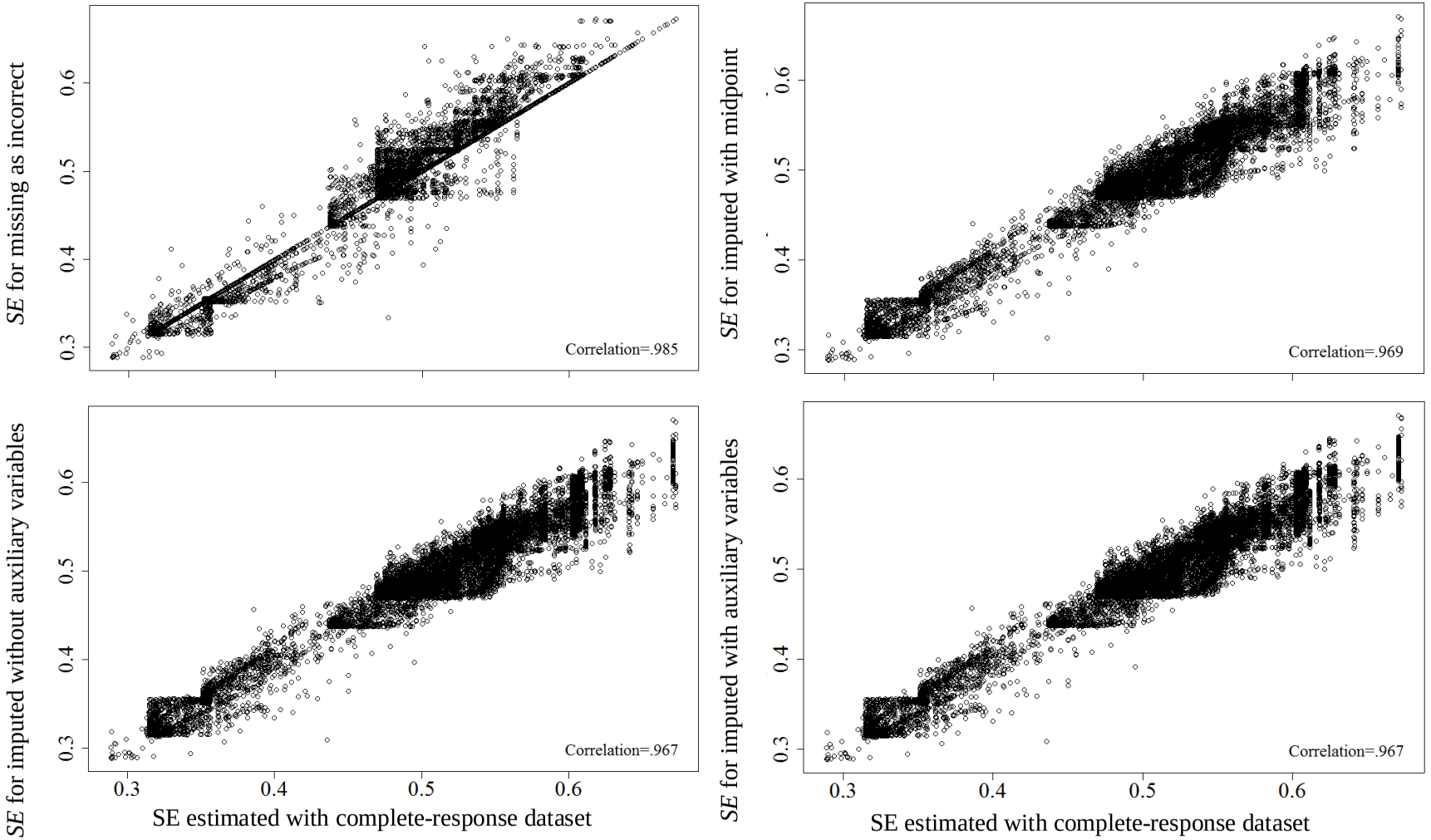

Figure 16. *SE* of estimated thetas under different conditions and SE of estimated theta using complete-response dataset, 2PL model.

Table 3.

Indices and coefficients estimated for comparison of missingness approaches using 2PL model

	Complete-response	Missing as incorrect	Imputed with midpoint	MI with auxiliary variables	MI w/o auxiliary variables
Proportion of coverage		.988	.991	.955	.956
Average confidence interval length	2.074	2.089	2.018	2.016	2.016
Confidence interval length, standard deviation	.312	.321	.283	.281	.281
Signed difference, mean		-.132	.297	.325	.324
Signed difference, standard deviation		.245	.340	.343	.344
Proportion of thetas correctly estimated		.686	.030	.062	.061
Proportion of thetas overestimated		.000	.807	.802	.803
Proportion of thetas underestimated		.314	.163	.136	.136
RMSD		.279	.452	.472	.473
Average standard error of estimated theta	.519	.522	.504	.504	.504
Standard error of estimated theta, standard deviation	. 078	.080	.071	.070	.070
Theta, mean	-.002	-.134	.295	.323	.322
Theta, standard deviation	.860	.799	.757	.767	.768
Correlation between confidence interval	1.000	.985	.969	.967	.967
Correlation between theta's absolute differences and number of missing responses		.332	.677	.692	.694
Correlation between estimated thetas	1.000	.959	.919	.917	.917
Correlation between *SE*	1.000	.985	.969	.967	.967
Correlation between *SE* and number of missing responses		.328	.190	.183	.183

CHAPTER V: DISCUSSION

This research aimed at exploring the performance of different approaches for handling missing data during theta estimation using large scale assessment data. Specifically, this study focused on the effectiveness of missing data approaches when the Rasch or two-parameter IRT models were used. The data used in this study come from SERCE, a large scale assessment. Item parameters for the Rasch and 2PL models were estimated using complete-response dataset and over which a missingness pattern was imposed as observed in the original dataset. Working with empirical data overcomes limitations observed with simulated data and sample size. All the missing responses were assumed to be omitted (i.e., not-answered responses towards the end of the test were not considered not-reached or ignorable). The results showed that the approaches do not differ much from each other but there are still some differences that are summarized in this chapter.

Multiple imputation data augmentation with and without auxiliary variables was used to generate decimal responses during the imputation process. These non-binary decimal numbers were used to estimate the theta values. In both the Rasch and 2PL models, the MIDA without auxiliary variables performed the same as when auxiliary variables were used. They returned similar theta estimates and the level of difference between the estimates from the complete-response dataset was the same. Overall, they equally overestimated examinees' ability with both the Rasch and 2PL models and showed the same average standard error of the estimate. Working with auxiliary variables when estimating thetas does not seem to improve the accuracy of the estimation.

One of the reasons for the similar performance could be low the level of correlation between the auxiliary variables (socioeconomic status and household cultural and educational condition) and the outcome variable (less than .30); the outcome variable was the number correct in the complete-response data set. Likewise, the correlation between the auxiliary variables and the estimated thetas was below .30 for both the Rasch and the 2PL models.

The literature (Collins et al., 2001; Enders, 2010) indicates that auxiliary variables are useful when the correlation with the missing analysis variable is larger than .40. Enders (2010) found that bias still exists when auxiliary variables are used, but they tend to decrease it, especially when the correlation with the missing analysis variable is larger than .50. Also, omitting auxiliary variables with correlation lower than .40 has a minimal impact on reducing the bias, especially when the missingness level is less than 25% (Collins et al., 2001).

Although they evaluated different parameters, this study showed that their conclusions hold for ability estimation as well. Therefore, an exploratory correlational analysis should be conducted in order to decide whether to include auxiliary variables when imputing ability estimates. Moreover, additional correlational analysis between the potential auxiliary variable and a binary missing indicator of the variable of interest should also be obtained. A high correlation is normally considered as evidence of MAR and thus should be included in the imputation process (Collins et al., 2001).

Imputing missing responses with a midpoint (i.e., .5) yielded also similar results as multiple imputation with data augmentation. The correlation between the estimated

theta using this approach and the estimated theta from the complete-response dataset was the same as those observed with MIDA for both IRT models. Likewise, this approach showed the same level of correlation between the missing level and the error in the estimation as with MIDA approaches. In other words, there was more error in the estimation of theta when there was higher level of missingness. This was true despite the fact that there was a low correlation between the theta estimated from the complete-response dataset and the level of missingness for both IRT models.

Imputing missing responses with midpoint, too, mainly overestimated the ability level but to a lower extent than either MIDA with and without auxiliary variables. The standard error of the estimate was the same as the previous approach. Consequently, it can be said that using the midpoint approach did not imply any loss in the accuracy of the ability estimation. Moreover, this approach has advantages over multiple imputation with data augmentation. The latter requires work with multiple files in order to maintain its stochastic nature. Working with several files to take into account the error in the estimation can be burdensome when this has to be combined with other types of analyses. Also, it may be confusing or tedious for secondary data analysts to deal with multiple files (Ender, 2010). Likewise, chances of error in the data analysis process increases as the number of files does. In contrast, the midpoint imputation allows the analyst to deal with only one dataset without compromising the quality of the estimates.

Unlike the three approaches described above, when the missing values were treated as incorrect the ability was either correctly estimated or underestimated. Moreover, the average error in the estimation was lower than what was observed in the

other approaches. In fact, treating missing responses as incorrect estimated the ability level of examinees without error at least two third of the times. The standard error of the estimate, although slightly higher than the other approaches, was closer to the observed with the complete-response dataset.

However, the error level in the cases where the thetas were underestimated was the highest for both IRT models. In other words, when this approach did not correctly estimate the thetas, it greatly underestimated it, on average. The reason for the good performance of this approach over the others is that most of the cases with missing responses had incorrect as their responses in the complete-data set. Therefore, treating missing as incorrect successfully imputed the expected answer two third of the times. The difference in performance between the imputed as incorrect and the other approaches was worse in the 2PL than in the Rasch model. Differences in the indices used to evaluate the approaches were larger for the former than for the latter.

Rose et al. (2010) found similar results as the ones presented for the missing as incorrect. Person's ability was underestimated when missing was treated as incorrect and the item parameters were estimated ignoring the missing responses (i.e., left in blank). When missing values were treated as incorrect, the average error in the ability estimation was similar to the level they observed when thetas were estimated with the complete-response dataset (compared to the true θ). Likewise, Ludlow and O'Leary (1999) showed that treating missing as incorrect led to better results than ignoring them in both item calibration and person ability estimation. In this study, the item parameters were estimated with the complete-response dataset and kept fixed throughout the study.

However, it is possible that the underestimation in this approach is due to differences in the "basic item statistics, such as the percent correct and the item total correlations between different stages of the analysis" (Rose et al., 2010, p. 4).

Regarding the ability level, both MIDA and midpoint imputation showed higher margin of error at the low level of ability estimates for both IRT models. In other words, the theta were more overestimated when they were below zero. Smaller errors were observed on the high ability level. Treating missing as incorrect yielded the opposite results. That is, the ability was more underestimated at the high level of ability (i.e., when theta estimates were above zero) than when the ability was low. In fact, a higher proportion of thetas estimated without error was observed at the low level of ability. The characteristics of the sample seems to explain this pattern, given the high number of cases with incorrect (i.e., zero) as the original response. Most of these cases (75%) were located at the low end of ability estimates.

The sample size should not represent a problem in this study, given the large number of examinees, although the ability distribution is positively skewed. Also, there appears not to be a relationship between sample size and the missingness approaches. The missingness level, on the other hand, seems to play a role. The correlation between this variable and the errors in the estimation for both MIDA and midpoint imputation was twice as large as it was with missing as incorrect. This high correlation is associated with the fact that the relationship between missingness level and ability gets inflated when any of the missing data handling approaches is used. While this correlation was low (less than -.10) for both IRT models when thetas estimated with the complete-response dataset were

used, it was twice or three times larger when the responses were treated as incorrect (e.g., r=-.144 for the Rasch model) and it became positive when the other approaches were used (e.g., r=.123 for the midpoint imputation in the Rasch model). This is associated to the fact that the former approach underestimated the ability performance while the last ones overestimated it, on average. Also, it is interesting to see that using proportion of number correct as a proxy of examinees' ability to estimate correlation with missingness level could be misleading when evaluating the ignorability of the missing responses (see data generation, step 3).

Overall, it seems that treating missing as incorrect yields a smaller average error in the person ability estimation, especially when the proportion of non-response per person is not so high (e.g., it was less than .20 in this study). In this approach, it is assumed that the examinees would provide a wrong answer to the unanswered items. Therefore, the likelihood of examinees getting the item right, regardless of their ability level, is reduced to zero with this approach. When the user is interested in cluster average such a country level performance, this approach seems to yield an acceptable estimate. Nevertheless, data analysts should be aware of the underestimation they will face and the subsequent inflation in the association between missingness and ability. Considering this, midpoint imputation might be more appropriate and more effective than multiple imputation. In addition, the use of auxiliary variables in the latter approach should not be considered unless there is a high correlation with the observed score.

Limitations

A limitation of this study was that the performance of these approaches did not consider the effect of missingness on item parameter estimation. Nevertheless, this study reported a significant and high correlation between item difficulty and missingness. The goal in this research was to examine the impact of missing data handling procedures in the estimation of the ability level for large samples. In large scale assessment surveys, however, item parameters also need to be estimated. Consequently, the effect of the quality of estimation of these parameters upon the person ability estimation have to be explored. As other authors have pointed out, there is a carry forward effect of the item estimation on the theta estimates (e.g., Ludlow & O'Leary, 1999; Oshima, 1994; Rose et al., 2010).

Another limitation of this study is the not control over the missingness mechanism. The missingness mechanism was not a condition in this study because the missingness pattern used in this study was taken from SERCE data. The low correlation between the missingness level and the estimated theta using the complete-response dataset may imply that the missingness mechanism more resembled either MCAR or MAR, but not MNAR. Further research is needed to compare the performance of these approaches when the missingness follow a MNAR pattern more closely. Moreover, it is possible that the missing mechanism differs from country to country in a large scale assessment. Rose et al. (2010) say that the comparison among "groups of respondents might be unfair if they differ in their amount of missing data and in the strength of the

relationship between the latent variable and the missing data" (p. 17). Consequently, this correlation should be estimated by country to rule out this issue.

A third limitation is that the performance of the missingness approaches was evaluated using only dichotomous items. It is most likely that performance of these approaches may differ when polytomous items are used, especially when the test contains a large proportion of these items. For example, it is possible that the missing as incorrect approach would underestimate ability more in these circumstances, given that there would be other potential answer options for the low ability examinees besides zero. Also, the IRT models were unidimensional which is how SERCE was designed. However, it is possible that the missingness approaches differ if multidimensional assumption is held, as is the case with PISA data.

Another limitation, although of the study but of M*plus*, is the fact that the program does not allow to distinguish between the planned and unplanned missing data. Therefore, imputation cannot be conducted on the whole response matrix. Instead, the matrix has to be split (in this case in blocks) in order to get imputed values only for the unplanned missing responses. In other words, multiple imputation was conducted on blocks rather than booklets to prevent the software from imputing unplanned missing data. As a consequence not all the observed answers of the examinees were considered at once when imputing their missing responses. Although a variable indicating the block order was included in the imputation process, it is likely that information about the student's ability contained in the items not included in the imputation process would have improved the performance of this approach.

Finally, it could be that not only the missing proportion but also missing values distribution conditions the missingness approaches. The distribution of the ratio number of missing values to total number of items was positively skewed in SERCE dataset. In other words, most of the examinees in this assessment had low proportion of missing responses (*M*=.176, *SD*=.120). The effect of the missingness distribution on missing data handling approaches needs to be explored in future studies.

REFERENCES

Abad, F. J., Olea, J., & Ponsoda, V. (2009). The Multiple-Choice Model: Some Solutions for Estimation of Parameters in the Presence of Omitted Responses. *Applied Psychological Measurement*, *33*(3), 200–221. doi:10.1177/0146621608320760

Ake, C. F. (2005). Rounding after multiple imputation with non-binary categorical covariates. In *Annual meeting of the SAS Users Group International* (pp. 1–11). Philadelphia, PA. Retrieved from http://www2.sas.com/proceedings/sugi30/112-30.pdf

Allison, P. (2012, July 9). *Why maximum likelihood is better than multiple imputation*. Retrieved from http://www.statisticalhorizons.com/ml-better-than-mi

Allison, P. D. (2001). *Missing data* (Vol. 136). Sage publications.

Allison, P. D. (2006). Multiple imputation of categorical variables under the multivariate normal model. In *Annual Meeting of the American Sociological Association* (pp. 1–20). Montreal, CA. Retrieved from http://www.statisticalhorizons.com/wp-content/uploads/2012/01/Allison.CatVarImp.pdf

Allison, P. D. (2012). Handling missing data by maximum likelihood. *SAS Global Forum*, 1–21. Retrieved from http://www.statisticalhorizons.com/wp-content/uploads/MissingDataByML.pdf

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Andreis, F., & Ferrari, P. A. (2012). Missing data and parameters estimates in multidimensional item response models. *Electronic Journal of Applied Statistical Analysis*, *5*(3), 431–437. doi:10.1285/i20705948v5n3p431

Andridge, R. R., & Little, R. J. A. (2010). A review of hot deck imputation for survey non response. *International Statistical Review*, *78*(1), 40–64. doi:10.1111/j.1751-5823.2010.00103.x.A

Bernaards, C. A., & Sijtsma, K. (1999). Factor analysis of multidimensional polytomous item response data suffering from ignorable item nonresponse. *Multivariate Behavioral Research*, *34*(3), 277–313. doi:10.1207/S15327906MBR3403

Bernaards, C. A., & Sijtsma, K. (2000). Influence of imputation and EM methods on factor analysis when item nonresponse in questionnaire data are nonignorable. *Multivariate Behavioral Research*, *35*(3), 321–364. doi:10.1207/S15327906MBR3503

Bernaards, C. A., Belin, T. R., & Schafer, J. L. (2007). Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine*, *26*, 1368–1382. doi:10.1002/sim

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*.

Breiman, L. (2001). *Random forests*. *Machine learning* (pp. 1–33). Berkeley, CA. Retrieved from http://link.springer.com/article/10.1023/A:1010933404324

Carpenter, Bartlett & Kenward, n.d. Missing data. Retrieved from http://missingdata.lshtm.ac.uk/

Chou, C. P., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: a Monte Carlo study. *British Journal of Mathematical and Statistical Psychology*, *44*(2), 347-357.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351. doi:10.1037//1082-989X.6.4.330

Cook, R. J., Zeng, L., & Yi, G. Y. (2004). Marginal analysis of incomplete longitudinal binary data: a cautionary note on LOCF imputation. *Biometrics*, *60*(3), 820-828. Retrieved from https://www.researchgate.net/profile/Richard_Cook4/publication/8373852_Marginal_Analysis_of_Incomplete_Longitudinal_Binary_Data_A_Cautionary_Note_on_LOCF_Imputation/links/09e4150b03ead7d5cd000000.pdf

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological methods*, *1*(1), 16. Retrieved from http://curran.web.unc.edu/files/2015/03/CurranWestFinch1996.pdf

Custer, M., Sharairi, S., & Swift, D. (2012). A Comparison of scoring options for omitted and not-reached items through the recovery of IRT parameter when utilizing the Rasch model and joint maximum likelihood estimation. In *Annual Meeting of the National Council on Measurement in Education*. Vancouver, CA. Retrieved from http://files.eric.ed.gov/fulltext/ED531171.pdf

De Ayala, R. J. (2003). The effect of missing data on estimating a respondent's location using ratings data. *Journal of Applied Measurement*, *4*(1), 1–9.

De Ayala, R. J. (2006). Estimating person locations from partial credit data containing missing responses. *Journal of Applied Measurement*, *7*(3), 278–291. Retrieved from http://europepmc.org/abstract/MED/16807494

De Ayala, R. J. (2009). T*heory and practice of item response theory*. New York, NY. Guilford Publications Inc.

De Ayala, R. J., Plake, B. S., & Impara, J. C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement*, *38*(3), 213–234. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2001.tb01124.x/full

de Waal, Ton. Pannekoek, Jeroen. Scholtus, Sander. (2011). Imputation under edit constraints. *In Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, New Jersey. Retrieved from http://libraries.unl.edu/

DeMars, C. (2002). Missing data and IRT item parameter estimation. *Annual Meeting of the American Educational Research*, 15. Retrieved from http://www.jmu.edu/assessment/wm_library/missdata.pdf

Dempster, A. P., Larid, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society: Series B*, *39*(1), 1–38. Retrieved from http://www.stat.missouri.edu/~dsun/9720/EM_JRSSB.pdf

Dong, Y., & Peng, C.-Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, *2*(1), 222. doi:10.1186/2193-1801-2-222

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement*, *19*(2), 143-166. Retrieved from http://conservancy.umn.edu/bitstream/handle/11299/117478/1/v19n2p143.pdf

Emenogu, B. C., Falenchuk, O., & Childs, R. A. (2010). The effect of missing data treatment on Mantel-Haenszel DIF detection. *Alberta Journal of Educational Research*, *56*(4), 459. Retrieved from http://crawl.prod.proquest.com.s3.amazonaws.com/fpcache/3bd250b01319b0b99bd2fae7a2df3f78.pdf?AWSAccessKeyId=AKIAJF7V7KNV2KKY2NUQ&Expires=1466623638&Signature=%2F%2Fvb%2BrrRE50VkCA2g4kfgBTAL2k%3D

Enders (October, 2015). Dealing with missing data. Fall 2015 Nebraska Methodology workshop.

Enders, C. K. (1999). *The relative performance of full-information maximum likelihood estimation for missing data in structural equation models*. University of Nebraska-Lincoln. Retrieved from http://0-search.proquest.com.library.unl.edu/docview/304512560/fulltextPDF/3E78C9F46394308PQ/1?accountid=8116

Enders, C. K. (2001). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological methods*, *6*(4), 352.

Enders, C. K. (2002). Applying the Bollen-Stine bootstrap for goodness-of-fit measures to structural equation models with missing data. *Multivariate Behavioral Research, 37*(3), 359-377.doi:10.1080/01621459.1976.10481472.

Enders, C. K. (2008). A Note on the Use of Missing Auxiliary Variables in Full Information Maximum Likelihood-Based Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(3), 434–448. doi:10.1080/10705510802154307

Enders, C. K. (2010). *Applied missing data analysis* (1st Ed.). New York: The Guilford Press.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, *8*(3), 430-457. Retrieved from http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1065&context=edpsychpapers

Enders, C. K., & Peugh, J. L. (2004). Using an EM covariance matrix to estimate structural equation models with missing data: Choosing an adjusted sample size to improve the accuracy of inferences. *Structural Equation Modeling, 11*(1), 1-19.

Eurostat. (2014). *Handbook on methodology of modern business statistics*. Retrieved from http://www.cros-portal.eu/sites/default/files/Imputation-02-M-Deductive Imputation v1.0_2.pdf

Fellegi, I. P., & Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association, 71*(353), 17–35.

Ferrari, P. A., Annoni, P., Barbiero, A., & Manzi, G. (2011). An imputation method for categorical variables with application to nonlinear principal component analysis.

Computational Statistics and Data Analysis Data Analysis, *55*(7), 2410–2420. doi:10.1016/j.csda.2011.02.007

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, *45*(3), 225–245. Retrieved from http://doi.wiley.com/10.1111/j.1745-3984.2008.00062.x

Finch, J. F., West, S. G., & MacKinnon, D. P. (1997). Effects of sample size and nonnormality on the estimation of mediated effects in latent variable models. *Structural Equation Modeling: A Multidisciplinary Journal*, *4*(2), 87-107.

Foy, P., Brossman, B., & Galia, J. (2011). *Scaling the TIMSS and PIRLS 2011 achievement data*. (pp. 1–28). Boston, MA. Retrieved from http://timssandpirls.bc.edu/methods/pdf/TP11_Scaling_Achievement.pdf

Foy, P., Martin, M. O., & Mullis, I. V. S. (2011). *Reviewing the TIMSS and PIRLS 2011 Achievement Item Statistics* (pp. 1–27). Boston, MA. Retrieved from http://timssandpirls.bc.edu/methods/pdf/TP11_Reviewing_Achievement.pdf

Gebregziabher, M., & DeSantis, S. M. (2010). Latent class based multiple imputation approach for missing categorical data. *Journal of Statistical Planning and Inference*, *140*(11), 3252–3262. doi:10.1016/j.jspi.2010.04.020.

Glas, C. A. W., & Pimentel, J. L. (2008). Modeling Nonignorable Missing Data in Speeded Tests. *Educational and Psychological Measurement*, *68*(6), 907–922. doi:10.1177/0013164408315262

Gold, M. S., & Bentler, P. M. (2000). Treatments of missing data: A Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization. *Structural Equation Modeling*, *7*(3), 319-355. Retrieved from http://www.tandfonline.com/doi/abs/10.1207/S15328007SEM0703_1?journalCode=hsem20

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling*, *10*(1), 80-100. Retrieved from https://methodology.psu.edu/media/techreports/01-48.pdf

Graham, J. W. (2012). *Missing data: Analysis and design*. Springer Science & Business Media.

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research*, *31*(2), 197-218.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, *8*(3), 206-213. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.595.7125&rep=rep1&type=pdf

Holman, R., & Glas, C. A. W. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *The British Journal of Mathematical and Statistical Psychology*, *58*(Pt 1), 1–17. doi:10.1348/000711005X47168.

Horton, N. J., Lipsitz, S. R., & Parzen, M. (2003). A Potential for Bias When Rounding in Multiple Imputation. *The American Statistician, 57*(4), 229–232. doi:10.1198/0003130032314

Hu, L. T., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological bulletin*, *112*(2), 351. Retrieved from http://www3.nd.edu/~kyuan/courses/sem/readpapers/Hu-Bentler-Kano-PB-92.pdf

Huisman, M. (2000). Imputation of missing item responses: Some simple techniques. *Quality and Quantity*, *34*(4), 331–351. Retrieved from http://link.springer.com/article/10.1023/A:1004782230065

Huisman, M., & Molenaar, I. W. (2001). Imputation of missing scale data with item response models. *Lecture Notes in Statistics*, *157*, 221–244. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4613-0169-1_13

Kelley, T. L. Ruch G. M. and Terman, L. M. (1922). *The Project Gutenberg EBook of Stanford Achievement Test*. Retrieved from http://archive.org/stream/stanfordachievem22425gut/pg22425.txt

Lee, K. J., & Carlin, J. B. (2010). Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology*, *171*(5), 624–32. doi:10.1093/aje/kwp425

Leite, W., & Beretvas, S. N. (2010). The performance of multiple imputation for Likert-type items with missing data. *Journal of Modern Applied Statistical Methods*, *9*(1), 64–74. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol9/iss1/8/?utm_source=digitalcommons.wayne.edu%2Fjmasm%2Fvol9%2Fiss1%2F8&utm_medium=PDF&utm_campaign=PDFCoverPages

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of American Statistical Association*, *83*(404), 1198–1202. Retrieved from http://medrescon.tripod.com/docs/little_paper.pdf

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York, J. Wiley & Sons.

Liu, G., & Gould, A. L. (2002). Comparison of alternative strategies for analysis of longitudinal trials with dropouts. *Journal of biopharmaceutical statistics*, *12*(2), 207-226.

Lord, F. M. (1973). *Estimation of latent ability and item parameter when there are omitted responses*. *Psychometrika* (p. 33). Princeton, NJ. Retrieved from http://link.springer.com/article/10.1007/BF02291471

Lord, F. M. (1980). Estimating ability and item parameter. In M. F. Lord (Ed.), *Applications of item response theory to practical testing problems* (pp. 179-191). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M. (1980). Omitted responses and formula scoring. In M. F. Lord (Ed.), *Applications of item response theory to practical testing problems* (pp. 225-231). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ludlow, L. H., & O'leary, M. (1999). Scoring Omitted and Not-Reached Items: Practical Data Analysis Implications. *Educational and Psychological Measurement*, *59*(4), 615–630. doi:10.1177/0013164499594004

Mallinckrodt, C. H., Clark, W. S., & David, S. R. (2001). Accounting for dropout bias using mixed-effects models. *Journal of biopharmaceutical statistics*, *11*(1-2), 9-21.

Martin, M.O. & Mullis, I.V.S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timssandpirls.bc.edu/methods/

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, *9*(4), 538–573. Retrieved from http://projecteuclid.org/euclid.ss/1177010269

Mislevy, R. J., & Wu, P. (1988). *Inferring examinee ability when some item responses are missing* (p. 80). Princeton, NJ. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA201421

Mislevy, R. J., & Wu, P.-K. (1996). *Missing Responses and IRT Ability Estimation: Omits, Choice, Time Limits, and Adaptive Testing.* Princeton, NJ. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA313823

Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M. G., Mallinckrodt, C., & Carroll, R. J. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, *5*(3), 445-464.

Muthén, L.K. and Muthén, B.O. (1998-2012). *Mplus User's Guide*. Seventh Edition. Los Angeles, CA: Muthén & Muthén

Not-reached item (n.d.). *In the NAEP glossary of term*. Retrieved from https://nces.ed.gov/nationsreportcard/glossary.aspx?nav=y

O'Muircheartaigh, C., & Moustaki, I. (1999). Symmetric pattern models: a latent variable approach to item non-response in attitude scales. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *162*(2), 177–194. doi:10.1111/1467-985X.00129

Organisation for Economic Co-operation and Development. (2009). *PISA 2006 Technical Report* (p. 418). Paris. Retrieved from http://www.oecd.org/pisa/pisaproducts/42025182.pdf

Organisation for Economic Co-operation and Development. (2012). *PISA 2009 Technical Report* (p. 390). OECD Publishing. doi:10.1787/9789264167872-en

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, *31*(3), 200–219. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.1994.tb00443.x/full

Pimentel, J. L. (2005). *Item response theory modeling with nonignorable missing data*. University of Twente, The Netherlands. Retrieved from http://doc.utwente.nl/50891/1/thesis_Pimentel.pdf

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of

regression models. *Survey Methodology*, *27*(1), 85–95. Retrieved from http://www.statcan.gc.ca/ads-annonces/12-001-x/5857-eng.pdf

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT)*. Princeton, NJ. Retrieved from https://www.ets.org/Media/Research/pdf/RR-10-11.pdf

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. Retrieved from http://biomet.oxfordjournals.org/content/63/3/581.short

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

Rutkowski, L., & Rutkowski, D. (2010). Getting it "better": the importance of improving background questionnaires in international large scale assessment. *Journal of Curriculum Studies*, *42*(3), 411–430. doi:10.1080/00220272.2010.487546

Savalei, V. (2008). Is the ML chi-square ever robust to nonnormality? A cautionary note with missing data. *Structural Equation Modeling*, *15*(1), 1-22. Retrieved from https://www.researchgate.net/profile/Victoria_Savalei/publication/228678533_Is_the_ML_Chi-Square_Ever_Robust_to_Nonnormality_A_Cautionary_Note_With_Missing_Data/links/0c9605345b23de44d8000000.pdf

Savalei, V., & Bentler, P. M. (2005). A statistically justified pairwise ML method for incomplete nonnormal data: A comparison with direct ML and pairwise ADF. *Structural Equation Modeling*, *12*(2), 183-214.

Savalei, V., & Bentler, P. M. (2009). A two-stage approach to missing data: Theory and application to auxiliary variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(3), 477-497. Retrieved from http://escholarship.org/uc/item/89f8q6jc

Savalei, V., & Falk, C. F. (2014). Robust two-stage approach outperforms robust full information maximum likelihood with incomplete nonnormal data. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(2), 280-302.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. doi:10.1037//1082-989X.7.2.147

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*. Retrieved from http://www.tandfonline.com/doi/abs/10.1207/s15327906mbr3304_5

Shculte Nordhold, Eric. Hoof Van Huijsduijnen, Jan. (1997). The treatment of item nonresponse during the editing of survey results. In Eurostat Seminar on New Techniques and Technologies for Statistics 1999 (Eds.). *New Techniques and Technologies for Statistics II: Proceedings of the Second Bonn Seminar*. IOS Press Inc. Retrieved from http://ec.europa.eu/eurostat/ramon/statmanuals/files/Handbook%20on%20surveys.pdf

Shin, S.-H. (2009). How to treat omitted responses in Rasch model-based equating. *Practical Assessment, Research & Evaluation*, *14*(1). Retrieved from http://www.pareonline.net/pdf/v14n1.pdf

Sijtsma, K., & van der Ark, L. A. (2003). Investigation and treatment of missing item scores in test and questionnaire data. *Multivariate Behavioral Research*, *38*(4), 505–528. doi:10.1207/s15327906mbr3804

Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological methods*, *6*(4), 317. Retrieved from http://www.education.umd.edu/EDMS/fac/Harring/651-Spring-2016/Readings/Sinharay-2001.pdf

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics (Oxford, England)*, *28*(1), 112–8. doi:10.1093/bioinformatics/btr597

Stocking, M. L., Eignor, D. R., & Cook, L. L. (1988). Factors affecting the sample invariant properties of linear and curvilinear observed-and true-score equating procedures. *ETS Research Report Series*, *1988*(2), i-71. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/j.2330-8516.1988.tb00297.x/pdf

Tabachnick, B. G., Fidell, L. S. (2013). *Using Multivariate Statistics*, 6th ed. Boston: Pearson.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, *49*(4), 501-519.

Thoemmes, F., & Rose, N. (2014). A Cautious Note on Auxiliary Variables That Can Increase Bias in Missing Data Problems. *Multivariate Behavioral Research*, *49*(5), 443–459. doi:10.1080/00273171.2014.931799

Trevino, E., Bogoya, D., Glejberman, D., Castro, M., Espinoza, G., Tamassia, C., Leigh, E. (2008). *Reporte técnico del SERCE*. Santiago de Chile, CL.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, *43*(1), 39-55.

Tutz, G. (1997). Sequential models for ordered responses. In *Handbook of modern item response theory* (pp. 139-152). Springer New York.

United Nation Educational, Scientific and Cultural Organization. (2009). *Los aprendizajes de los estudiantes de América Latina y el Caribe. Reporte Técnico.* Retrieved from http://unesdoc.unesco.org/images/0019/001902/190297s.pdf

United Nation Educational, Scientific and Cultural Organization. (n.d.). First regional comparative and explanatory study, PERCE. Retrieve from http://www.unesco.org/new/en/santiago/education/education-assessment/first-regional-comparative-and-explanatory-study/

United Nation Educational, Scientific and Cultural Organization. (n.d.). Quality of education assessment. Retrieved from http://portal.unesco.org/geography/en/ev.php-URL_ID=7732&URL_DO=DO_TOPIC&URL_SECTION=201.html

United Nation Educational, Scientific and Cultural Organization. (n.d.). Second regional comparative and explanatory study, SERCE. Retrieved from http://www.unesco.org/new/en/santiago/education/education-assessment/second-regional-comparative-and-explanatory-study-serce/

United Nation Educational, Scientific and Cultural Organization. (2010). *Compendio de los manuales del SERCE*. Retrieved from http://unesdoc.unesco.org/images/0019/001919/191940s.pdf

United Nation Educational, Scientific and Cultural Organization. (2006). *Second regional comparative and explanatory study, SERCE, datasets* [Dataset]. Retrieved from http://www.unesco.org/new/en/santiago/education/education-assessment-llece/perce-serce-databases/

United Nation Educational, Scientific and Cultural Organization. (n.d.). Third regional comparative and explanatory study, TERCE. Retrieved from http://www.unesco.org/new/en/santiago/education/education-assessment-llece/third-regional-comparative-and-explanatory-study-terce/

Valdez, H., Trevino, E., Acevedo, C. G., Castro, M., Carrillo, S., Costilla, R., Bogoya, D., Pardo, C. (2008). *Primer reporte de los resultados del SERCE*. Santiago de Chile, CL. Retrieved from http://unesdoc.unesco.org/images/0016/001606/160660s.pdf

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research, 16*, 219–242. Retrieved from http://smm.sagepub.com/content/16/3/219.short

van Buuren, S. (2010). Item imputation without specifying scale structure. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 6*(1), 31–36. doi:10.1027/1614-2241/a000004.

van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine, 18*(6), 681-694. Retrieved from http://www.stefvanbuuren.nl/publications/Multiple%20imputation%20-%20Stat%20Med%201999.pdf

van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med, 18*(6), 681–694. doi:10.1002/(SICI)1097-0258(19990330)18:6<681::AID-SIM71>3.0.CO;2-R [pii].

van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation, 76*(12), 1049–1064. doi:10.1080/10629360600810434

van den Wollenberg, A. L. (1979). *The Rasch model and time-limit tests: An application and some theoretical contributions*. Katholieke Universiteit te Nijmegen.

van Ginkel, J. R., van der Ark, L. A., & Sijtsma, K. (2007). Multiple imputation of item scores in test and questionnaire data, and influence on psychometric results. *Multivariate Behavioral Research, 42*(2), 387–414. doi:10.1080/00273170701360803

Verhelst, J., Abs, R., Vandeweghe, M., Mockel, J., Legros, J. J., Copinschi, G.,Mabler, C., Velkeniers, B., Vanhaelst, L., Van Aelst, A., De Rijdt, D., Stevenaert, A., & Beckers, A. (1997). Two years of replacement therapy in adults with growth hormone deficiency. *Clinical endocrinology, 47*(4), 485-494.

Vermunt, J. K., van Ginkel, J. R., van der Ark, L. A., & Sijtsma, K. (2008). Multiple Imputation of Incomplete Categorical Data Using Latent Class Analysis. *Sociological Methodology*, 369–397. doi:10.1111/j.1467-9531.2008.00202.x

Wayman, J. C., & Swaim, R. C. (2002). Practical Considerations in Constructing a Multiple Imputation Model A Data Example. In *Annual Meeting of the American Educational Research Association, New Orleans, LA*.

Wolkowitz, A. A., & Skorupski, W. P. (2013). A Method for Imputing Response Options for Missing Data on Multiple-Choice Assessments. *Educational and Psychological Measurement*, *73*(6), 1036–1053. doi:10.1177/0013164413497016

Yuan, K. H. (2007). Normal theory ML for missing data with violation of distribution assumptions. *Manuscript submitted for publication*.

Yuan, K., & Lu, L. (2008). SEM with missing data and unknown population distributions using two-stage ML: Theory and its application. *Multivariate Behavioral Research*, *43*(4), 621–652. doi:10.1080/00273170802490699

Yuan, K.-H. (2009). Normal distribution based pseudo ML for missing data: With applications to mean and covariance structure analysis. *Journal of Multivariate Analysis*, *100*(9), 1900–1918. doi:10.1016/j.jmva.2009.05.001

Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/0081-1750.00078/abstract

Yuan, K.-H., Yang-Wallentin, F., & Bentler, P. M. (2012). ML versus MI for Missing Data with Violation of Distribution Conditions. *Sociological Methods and Research*, *41*(4), 598–629. doi:10.1177/0049124112460373

Zimowsky, M., Mustaki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG* (Version 3.0.2327.2) [Computer program]. Mooresville, IN: Scientific Software.

APPENDIX A. Missing data handling methods

1. Complete data analysis
    1.1. Complete-case analysis (listwise deletion)
    1.2. Available-case analysis (pairwise deletion)
2. Incomplete data analysis
    2.1. Single imputation
        a) Unconditional mean imputation
        b) Person mean imputation
        c) Regression imputation
        d) Stochastic regression imputation
        e) Hot-deck imputation
        f) Cold-deck imputation
        g) Similar response pattern imputation
        h) Last observation carried forward
        i) Worse observation carried forward
    2.2. Maximum likelihood estimation
        a) Expectation-maximization (EM)
    2.3. Multiple imputation
        a) Fully conditional specification (FCS)
        b) Multiple imputation with data augmentation (MIDA)

APPENDIX B. Items retained (✓) or removed (✗) based on item analysis

Item	Rasch	2PL	Item	Rasch	2PL	Item	Rasch	2PL
DM6B1IT01	✓	✓	DM6B3IT01	✗	✓	DM6B5IT01	✗	✗
DM6B1IT02	✓	✓	DM6B3IT02	✗	✗	DM6B5IT02	✗	✗
DM6B1IT03	✓	✓	DM6B3IT03	✗	✗	DM6B5IT03	✓	✓
DM6B1IT04	✓	✓	DM6B3IT04	✓	✓	DM6B5IT04	✗	✗
DM6B1IT05	✓	✓	DM6B3IT05	✓	✓	DM6B5IT05	✗	✗
DM6B1IT06	✓	✓	DM6B3IT06	✓	✓	DM6B5IT06	✗	✗
DM6B1IT07	✓	✗	DM6B3IT07	✓	✓	DM6B5IT07	✗	✗
DM6B1IT08	✓	✗	DM6B3IT08	✗	✗	DM6B5IT08	✗	✗
DM6B1IT09	✓	✓	DM6B3IT09	✗	✗	DM6B5IT09	✓	✓
DM6B1IT10	✓	✓	DM6B3IT10	✗	✗	DM6B5IT10	✗	✗
DM6B1IT11	✓	✓	DM6B3IT11	✗	✗	DM6B5IT11	✓	✓
DM6B1IT12	✓	✓	DM6B3IT12	✓	✓	DM6B5IT12	✗	✗
DM6B1IT13	✓	✓	DM6B3IT13	✗	✗	DM6B5IT13	✗	✗
DM6B1IT14	✓	✓	DM6B3IT14	✗	✗	DM6B5IT14	✓	✓
DM6B1IT15	✓	✓	DM6B3IT15	✗	✗	DM6B5IT15	✗	✗
DM6B1IT16	✓	✓	DM6B3IT16	✗	✗	DM6B5IT16	✗	✗
DM6B2IT01	✓	✓	DM6B4IT01	✓	✓	DM6B6IT01	✗	✗
DM6B2IT02	✓	✓	DM6B4IT02	✓	✓	DM6B6IT02	✓	✗
DM6B2IT03	✓	✓	DM6B4IT03	✓	✓	DM6B6IT03	✓	✗
DM6B2IT04	✓	✓	DM6B4IT04	✗	✗	DM6B6IT04	✗	✗
DM6B2IT05	✓	✓	DM6B4IT05	✗	✗	DM6B6IT05	✗	✗
DM6B2IT06	✓	✓	DM6B4IT06	✗	✗	DM6B6IT06	✗	✗
DM6B2IT07	✓	✓	DM6B4IT07	✓	✓	DM6B6IT07	✓	✓
DM6B2IT08	✓	✓	DM6B4IT08	✗	✗	DM6B6IT08	✗	✗
DM6B2IT09	✓	✓	DM6B4IT09	✓	✗	DM6B6IT09	✗	✗
DM6B2IT10	✓	✓	DM6B4IT10	✗	✗	DM6B6IT10	✗	✗
DM6B2IT11	✓	✓	DM6B4IT11	✓	✓	DM6B6IT11	✗	✗
DM6B2IT12	✓	✓	DM6B4IT12	✓	✓	DM6B6IT12	✗	✗
DM6B2IT13	✓	✓	DM6B4IT13	✓	✗	DM6B6IT13	✗	✗
DM6B2IT14	✓	✓	DM6B4IT14	✓	✓	DM6B6IT14	✓	✓
DM6B2IT15	✓	✓	DM6B4IT15	✓	✓	DM6B6IT15	✓	✓
DM6B2IT16	✗	✓	DM6B4IT16	✓	✓	DM6B6IT16	✓	✓

APPENDIX C. Item parameters per IRT model

Rasch model

Item	Difficulty	SE	Item	Difficulty	SE	Item	Difficulty	SE
M01	-2.528	0.067	M20	0.525	0.058	M39	0.933	0.027
M02	-1.356	0.054	M21	-0.028	0.054	M40	0.638	0.028
M03	0.495	0.059	M22	-2.157	0.064	M41	1.821	0.035
M04	-1.569	0.056	M23	-1.638	0.055	M42	-0.857	0.025
M05	-0.631	0.055	M24	-3.807	0.094	M43	0.597	0.025
M06	-0.239	0.056	M25	1.315	0.068	M44	0.744	0.026
M07	1.437	0.071	M26	-1.887	0.056	M45	2.553	0.049
M08	0.614	0.056	M27	-0.412	0.049	M46	1.616	0.037
M09	0.229	0.057	M28	-0.541	0.057	M47	1.440	0.033
M10	1.448	0.079	M29	0.125	0.051	M48	-1.092	0.026
M11	1.459	0.078	M30	0.932	0.064	M49	-1.840	0.028
M12	0.388	0.055	M31	1.893	0.081	M50	-1.201	0.026
M13	0.850	0.061	M32	0.150	0.029	M51	-0.660	0.025
M14	0.014	0.054	M33	-1.321	0.029	M52	1.365	0.037
M15	0.979	0.064	M34	0.992	0.032	M53	1.074	0.034
M16	0.421	0.061	M35	-0.601	0.029	M54	-0.024	0.031
M17	-3.799	0.090	M36	0.772	0.031	M55	2.078	0.051
M18	-1.131	0.056	M37	-2.700	0.032	M56	3.965	0.108
M19	-1.369	0.056	M38	-1.667	0.026	M57	1.190	0.039

* item discrimination is 1 for all the items.

2PL model

Item	Discrimination	SE	Difficulty	SE	Item	Discrimination	SE	Difficulty	SE
M01	1.133	0.119	-1.752	0.133	M28	2.017	0.108	1.065	0.038
M02	1.058	0.093	-0.689	0.060	M29	2.097	0.135	1.681	0.061
M03	0.769	0.072	1.484	0.147	M30	1.933	0.297	3.210	0.218
M04	1.426	0.122	-0.769	0.051	M31	1.145	0.080	-3.105	0.166
M05	1.634	0.128	-0.024	0.039	M32	0.929	0.038	0.832	0.042
M06	1.732	0.125	0.259	0.041	M33	1.033	0.044	-0.660	0.032
M07	0.965	0.078	0.944	0.093	M34	0.597	0.036	2.465	0.143
M08	1.135	0.090	1.970	0.142	M35	1.382	0.048	0.023	0.022
M09	0.883	0.078	2.395	0.200	M36	0.540	0.034	2.329	0.143
M10	0.363	0.053	2.725	0.431	M37	0.695	0.042	-2.775	0.148
M11	0.357	0.055	3.896	0.615	M38	1.151	0.042	-0.952	0.032
M12	0.560	0.061	1.189	0.161	M39	0.398	0.031	3.461	0.260
M13	0.438	0.057	3.472	0.463	M40	1.534	0.048	0.984	0.027
M14	1.208	0.093	0.959	0.084	M41	1.517	0.050	-0.182	0.018
M15	1.289	0.120	-2.479	0.167	M42	0.335	0.027	3.194	0.261
M16	2.185	0.112	-0.302	0.028	M43	1.683	0.084	2.307	0.068
M17	2.173	0.115	-0.451	0.029	M44	2.366	0.096	1.433	0.028
M18	1.407	0.069	0.933	0.050	M45	1.305	0.048	1.755	0.049
M19	1.411	0.070	0.503	0.041	M46	1.196	0.045	-0.405	0.024
M20	3.375	0.217	-0.837	0.027	M47	1.510	0.058	-0.930	0.028
M21	1.524	0.082	-0.705	0.041	M48	1.063	0.039	-0.543	0.029
M22	2.704	0.249	-1.735	0.067	M49	1.120	0.041	-0.022	0.023
M23	1.631	0.085	1.441	0.061	M50	1.077	0.049	0.584	0.035
M24	1.490	0.087	-0.895	0.045	M51	2.186	0.125	1.742	0.050
M25	0.986	0.057	0.263	0.052	M52	1.979	0.190	2.969	0.147
M26	2.308	0.108	0.072	0.027	M53	1.681	0.078	1.330	0.041
M27	1.022	0.059	0.756	0.059					

Note: Do not compare the item parameters across models because they are different items (e.g., M01 from the Rasch is not the same as M01 from the 2PL).

Endnotes

[1] *R* software has a module called multivariate imputation by chained equations (MICE) that implements this method. Given its popularity, sometimes FCS is called MICE.

[2] The National Assessment of Educational Progress (NAEP) defines a not-reached item as the one "to which the student did not respond because the time limit was up for the section of the assessment on which s/he was working. After the first "not reached" item, the student will have no responses to any further questions on that section of the assessment" (NAEP Glossary, n.d.). Therefore, the first item with missing response is treated as [intentionally] omitted and the following non-responses are treated as not administered (Mislevy and Wu, 1988). The Australian Council for Educational Research (ACER) defines not-reached items when there are more than two blank answers.

[3] Argentina, Brazil, Chile, Colombia, Costa Rica, Cuba, Ecuador, El Salvador, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Dominican Republic, Uruguay, and the Mexican State of Nuevo Leon.

[4] Website: http://www.unesco.org/new/en/santiago/education/education-assessment/, and for the SERCE data: http://www.unesco.org/new/fileadmin/MULTIMEDIA/FIELD/Santiago/zip/bcf362e6.zip

[5] In TIMSS, 2011, 3.2% and 4.5% of the students have omitted and not-reached responses, respectively. In PIRLS (2011) 8.9% of students omitted responses(Foy et al., 2011; Organisation for Economic Co-operation and Development, 2012).

[6] It is not possible to talk about not-reached and omitted responses in rating scale data, therefore non-answered items are referred as missing responses.

[7] RDS replaces a missing value with a random draw from the permitted response options. IAS imputes (a) the incorrect answer, when item is scored as right or wrong, or (b) the answer that is socially most undesirable (i.e., worst case scenario) for attitude items. IMS imputes the missing values with the mean of observed cases in the item. PMS replaces missing values with the average of the observed responses for each case. CIM adjusts the item mean by taking into account the respondent's ability. ICS imputes the missing value with the observed responses on the item with which the item with missing values has the highest correlation. HNC uses as the donor the first complete case after the incomplete case. HDD uses the complete case for which the distance from the incomplete case is minimized. HDR first selects several donors with small distance from the incomplete case. Then, one of them is randomly selected (Huisman, 2000).

[8] "Even though Schafer (1997) provided a way to combine likelihood ratio test statistics in MI, no empirical studies have evaluated the performance of this pooled likelihood ratio test under various data condition. Also, this test has not been incorporated into popular statistical packages" (Dong & Peng, 2013, p. 15)

[9] There is a website that more formally tracks the work done with MI, http://www.stefvanbuuren.nl/mi/index.html. However, this statement is done basically comparing the number of papers that either have the methods as part their title or they are mentioned in the document.

[10] Mean conditional on the covariates (CM): "imputes the mean based on the available scores across all items of all persons within the same covariate class, and imputes this mean for each missing in this covariate class". Overall mean (OM): imputation based on the data matrix mean. Two-way imputation (TW): the imputation for the missing observation (i, j) = IMS + PMS – OM (Bernaards & Sijtsma, 2000). The two-way imputation with normally distributed error (TW-E) is an imputation method that corrects both for person effect and item effect, and adds a random error drawn from a normal distribution ($\mu=0$, σ_ε^2) to the imputation process. The corrected item-mean with normally distributed error (CIM-E) implies that "the item mean is corrected for person i's score level relative to the mean of the items to which he/she responded. Normally distributed errors are added to CIM_{ij} using a procedure similar to the one used for adding normally distributed errors in method TW-E" (van Ginkel et al., 2007, pp. 391-393).

[11] The factor loading recovery was measured with the Tucker's ϕ (Burt, 1948; Tucker, 1951) and the $\overline{D^2}$ in Bernaards and Sitjsma (1999). In Bernaards and Sijtsma's (2000) study $\overline{D^2}$ and $\Pi\gamma$ (i.e., the product of estimated eigenvalues) were the indicators. The Tucker's ϕ is a coefficient of congruence that

measures the similarities between the factors derived from factor analysis. It is basically a correlation coefficient. The $\overline{D^2}$ index is the average of the D^2 across all the sample replications within each condition. D^2 is the sum of squared differences, divided by the number of extracted factors based on the complete data and the corresponding factor loadings based on the imputed datasets using the methods aforesaid (Bernaards & Sijtsma, 1999, 2000).

[12] BIC: Bayesian Information Criterion, AIC: Akaike Information Criterion, and AIC3 is a modified index of AIC Vermunt, van Ginkel, van der Ark, and Sijtsma (2008).

[13] Sijtsma and van der Ark (2003) study is based on two main parts. Only one part is presented in this document. The second part of the study refers to two methods to determine the missingness mechanism, originally proposed by Huisman (1999). One of them is done at the data matrix level (the Huisman's (1999) asymptotic test), while the second method does it at the item level. For details, see Sijtsma and van der Ark's (2003) publication.

[14] "R_{1c} tests whether the response functions of the *J* items are logistic with the same slope against the alternative that they deviate from these conditions, and statistic Q_2 tests whether the test is unidimensionality against the alternative of multidimensionality" (Sijtsma & van der Ark, 2003, p. 520).

[15] SERCE missing data were recoded following the procedure described by other large-scale assessments (e.g., PISA, TIMSS, and PIRLS). That is, the first missing response in the blank-response string was considered omitted and the rest are coded as not-reached. For example, a student's pattern response such as *43231Z1Z43442Z3ZZZZZZZZZ* (where "*Z*" is SERCE's code for missing responses) was recoded as *43231Z1Z43442Z3ZRRRRRRRR*, where "*R*" are not-reached responses. Notice that the firs "*Z*" was kept, given that this is normally taken as reached, thus intentionally omitted (Mislevy & Wu, 1988).

[16] Thanks to Yem Ahiatsi for writing the algorithm.

[17] Thanks to Dr. Rafael De Ayala for writing the algorithm.

[18] In the imputation with regression model, variables with non-missing values are considered covariate in the imputation process.