*Journal of Educational Measurement*
*Fall 2008, Vol. 45, No. 3, pp. 225–245*

# Estimation of Item Response Theory Parameters in the Presence of Missing Data

**Holmes Finch**
*Ball State University*

*Missing data are a common problem in a variety of measurement settings, including responses to items on both cognitive and affective assessments. Researchers have shown that such missing data may create problems in the estimation of item difficulty parameters in the Item Response Theory (IRT) context, particularly if they are ignored. At the same time, a number of data imputation methods have been developed outside of the IRT framework and been shown to be effective tools for dealing with missing data. The current study takes several of these methods that have been found to be useful in other contexts and investigates their performance with IRT data that contain missing values. Through a simulation study, it is shown that these methods exhibit varying degrees of effectiveness in terms of imputing data that in turn produce accurate sample estimates of item difficulty and discrimination parameters.*

Psychometricians and other measurement professionals are familiar with the phenomenon of missing item responses for both cognitive and affective assessments. For example, examinees may leave one or more items unanswered either inadvertently or because they do not know the answer and are afraid to guess. Respondents to a questionnaire might feel inhibited in answering items dealing with a sensitive topic, leading to missing data. Much research has been conducted regarding the impact of missing data on statistical analyses in general and a variety of methods have been developed for dealing with the problem. The interested reader is encouraged to see Schafer and Graham (2002) for a comprehensive review of methods for dealing with missing data. In addition to the Schafer and Graham paper, there are a number of other comprehensive discussions regarding specific types of missing data that researchers might see in practice (Bernaards & Sijtsma, 1999; Peng & Zhu, 2005; Schafer, 1997; Schafer & Olsen, 1998; Sinharay, Stern, & Russell, 2001). Data that are missing completely at random (MCAR) can be thought of as having no systematic cause; i.e., the missing data are a simple random sample of the observed data (Schafer, 1997, p. 11). When data are missing at random (MAR), the probability of a value being missing is dependent on some measurable characteristic of the individual but not on the missing value itself. Schafer (p. 11) points out that for data to be MAR, the variable associated with the probability of data being missing must be observed. Finally, for values missing not at random (MNAR), the likelihood of a variable value being missing is directly related to the value of the variable itself. In the context of actual testing data, researchers have identified a variety of reasons for item responses to be missing, with the most likely reasons including omission because the examinee believes that he/she does not know the answer (De Ayala, Plake, & Impara, 2001), or through inadvertently skipping the item (Huisman & Molenaar, 2001). The former case may represent an example of MAR (if the nonresponse is related to the underlying latent trait but not the actual item response had it been given) or MNAR (if nonresponse is directly related to the actual item response had it been given), while the latter case represents MCAR data, in which no systematic mechanism is associated with nonresponse.

Mislevy and Wu (1988) reported on the estimation of ability when examinees did not respond to all items and found support for the contention that when the mechanism behind missing data is unrelated to the parameters of interest (in their case examinee ability), the quality of estimation is largely unaffected. On the other hand, when the reason that an examinee did not respond to an item cannot be verified (which is perhaps most often the case), Mislevy and Wu (1996) suggested that, if possible, the missing data mechanism be incorporated into parameter estimation so as to avoid potentially biased results. If this inclusion cannot be done, it is very possible that examinee and item parameter estimates will be biased.

This study was conducted to assess the impact of various approaches for dealing with missing data on the estimation of item parameters for the three-parameter logistic (3PL) model. Prior simulation-based research, described below, has examined the impact of missing item responses on some of these estimates (e.g., difficulty) but not methods of data imputation, while a variety of methods for dealing with missing item response data have been used with individual data sets but not in a Monte Carlo study. This study furthers this research by comparing several of these approaches for dealing with missing item response data using simulation, as well as an analysis of a real data set with missing item responses.

## Methods of Dealing with Missing Data

Several methods for dealing with missing data are examined in this study, each of which is described briefly below. All of these methods are available in standard statistical software such as SAS, SPSS, or BILOG, and some involve data imputation. The methods not involving imputation, including treating the missing items as not presented (NP), incorrect (IN) or fractionally correct (FR), can be carried out directly with the item parameter estimation software BILOGMG (Zimowski, Muraki, Mislevy, & Bock, 2003). Imputation involves the estimation, based upon other sources of information such as nonmissing observations in the data set, of what a missing data value might have been. For example, in the context of testing, missing item response values can be imputed using an examinee's responses to other items on the instrument as well as the responses of other examinees. The approaches to imputation described below represent common and/or newer methods, though they are not an exhaustive set. They were selected because prior research has demonstrated their potential effectiveness in other contexts, so they may be worthwhile candidates for use with item response data. The interested reader is encouraged to refer to Schafer and Graham (2002) for a more comprehensive review of imputation methods.

### *Corrected Item Mean Substitution (CM) Imputation*

CM imputation (Bernaards & Sijtsma, 2000; Huisman & Molenaar, 2001; Sijtsma & van der Ark, 2003) involves the calculation of a weight function reflecting the relative performance of an examinee on their nonmissing item responses and the application of the weight to mean performance on the item across examinees. First, a person mean is calculated as:

$$PM_i = \frac{\sum_j x_{ij}}{J_i}, \tag{1}$$

where
$x_{ij}$ = response to item $j$, where $x_{ij}$ is not missing, for examinee $i$.
$J_i$ = number of nonmissing items for examinee $i$.

Similarly, an item mean (IM) is also calculated:

$$IM_j = \frac{\sum_i x_{ij}}{I_j}, \tag{2}$$

where $I_j$ = number of individuals with nonmissing response for item $j$.

Finally, CM is calculated as:

$$\tilde{x}_{ij} = \left[ \frac{PM_i}{\frac{1}{\#\text{obs}(i)} \sum_j IM_j} \right] IM_j, \tag{3}$$

where
$\tilde{x}_{ij}$ = imputed value for examinee $i$, item $j$.
#obs($i$) = number of nonmissing item responses for examinee $i$.

In short, *PM* is the mean item response for a single examinee across all nonmissing items while *IM* is the mean nonmissing item response for a single item across examinees. The weight (appearing in the brackets) will be larger than 1.0 for examinees with higher than average scores on the test, and less than 1.0 for those with lower than average test performance. This takes into account the relative performance of the examinee by providing higher imputed values for above average performance and lower imputed values for below average performance (Bernaards & Sijtsma, 2000). For dichotomous item responses, this translates into a greater likelihood of imputing a correct response for individuals whose relative performance is higher than most of their peers. The imputed value for a missing item response is this weighted mean of the nonmissing item responses rounded to either 0 or 1.

### *Response Function (RF) Imputation*

A second method for imputing missing item response data included in this study is response function (RF) imputation (Sijtsma & van der Ark, 2003). This approach is nonparametric in nature, in that while it assumes the presence of an underlying ability parameter, $\theta$, it does not assume anything about the item parameters nor does it attempt to estimate them using a likelihood function. In order to impute a value of missing item $j$ for individual $i$, a summary score is calculated based on the nonmissing item responses, $\hat{R}_{(-j)i}$, which is known as the rest score. Formally, this score takes the form

$$\hat{R}_{(-j)i} = PM_i(J - 1), \tag{4}$$

where $J$ represents the total number of items.

The probability of imputing a correct response to item $j$ for individual $i$ with a given integer value on $\hat{R}_{(-j)i}$ is estimated by the proportion of examinees with rest score $\hat{R}_{(-j)i}$ and without missing data on the target item who answered the item correctly. If the examinee in question does not have an integer value for $\hat{R}_{(-j)i}$ the proportion of examinees answering item $j$ correctly for integer values below and above the value of $\hat{R}_{(-j)i}$ are used, and linear interpolation is employed to find the proportion of correct responses, which serves as an estimate of the probability of a correct response. The imputed item response is then drawn from the Bernoulli distribution using this estimate as the parameter value. A more complete description of this method can be found in Sijtsma and van der Ark (2003).

### *Multiple Imputation (MI)*

MI has been described in very complete detail in several places (Leite & Beretvas, 2004; Schafer, 1997; Schafer & Graham, 2002; Schafer & Olsen, 1998; Sinharay et al., 2001). MI, first proposed by Rubin (1987), was developed as an alternative to earlier approaches to imputation such as mean substitution, Hot Deck imputation, regression-based imputation and conditional distribution imputation (Huisman & Molenaar, 2001; Madow, Nisselson, & Olkin, 1983). Unlike these single imputation techniques, MI accounts for the inherent uncertainty in sampling from a population by introducing a degree of randomness to the imputations and creating $m$ imputed data sets, each of which can be analyzed in standard ways. MI can incorporate information from other variables into the imputation process to provide more accurate values.

The use of MI requires an assumption about the probability model underlying a set of data, such as multivariate normality (frequently used for continuous variables) or a multinomial distribution (common with categorical variables).[1] Once a probability model is chosen, parameter estimates are obtained using the Bayesian posterior distribution based upon the likelihood function of the proposed model, the observed data, and a prior distribution. The Markov Chain Monte Carlo (MCMC) method of data augmentation is employed to arrive at the posterior distribution from which the imputed values can be drawn. This imputation process is repeated $M$ times to create independent data sets (Schafer & Olsen, 1998). Each of these data sets is then subjected to the analysis of interest, such as the Item Response Theory (IRT) parameter estimation featured in this paper. The results of the $M$ separate analyses (e.g., parameter estimates) are then combined into a single value as

$$\bar{Q} = \frac{\sum_{m} \hat{Q}_m}{M}. \tag{5}$$

The variance for these estimates is composed of two parts: between imputation variance and within imputation variance. Between imputation variance takes the form

$$B = \frac{\sum_{m} (\hat{Q}_m - \bar{Q})^2}{M - 1}. \tag{6}$$

The within imputation variance, $\bar{U}$, is the mean of estimated variances across the $M$ imputations. The total variance for MI is then calculated as

$$T = \bar{U} + \left(1 + \frac{1}{M}\right) B. \tag{7}$$

Schafer (1997) described an imputation approach specifically for categorical data based on the multinomial distribution. While in theory it is most appropriate for categorical data, he pointed out that for more than just a few variables the saturated log-linear model upon which it is based is severely degraded, making it impractical for use with most real world problems (p. 239). Schafer went on to suggest that using the normal-based approach to MI described above appeared to work well for many categorical data problems. Other researchers have shown that imputing ordinal data using the normal MI model yielded acceptable results when as much as 30% of the data were missing (Leite & Beretvas, 2004; Schafer, Khare, & Ezzati-Rice, 1993). When the normal model is used to impute categorical data, it is recommended that noninteger values be rounded off so that the imputed data conform to the nature of the actual data (i.e., ordinal or dichotomous integers) (Ake, 2005; Schafer, 1997). A preliminary analysis for the current study found that imputation using the multinomial model was not possible for more than 10 items. Therefore, given prior research and this inability to examine larger data sets with the multinomial MI model, the normal-based approach was used in this study.

### *EM Algorithm (EM)*

There are a number of excellent descriptions of the EM algorithm in the context of data imputation (Little & Rubin, 2002; Schafer, 1997), and for that reason the following will be kept very brief. The EM algorithm involves an iterative process in which initial estimates of the missing data values are obtained (typically using regression with a random error term added) and then an estimate of the covariance matrix and set of means is obtained. Next, the newly obtained covariance matrix and mean vector are used in another regression analysis to obtain new estimates for the missing values. These steps are iterated repeatedly until the change in the covariance matrix is minimal, at which point the algorithm stops and the imputed values are ready for use in other analyses. In the context of categorical data analysis, it may be necessary to round these values as they will typically not be integers. The SAS V 9.1 software PROC MI procedure (SAS Institute, 2004) was used in the current study to conduct EM imputation.

### *Nonimputation-Based Approaches*

One nonimputation-based approach for dealing with missing item responses is to treat them as not presented (NP), an option available in BILOGMG (Zimowski et al., 2003). In this case, item responses that were missing for an individual were simply not included in the estimation of item parameters, as if the student had never been given the opportunity to answer the items. A second missing data approach available to the users of BILOGMG is to treat the missing item response as incorrect (IN). In other words, any missing item response is treated as an incorrect response in the scoring of the test and in the estimation of item and person parameter values. The final nonimputation-based method for dealing with missing item responses is to treat them as fractionally correct (FR). For example, when using the 3PL model this would mean that when there are 5 alternatives, the omitted response would be scored fractionally correct with the fraction 1/5 (Zimowski et al., 2003). Each of these approaches was investigated in the current research.

### *Prior Research*

Some of the earliest research regarding missing item responses and item parameter estimation was done by Lord (1974), who demonstrated that omitted items should not be treated as IN when one is interested in accurately estimating either examinee ability or item parameters. He argued instead for a method that provides comparable estimates that one would obtain if an examinee had responded totally randomly to the omitted item. Lord (1983) extended this work to account for the fact that the probability of a correct response given a random response mechanism is an increasing function based on examinee ability. He continued to argue that treating an omitted item response as IN is not the optimal approach, despite its common use in practice.

A number of studies (summarized in Schafer & Graham, 2002 and in Sinharay et al. 2001) have examined the performance of MI with continuous data. Much less work has been done using MI with categorical data (either ordinal or dichotomous). As mentioned above, Schafer (1997) suggested that normal-model-based methods may work well for categorical data provided that imputed values are rounded to integers within the original range of the data (p. 147). Ake (2005) investigated the impact of using MI based on the normal model with ordinal data and found that when imputed values were rounded to an integer value estimates of the mean suffered from increased bias as the percent of MCAR data increased. On the other hand, Ake found that when rounding was not done there was very little bias in the mean estimate. Allison (2006) furthered Ake's work by examining the use of normal based MI with dichotomous data. He found that when values were rounded to the nearest integer (either 0 or 1), the resulting estimate of the mean was biased, but that when no rounding occurred very little if any bias was found. Leite and Beretvas (2004) examined the performance of the normal model MI approach in terms of reproducing correlations among Likert type ordinal data in a simulation study for MCAR, MAR, and MNAR data. Their results show that imputations based on the normal MI model were able to accurately reproduce inter-item correlation values for the Likert data, particularly with lower levels of missing data. This discrepancy in results between Ake and Leite and Beretvas may be due in part to the fact that in the former case, ordinal data were dummy coded and represented as $m - 1$ dichotomous variables, where $m$ was the number of categories.

A number of the methods for imputation described above, including CM and RF, were designed specifically for discrete item response data, unlike the EM or MI approaches (Bernaards & Sijtsma, 1999). Bernaards and Sijtsma examined the ability of several of these categorical imputation techniques to produce complete sets of ordinal data that could provide accurate estimation of factor loadings. They found that the CM technique was superior to random imputation, mean imputation, and listwise deletion. In extending this research, Bernaards and Sijtsma (2000) found that the CM and EM algorithm techniques accurately estimate factor loadings for ordinal data. They stated that the EM approach was the optimal method for Likert data but went on to argue that CM (among other similar approaches) offers a viable, simpler alternative to the more complex methods for imputing categorical missing data. Sijtsma and van der Ark (2003) extended the work of Bernaards and Sijtsma, including the RF method, and examined different outcome variables using dichotomous data. They studied differences in values between complete and imputed data sets for the following statistics: Cronbach's $\alpha$, Mokken's scalability coefficient, and the goodness-of-fit chi-square statistics for the Rasch model. They found that the RF technique provided the most accurate results, particularly for larger sample sizes. Sijtsma and van der Ark called for future research examining the performance of MI with dichotomous data.

Huisman and Molenaar (2001) investigated the performance of a number of imputation methods in the item response context, including the Hot Deck, CM, and model-based IRT methods such as Mokkken scaling and imputation using a 1-parameter logistic (1PL) model. Their results indicated that methods based on the 1PL model provided accurate ability estimation, while the Mokken and CM techniques performed only slightly worse. They pointed out that the latter methods are somewhat easier to carry out than those based on the IRT model. The Hot Deck method was found to be a generally poor performer.

Smits, Mellenbergh, and Vorst (2002) investigated the performance of a number of imputation techniques using a data set of course grades. These methods included unweighted and weighted mean substitution, as well as use of the GPA for nonmissing course grades, correlation substitution, regression and stochastic regression imputation, EM, and MI. While they found that "No single method emerged as really superior" (p. 204), they concluded their paper by suggesting that MI and stochastic regression produced more consistent predictions than others, and for that reason might be viewed as preferable. They also noted that in the two prediction studies they conducted relatively small differences among methods were seen.

In investigating the impact of strategies for dealing with missing data in the context of reliability estimation for Likert data (using Cronbach's alpha), Enders (2004) found that EM was superior to listwise deletion, pairwise deletion, and mean imputation in terms of the bias, root mean square error, and confidence interval coverage for coefficient alpha. This study used simulated samples of 200 subjects with 7 item responses and approximately 11% total missing data. Enders suggested that researchers recognize the impact of missing data on the estimation of reliability and that EM be used for imputing missing item responses.

DeMars (2002) reported on work examining the impact of missing data on the estimation of item difficulty parameters using both joint maximum likelihood estimation (JMLE) and marginal maximum likelihood estimation (MMLE) for both MCAR and MAR data, where simulees at lower ability levels were more likely to leave target items unanswered. She simulated data with two groups, one having lower mean ability than the other. She measured the bias in difficulty estimates for items with missing data and found that when the ability groupings were ignored, greater bias occurred. The degree of bias was generally not very large, however, in all cases between 0.1 and 0.15. The proportion of missing data ranged from 0.39 to 0.69 for the target items.

De Ayala et al. (2001) investigated the impact of missing data, and several approaches for dealing with it, on estimation of the latent trait values for examinees in the 3PL IRT context. They reported that for estimation of the latent trait treating missing data as NP generally led to more accurate results than treating them as IN. In addition, they suggested that another potentially useful strategy for dealing with missing item responses would be to treat them as FR. They reiterated the suggestion of Lord (1974, 1983) not to treat missing values as IN, which seems to result in biased ability estimates.

The goal of the current study was to ascertain the impact of these methods for dealing with missing data on the estimation of item parameters in the IRT context. Of specific interest was the accuracy of estimation for item difficulty, discrimination, and pseudo-chance parameters for 3PL data. The techniques for dealing with missing item responses included IN, FR, NP, or using one of the imputation methods from among the CM, EM, MI, and RF. CM and RF were selected because they were designed specifically for categorical item response data and have demonstrated promising results in prior research. EM and MI have been suggested as the optimal approaches for imputing missing data in many contexts (Schafer & Graham, 2002) and thus were included in the current study.

## Methods

A simulation study was conducted to ascertain the parameter estimation accuracy of standard IRT software when various methods for handling missing data were used. In addition, analysis was conducted on an actual set of testing data containing some missing item responses. All item parameter estimates were provided by BILOGMG, v 3.0 (Zimowski et al., 2003). For each combination of the study conditions described below, 100 replication were conducted. For each replication, 20 items were simulated from a 3-parameter logistic (3PL) model using the IRT-LAB software (Penfield, 2003), with 4 of these being the target items for which data were made missing. The item parameter values for all of the items were taken from Narayanan and Swaminathan (1994). These item parameters were selected because they have been well studied, are considered typical for item response data and thus should not introduce any artifacts into the analyses due to item atypicality. Following is a discussion of the manipulated factors in this study.

TABLE 1
*Item Parameter Values for Target Items*

| Item | Discrimination | Difficulty | Pseudo-chance |
|---|---|---|---|
| 1 | .44 | −.30 | .17 |
| 2 | 1.02 | 1.28 | .22 |
| 3 | .76 | −2.70 | .21 |
| 4 | 1.32 | .57 | .18 |

### *Sample Size*

Two sample size conditions were simulated in this study, 500 and 1000. Previous studies have examined a broad variety of sample sizes, from 100 through 1000 (Bernaards & Sijtsma, 2000; DeMars, 2002; Huisman & Molenaar, 2001). Results from these studies have demonstrated that larger sample sizes were generally associated with better performance for statistical methods making use of imputed data. The sample sizes selected for this study were meant to replicate those from prior research, as well as to reflect real data conditions found in practice with IRT models.

### *Percent Missing Data*

Three conditions for the percent of missing data were simulated: 5%, 15%, and 30%. These values were selected based upon those used in prior research, where percent missing ranged from 1% to 40% (e.g., Ake, 2005; Enders, 2004; Peng & Zhu, 2005; Sitjsma & van der Ark, 2003). These earlier studies found that typically, statistical analyses based on imputed data performed better with less missing data.

### *Target Items*

Four target items were simulated with missing data. These were selected from the 20 items included in the study in order to allow for a variety of difficulty, discrimination, and pseudo-guessing values. The levels of the item parameters in the population appear in Table 1. It is recognized that these are only four possibilities among a truly infinite number of item parameter values; however, we believe that these represent typical values seen in practice (e.g., Narayanan & Swaminathan, 1994), therefore making the results from this study informative to practitioners who might have an interest in the methods of imputation discussed here.

### *Type of Missing Data*

Previous studies have examined the performance of various imputation methods with data that are both MAR and MNAR (e.g., DeMars, 2002; Huisman & Molenaar, 2001; Peng & Zhu, 2005; Sijtsma & van der Ark, 2003). Generally speaking, MNAR data create the greatest difficulties in terms of accurately imputing values that could then be used successfully in other statistical analyses. In this study, two missing data conditions were simulated.

For the first condition, the probability of a missing value was inversely related to the sum of the number of correct items, other than the target. This type of missing data mechanism may be thought of as MAR because the probability of having a missing value is not related directly to the item response, but rather to another variable that can be observed (number of items correct). Borrowing on the methodologies outlined in De Ayala et al. (2001) and Enders (2004) for creating data sets with missing observations, the number of items correct score was calculated for each examinee on all but the target items. The simulees were then divided into four fractiles based on this number correct score (0–3, 4–7, 8–11, 12–16). Members of each fractile were assigned a probability of a missing response, with lower scores having a higher probability of a missing value. The average of these probabilities across the fractiles was equal to the desired proportion missing (i.e., 0.05, 0.15, or 0.30). For each simulee, a random uniform (0, 1) value was generated and compared with the probability of a missing response. In order to ensure that the correct total proportion of missing data was maintained, the results of the aforementioned steps were monitored. For nearly all of the replications, the proportion of actual missing data was within 0.005 of the desired probability. For those few replications where this was not the case, the process of assigning missing values was redone from the beginning and the resulting data sets were not used until the proportion of missing was within 0.005 of the desired probability. Using this approach for generating missing data, simulees with higher number correct score values were less likely to have a missing response than were those with lower such values and the total proportion with missing data was kept at the desired level. This condition was meant to simulate the situation where examinees with less knowledge of the subject being tested were more likely to leave an item unanswered. It will be denoted as MAR in this study.

The second type of missing data generated in this study was MNAR. In this case, the probability of having a missing response was directly related to whether the individual actually got the item correct or not in the initial data generation phase. In order to create a data matrix with missing responses, first a complete item response matrix was generated (i.e., no missing data were present). Each simulee was then assigned a probability of a missing response, with individuals having an incorrect target item response in the initial data generation being assigned a higher probability of that item response being missing. The mean of these probabilities of a missing item response was equal to the overall probability for that specific study condition (i.e., 0.05, 0.15, or 0.30). As a simple example, take a data set with 10 examinees in which, for the initial data generation, 5 have a correct response to the target item and 5 have an incorrect response, and the probability of a missing value is 0.15. In this case, the 5 individuals with an incorrect response will be assigned a probability of a missing value of 0.20 while the 5 individuals with a correct response will be assigned a probability of a missing value of 0.10, so that the average probability missing was 0.15. For each simulee, a uniform random variable (0, 1) was generated and if the resulting value was lower than the assigned probability of a missing value the item response was made missing.

### *Methods of Dealing with Missing Data*

Data were imputed using the EM, RF, CM, and MI methods described above, and the NP, IN, and FR approaches were also used. As a baseline for comparison, parameter estimates for the complete data (CL) were also obtained. The RF and CM methods were carried out using macros written for SPSS, as described in van Ginkel and van der Ark (2005), while MI and EM were conducted with SAS version 9.1 (SAS Institute, 2004). Because the data were all dichotomous in nature and were to be analyzed using BILOGMG to obtain item parameter estimates, imputed responses were rounded to the nearest integer values of 0 or 1. For the MI approach, the number of items responded to correctly was used in imputation for the MAR condition.

The outcomes of interest in this study were the bias and standard errors of item discrimination, difficulty, and pseudo-guessing parameter estimates. Bias for a given parameter, $\theta$, was defined as:

$$\theta - \hat{\theta} \tag{8}$$

where
$\theta =$ actual value of the parameter (i.e., discrimination, difficulty or pseudo-guessing).
$\hat{\theta} =$ estimated value of the parameter.

Analysis of variance (ANOVA) and variance components analysis were used to ascertain which of the manipulated factors, or interactions of these factors, significantly influenced the estimation bias for the discrimination and difficulty parameters.

## Results

### *Item Discrimination*

Results of the ANOVA and variance components analyses indicated that the highest order statistically significant ($\alpha = 0.05$) term accounting for at least 8% of the variation in estimation bias was the interaction between the type of missing data by percent of data missing by imputation method. In addition to this interaction, other significant terms in the model that accounted for at least 10% of the variation in bias were the main effect of imputation method (accounting for 12.5% of bias variation) and the interactions of missing data type by percent of missing data (12.5%) and type of missing data by imputation method (25%). Given the significance of the three-way interaction that subsumes these lower-order effects, it will be the focus of the following discussion.

The bias values for the significant interaction of imputation method by percent and type of missing data appear in Table 2. These results indicate that across all incomplete data conditions, bias in the MNAR condition was greater than that for MAR. Furthermore, estimation bias was greatest in both conditions when the missing item responses were treated as IN. As can be seen in the table, there was generally an increase in estimation bias as the percent of missing data increased (except for the complete data condition, in which no missing data were present). For those cases where the discrimination parameter in the population was less than 1.0, all the methods overestimated the value, while when the parameter was greater than 1.0, they all underestimated it. This result may be due to the fact that in the estimation procedure the prior value of *a* in BILOG is set to 1.0 by default (which was the approach used here), so that the estimated values tend toward this prior, resulting in some overestimation when the parameter value is less than 1.0 and underestimation when it is

TABLE 2
*Bias in Item Discrimination by Method of Imputation, Type of Missing Data, and Percent of Missing Data*

| Type | Percent | CL | IN | FR | NP | RF | EM | CM | MI |
|---|---|---|---|---|---|---|---|---|---|
| Item 1 ($a = 0.44, b = -0.33, c = 0.17$) | | | | | | | | | |
| MAR | 5 | .07 | .23 | .07 | .07 | .07 | .09 | .09 | .08 |
| | 15 | .07 | .47 | .08 | .09 | .09 | .14 | .16 | .08 |
| | 30 | .08 | .84 | .13 | .13 | .11 | .22 | .27 | .09 |
| MNAR | 5 | .06 | .18 | .16 | .14 | .15 | .18 | .18 | .09 |
| | 15 | .08 | .58 | .24 | .23 | .21 | .27 | .28 | .17 |
| | 30 | .06 | .91 | .35 | .31 | .30 | .34 | .40 | .27 |
| Item 2 ($a = 1.02, b = 1.28, c = 0.22$) | | | | | | | | | |
| MAR | 5 | −.03 | −.14 | −.04 | −.04 | −.06 | −.07 | −.12 | −.03 |
| | 15 | −.02 | −.29 | −.06 | −.08 | −.10 | −.12 | −.19 | −.04 |
| | 30 | −.04 | −.48 | −.09 | −.10 | −.17 | −.23 | −.33 | −.05 |
| MNAR | 5 | −.02 | −.20 | −.09 | −.08 | −.11 | −.14 | −.16 | −.12 |
| | 15 | −.04 | −.32 | −.17 | −.16 | −.16 | −.20 | −.23 | −.18 |
| | 30 | −.03 | −.53 | −.32 | −.29 | −.34 | −.35 | −.42 | −.23 |
| Item 3 ($a = 0.76, b = -2.7, c = 0.21$) | | | | | | | | | |
| MAR | 5 | .06 | .15 | .06 | .06 | .08 | .08 | .13 | .07 |
| | 15 | .07 | .36 | .09 | .07 | .10 | .13 | .19 | .08 |
| | 30 | .06 | .79 | .11 | .09 | .12 | .18 | .24 | .10 |
| MNAR | 5 | .07 | .22 | .19 | .19 | .19 | .21 | .20 | .16 |
| | 15 | .07 | .49 | .24 | .24 | .25 | .28 | .34 | .22 |
| | 30 | .07 | .90 | .29 | .28 | .29 | .37 | .44 | .29 |
| Item 4 ($a = 1.32, b = 0.57, c = 0.18$) | | | | | | | | | |
| MAR | 5 | −.02 | −.09 | −.06 | −.05 | −.04 | −.06 | −.07 | −.04 |
| | 15 | −.02 | −.22 | −.09 | −.09 | −.07 | −.13 | −.15 | −.07 |
| | 30 | −.02 | −.52 | −.12 | −.15 | −.16 | −.25 | −.25 | −.09 |
| MNAR | 5 | −.01 | −.27 | −.13 | −.09 | −.13 | −.17 | −.19 | −.13 |
| | 15 | −.03 | −.54 | −.20 | −.19 | −.18 | −.26 | −.27 | −.17 |
| | 30 | −.04 | −.70 | −.41 | −.38 | −.34 | −.44 | −.39 | −.37 |

*Note.* CL = complete data, IN = incorrect, FR = fractionally correct, NP = not presented, RF = response function, EM = EM algorithm, CM = corrected mean, MI = multiple imputation.

greater than 1.0. Indeed, this result is also apparent in the CL case for all four items. Given these results, users of BILOG-MG with missing data may consider providing the software with prior values for the discrimination parameter (an available option), such as those based on the biserial correlation between the item response and the number of items correct.

As stated above, treating the missing item responses as IN resulted in the greatest bias across type of missing data and percentage missing. The next most biased approaches to dealing with the missing item responses in general appear to be EM and CM. The FR, NP, RF, and MI approaches resulted in relatively less bias, with MI performing somewhat better overall particularly in the MAR condition. Indeed, for the target items (particularly 1, 2, and 3), the discrimination bias found in the MI condition was very comparable to that for the complete data condition for MAR data where the MI used the number of items correct in the imputation process. Item discrimination bias values for the imputation methods by sample size were not significantly different from one another across the methods, nor was there a significant interaction between sample size and method. For this reason, these results are discussed no further here.

Based on results in Table 3, the standard errors for the discrimination parameter estimates were fairly comparable across the methods used for handling the missing data. There was a slight increase in the standard errors of all of the methods, except IN, when the percent of missing data increased. Finally, the standard errors were slightly lower in the MNAR than the MAR condition for all of the methods as well.

### *Item Difficulty*

As with item discrimination bias, ANOVA and variance components analysis were used to identify significant main effects and interactions from among the manipulated variables and to determine the amount of variation in bias accounted for by each. The highest-order significant interaction was method of imputation by type of missing data, which accounted for between 16% and 22% of variation in bias across the four target items. In addition, the main effects of method and type of missing data were also statistically significant and accounted for more than 15% each of variation in bias each. No other term was both statistically significant and accounted for more than 10% of the variance in bias for any of the target items. Table 4 displays the bias results for the combination of method of imputation by type of missing data.

For all of the methods studied here, with the exception of IN, there was a small underestimation in item difficulty in the MAR condition, with MI exhibiting the least such bias. In general, the magnitude of bias in the MAR case is similar to that reported by DeMars (2002). On the other hand, when the data were MNAR, all the methods except IN produced more negatively biased estimates of item difficulty (i.e., they indicated that the item was easier than it actually was). On the other hand, for IN the difficulty estimates were overestimated (i.e., items were estimated to be more difficult than they actually were) regardless of whether the missing data mechanism was MAR or MNAR, with bias being more pronounced in the MNAR case. Among the other methods, MI typically had among the lowest (if not the lowest) estimates of bias, while the FR, NP, RF, and CM methods were all very comparable. The bias results by the sample size were not statistically significant for any of the four items examined here. In addition, the percent of missing data accounted for less than 2% in all cases, as did the interaction of percent missing data and type of missing data. As with item discrimination, this lack of significant results means that these terms are discussed no further.

The standard error of the item difficulty estimates by the interaction of imputation method and type of missing data appear in Table 5. In general, there was less variation in the estimates under the MNAR condition than the MAR for all methods for dealing with missing data. Treating the missing responses as IN resulted in the least

TABLE 3

*Standard Error of Item Discrimination by Method of Imputation, Type of Missing Data and Percent of Missing Data*

| Type | Percent | CL | IN | FR | NP | RF | EM | CM | MI |
|---|---|---|---|---|---|---|---|---|---|
| Item 1 ($a = 0.44, b = -0.33, c = 0.17$) | | | | | | | | | |
| MAR | 5 | .07 | .06 | .08 | .07 | .09 | .10 | .09 | .08 |
| | 15 | .05 | .06 | .09 | .08 | .11 | .11 | .08 | .09 |
| | 30 | .07 | .05 | .12 | .10 | .12 | .12 | .11 | .08 |
| MNAR | 5 | .06 | .06 | .08 | .08 | .06 | .09 | .06 | .04 |
| | 15 | .06 | .07 | .10 | .11 | .09 | .10 | .08 | .06 |
| | 30 | .06 | .06 | .13 | .13 | .11 | .10 | .09 | .07 |
| Item 2 ($a = 1.02, b = 1.28, c = 0.22$) | | | | | | | | | |
| MAR | 5 | .10 | .09 | .11 | .11 | .13 | .10 | .09 | .07 |
| | 15 | .10 | .08 | .12 | .13 | .15 | .10 | .08 | .10 |
| | 30 | .09 | .07 | .16 | .15 | .18 | .13 | .11 | .12 |
| MNAR | 5 | .07 | .05 | .07 | .08 | .10 | .08 | .07 | .06 |
| | 15 | .07 | .06 | .08 | .08 | .13 | .09 | .07 | .08 |
| | 30 | .08 | .06 | .11 | .10 | .15 | .11 | .09 | .09 |
| Item 3 ($a = 0.76, b = -2.7, c = 0.21$) | | | | | | | | | |
| MAR | 5 | .11 | .12 | .12 | .10 | .12 | .11 | .12 | .08 |
| | 15 | .11 | .13 | .13 | .11 | .13 | .11 | .14 | .10 |
| | 30 | .11 | .13 | .14 | .13 | .14 | .12 | .17 | .14 |
| MNAR | 5 | .10 | .09 | .08 | .08 | .09 | .07 | .11 | .06 |
| | 15 | .13 | .10 | .09 | .10 | .09 | .09 | .12 | .07 |
| | 30 | .11 | .10 | .10 | .13 | .10 | .10 | .14 | .09 |
| Item 4 ($a = 1.32, b = 0.57, c = 0.18$) | | | | | | | | | |
| MAR | 5 | .12 | .11 | .12 | .13 | .13 | .12 | .12 | .11 |
| | 15 | .11 | .12 | .13 | .14 | .14 | .14 | .12 | .10 |
| | 30 | .12 | .13 | .15 | .17 | .15 | .19 | .16 | .12 |
| MNAR | 5 | .13 | .12 | .10 | .11 | .12 | .10 | .10 | .08 |
| | 15 | .14 | .13 | .11 | .13 | .13 | .12 | .11 | .09 |
| | 30 | .13 | .12 | .12 | .14 | .12 | .13 | .12 | .11 |

*Note.* CL = complete data, IN = incorrect, FR = fractionally correct, NP = not presented, RF = response function, EM = EM algorithm, CM = corrected mean, MI = multiple imputation.

amount of variation, a point that will be discussed further below. There was relatively little difference among the standard errors across the other methods, though item 3, which had the lowest difficulty parameter value, was associated with markedly greater variation for all of the techniques than were the other items, including for the CL estimates.

### *Pseudo-Chance Parameter*

The results of the ANOVA and variance components analyses for the pseudo-chance parameter value showed that for each of the target items, the interaction of type of missing data and method for dealing with missing data was the highest-order

TABLE 4
*Bias of Item Difficulty by Method of Imputation and Type of Missing Data*

| Type | CL | IN | FR | NP | RF | EM | CM | MI |
|---|---|---|---|---|---|---|---|---|
| Item 1 ($a = 0.44, b = -0.33, c = 0.17$) | | | | | | | | |
| MAR | .02 | .37 | −.10 | −.06 | −.06 | −.16 | −.08 | −.06 |
| MNAR | .01 | .96 | −.49 | −.54 | −.56 | −.61 | −.54 | −.51 |
| Item 2 ($a = 1.02, b = 1.28, c = 0.22$) | | | | | | | | |
| MAR | −.01 | .13 | −.02 | −.04 | −.03 | −.06 | −.03 | −.02 |
| MNAR | .01 | .44 | −.39 | −.37 | −.34 | −.42 | −.30 | −.30 |
| Item 3 ($a = 0.76, b = -2.7, c = 0.21$) | | | | | | | | |
| MAR | −.02 | .91 | −.09 | −.10 | −.10 | −.17 | −.12 | −.06 |
| MNAR | −.02 | 1.54 | −.19 | −.17 | −.20 | −.26 | −.24 | −.14 |
| Item 4 ($a = 1.32, b = 0.57, c = 0.18$) | | | | | | | | |
| MAR | .01 | .17 | −.06 | −.08 | −.06 | −.11 | −.05 | −.03 |
| MNAR | .01 | .60 | −.37 | −.38 | −.38 | −.42 | −.36 | −.31 |

*Note.* CL = complete data, IN = incorrect, FR = fractionally correct, NP = not presented, RF = response function, EM = EM algorithm, CM = corrected mean, MI = multiple imputation.

TABLE 5
*Standard Error of Item Difficulty by Method of Imputation and Type of Missing Data*

| Type | CL | IN | FR | NP | RF | EM | CM | MI |
|---|---|---|---|---|---|---|---|---|
| Item 1 ($a = 0.44, b = -0.33, c = 0.17$) | | | | | | | | |
| MAR | .20 | .12 | .21 | .21 | .26 | .26 | .18 | .20 |
| MNAR | .20 | .04 | .11 | .09 | .14 | .15 | .12 | .11 |
| Item 2 ($a = 1.02, b = 1.28, c = 0.22$) | | | | | | | | |
| MAR | .15 | .13 | .17 | .16 | .17 | .15 | .15 | .14 |
| MNAR | .16 | .05 | .10 | .06 | .13 | .07 | .13 | .08 |
| Item 3 ($a = 0.76, b = -2.7, c = 0.21$) | | | | | | | | |
| MAR | .25 | .15 | .31 | .34 | .33 | .37 | .28 | .30 |
| MNAR | .27 | .06 | .18 | .16 | .18 | .20 | .13 | .19 |
| Item 4 ($a = 1.32, b = 0.57, c = 0.18$) | | | | | | | | |
| MAR | .21 | .11 | .19 | .20 | .22 | .23 | .20 | .18 |
| MNAR | .22 | .09 | .11 | .11 | .15 | .17 | .10 | .12 |

*Note.* CL = complete data, IN = incorrect, FR = fractionally correct, NP = not presented, RF = response function, EM = EM algorithm, CM = corrected mean, MI = multiple imputation.

significant term, and accounted for more than 75% of the variance in bias. None of the other terms were both statistically significant and accounted for more than 5% of the observed variation in estimation bias of the pseudo-chance parameter.

The bias in the pseudo-guessing parameter estimates appears in Table 6. For all methods except IN, the bias in the $c$ parameter was very close to that for CL. In addition, there was little difference from MAR to MNAR. In general, the methods

TABLE 6

*Bias of Pseudo-Chance Parameter Estimates by Method of Imputation and Type of Missing Data*

| Type | CL | IN | FR | NP | RF | EM | CM | MI |
|---|---|---|---|---|---|---|---|---|
| | | Item 1 ($a = 0.44, b = -0.33, c = 0.17$) | | | | | | |
| MAR | .01 | −.05 | .03 | .03 | .04 | .01 | .02 | .02 |
| MNAR | .02 | −.02 | .03 | .03 | .04 | .02 | .03 | .03 |
| | | Item 2 ($a = 1.02, b = 1.28, c = 0.22$) | | | | | | |
| MAR | .01 | −.08 | .01 | .02 | .01 | .04 | .01 | .01 |
| MNAR | .01 | −.05 | .01 | .02 | .03 | .03 | .02 | .02 |
| | | Item 3 ($a = 0.76, b = -2.7, c = 0.21$) | | | | | | |
| MAR | .01 | −.03 | .02 | .02 | .02 | .03 | .02 | .01 |
| MNAR | .01 | −.03 | .03 | .02 | .03 | .03 | .03 | .02 |
| | | Item 4 ($a = 1.32, b = 0.57, c = 0.18$) | | | | | | |
| MAR | .01 | −.07 | .01 | .01 | .01 | .03 | .01 | .01 |
| MNAR | .01 | −.05 | .02 | .01 | .02 | .04 | .01 | .01 |

*Note.* CL = complete data, IN = incorrect, FR = fractionally correct, NP = not presented, RF = response function, EM = EM algorithm, CM = corrected mean, MI = multiple imputation.

examined here (again, except for IN) were associated with a slight positive bias in the pseudo-chance parameter. When the missing response was treated as IN, the resulting pseudo-guessing values were underestimates of the actual value.

As can be seen in Table 7, the standard errors of the pseudo-guessing parameter estimates were generally very low, and very comparable across the methods of dealing with the missing data and the four items. As with the difficulty parameter estimates, the variation for the *c* estimates was lower in the MNAR condition, though only slightly so. There were not marked differences in the standard errors among the methods examined here.

## Discussion

The results of this study appear to support the assertion by Lord (1974) that treating missing item responses as IN is not optimal. Indeed, for the estimation of both item discrimination and difficulty, the IN approach was associated with much greater bias than were any of the other methods investigated here. Using the FR technique, with a fraction of 1/5, resulted in much lower estimation bias than did IN, similar to results reported by De Ayala et al. (2001) for estimating latent abilities. Among the imputation methods examined here, it appears that there was not necessarily a superior approach, though MI was often associated with slightly lower estimation bias than the other techniques. This outcome is in keeping with Smits et al. (2002), who found that when imputing missing grades, "No single method emerged as really superior" (p. 204), though they did recommend the use of data augmentation, the type of MI used in this study. At the same time, they found that the methods did not differ substantially in terms of their imputed values, much as was the case here.

TABLE 7

*Standard Error of Pseudo-Chance Parameter by Method of Imputation and Type of Missing Data*

| Type | CL | IN | FR | NP | RF | EM | CM | MI |
|---|---|---|---|---|---|---|---|---|
| | | | Item 1 ($a = 0.44, b = -0.33, c = 0.17$) | | | | | |
| MAR | .04 | .02 | .03 | .03 | .04 | .03 | .03 | .03 |
| MNAR | .03 | .01 | .02 | .01 | .02 | .01 | .01 | .01 |
| | | | Item 2 ($a = 1.02, b = 1.28, c = 0.22$) | | | | | |
| MAR | .03 | .02 | .03 | .03 | .04 | .03 | .03 | .03 |
| MNAR | .03 | .01 | .01 | .01 | .04 | .02 | .03 | .02 |
| | | | Item 3 ($a = 0.76, b = -2.7, c = 0.21$) | | | | | |
| MAR | .01 | .02 | .01 | .01 | .01 | .02 | .01 | .01 |
| MNAR | .01 | .01 | .01 | .01 | .01 | .01 | .01 | .01 |
| | | | Item 4 ($a = 1.32, b = 0.57, c = 0.18$) | | | | | |
| MAR | .04 | .03 | .03 | .03 | .05 | .04 | .04 | .03 |
| MNAR | .04 | .01 | .01 | .01 | .03 | .01 | .03 | .01 |

*Note*. CL = complete data, IN = incorrect, FR = fractionally correct, NP = not presented, RF = response function, EM = EM algorithm, CM = corrected mean, MI = multiple imputation.

With respect to MI, it was not always associated with unbiased estimates, which is consistent with prior work that found when it is used with categorical variables and rounding of imputed values, as was the case here, some estimation bias is present (Ake, 2005; Allison, 2006).

In terms of the types of missing data, greater bias was clearly associated with item responses that were MNAR. It is important to keep in mind that in the MNAR condition data were more likely to be missing if they were incorrect in the original data generation phase of the study. Item difficulty values were uniformly underestimated in the MNAR case, except when missing data were treated as IN. This latter result would appear to be due to the fact that every missing response was treated as IN even if it was actually correct in the complete data matrix (before some items were simulated as missing), so that when the difficulty parameter was estimated more examinees were counted as having answered the item incorrectly than was the case in actuality, leading to the overestimation of item difficulty. This outcome was most notable for the easier items (1 and 3), though it was present in all cases. On the other hand, all of the other approaches for dealing with missing data were associated with an underestimation of item difficulty in the MNAR case.

In order to investigate this result somewhat more deeply, selected individual replications from a few of the MNAR data conditions were examined and the proportion of cases correct (and incorrect) for the target items were calculated under each of the missing data methods examined here. In every case, the proportion of correct responses was higher for the missing data methods (except IN) than for the original data set where no missing data existed. Furthermore, in all cases, MI had the closest percent correct to that for the complete data set while EM was the furthest away from the complete data case (and always higher). In the MNAR condition, EM was also typically associated with the largest underestimation of item difficulty of the methods studied here.

Because simulees were more likely to be assigned missing item responses if they had an incorrect response in the original complete data set, the resulting data matrices with missing values tended to have a higher percent of correct nonmissing values than was true for the original data sets. For example, in one of the selected replications, the proportion of correct responses to item 1 was 0.656 in the original data matrix. After the missing data were simulated for this data set, the proportion of nonmissing values with correct responses was 0.758. This pattern was very similar throughout the individual replications examined. The various imputation methods use the nonmissing observations in some way to impute for the missing responses, meaning that these imputations were based on an inflated proportion of correct responses in the MNAR case, which appears to have resulted in the underestimation of item difficulty. A similar pattern is evident for the MAR data, though the degree of bias was much less severe. In this case, individuals at lower ability levels were more likely to have missing data values than those at higher abilities, so that again, the resulting data matrices (after missingness was simulated) tended to have a higher percent of correct responses than did the complete data sets, which seems to have led to an underestimation of item difficulty. However, this outcome was not as marked as when the missing values were directly tied to the item responses as in MNAR.

With respect to the item discrimination values, all of the methods studied here were associated with overestimation for items where *a* in the population was less than 1.0 and underestimation where *a* was greater than 1.0, including for the complete data sets. As discussed briefly above, the default prior value used by BILOGMG in estimating item discrimination is 1.0, so that the estimated values when the actual *a* was less than 1.0 tended upward toward this prior, while for cases where *a* is greater than 1.0, they tended downward. All of the techniques studied here produced results following this pattern, and when the percent of missing data was low, the levels of bias for several of the methods were similar to that for the complete data. Among the various approaches, MI was associated with results most similar to those seen with the complete data, while IN was associated with the greatest bias for both MAR and MNAR data. Such a result would not be surprising in the MAR case, given that MI used the number of items correct score (which was directly associated with missing item responses) in the imputation procedure, unlike the other approaches.

There was relatively little difference among the pseudo-guessing parameter estimates associated with the methods studied here. As described above, IN had greater bias, and the bias was negative, whereas the other approaches all had slight positive bias. The relative lack of difference in performance across items might have been due to the similarity in their *c* values in the population.

While the results of this study do not provide definitive evidence that one method for dealing with missing item responses is superior to the others, they do suggest that some approaches may not be optimal. Specifically, treating missing item responses as IN would seem to result in greater bias for both difficulty and discrimination parameter estimates. In addition, EM was associated with somewhat greater bias than was true for other imputation methods such as MI and RF. This result may be due in large part to the fact that the EM approach relies on an assumption of multivariate normality that clearly does not apply to dichotomous item responses. While this assumption is also made for the MI model used here, Schafer (1997) has shown that the normal-based approach often works for categorical data, and the current results appear to bear this out. In addition, MI was able to make use of the number correct score, which EM was not, to provide more information to the imputation process in the MAR condition and that appeared to be helpful. Prior simulation-based research in which the EM algorithm has been used to impute missing values for categorical data were either done with polytomous responses (5 levels) based on a normally distributed continuum and with a symmetric ordinal distribution (Enders, 2004) or with 5 level polytomous data where the imputed values were not made to conform to the ordinal form of the original data (Bernaards & Sitjsma, 1999). In either case, the type of data was very different from the dichotomous 3PL model item responses that were used in this study.

In considering which of these approaches to missing item responses one should use, it is important to note that while MI had slightly lower bias results across most conditions studied here, both FR and NP had only slightly higher bias values and comparable standard errors. In addition, because these methods are a part of the item parameter estimation software itself, they do not require any additional analyses, unlike MI. Of the two methods designed specifically for categorical data (RF and CM), RF had bias results comparable to both FR and NP, but somewhat higher standard errors in many (though not all) conditions, while CM was associated with more bias in item discrimination values and generally comparable bias for difficulty estimation. For this reason, neither of these techniques would seem to be superior to those available in the BILOGMG software. In sum, while these results do not seem to lift one method up as superior to all others, they do support the use of FR, NP, and MI, if the researcher is interested in limiting item parameter estimation bias due to missing responses.

## Note

[1] Note that other such models are possible but are beyond the scope of this article.

## References

Allison, P. D. (2006, March). Imputation of categorical variables with PROC MI. Paper presented at the annual meeting of the SAS Users Group International, San Francisco, CA.

Ake, C. F. (2005, April). Rounding after multiple imputation with non-binary categorical covariates. Paper presented at the annual meeting of the SAS Users Group International, Philadelphia, PA.

Bernaards, C. A., & Sijtsma, K. (1999). Factor analysis of multidimensional polytomous item response data from ignorable item nonresponse. *Multivariate Behavioral Research, 34*, 277–314.

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

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*, 213–234.

DeMars, C. (2002). Incomplete data and item parameter estimates under JMLE and MML estimation. *Applied Measurement in Education, 15*, 15–31.

Enders, C. K. (2004). The impact of missing data on sample reliability estimates: Implications for reliability reporting practices. *Educational and Psychological Measurement, 64*, 419–436.

Huisman, M., & Molenaar, I. W. (2001). Imputation of missing scale data with item response models. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 221–244). New York: Springer.

Leite, W. L., & Beretvas, S. N. (2004, April). The performance of multiple imputation for Likert-type items with missing data. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Hoboken, NJ: John Wiley and Sons, Inc.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39*, 247–264.

Lord, F. M. (1983). Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika, 48*, 477–482.

Madow, W. G., Nisselson, H., & Olkin, I. (Eds.) (1983). *Incomplete data in sample surveys, Volume 1: Report and case studies*. New York: Academic Press.

Mislevy, R. J., & Wu, P.-K. (1988). *Inferring examinee ability when some item responses are missing* (ERIC Document Reproduction Service No. ED 395 017). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Wu, P.-K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (Research Report RR-96-30-ONR). Princeton, NJ: Educational Testing Service.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*, 315–328.

Penfield, R. D. (2003). IRT-Lab: Software for research and pedagogy in item response theory. *Applied Psychological Measurement, 27*, 301–302.

Peng, C.-Y. J., & Zhu, J. (2005, April). Comparison of two methods for handling missing covariates in logistic regression. Paper presented at the annual meeting of the American Educational Research Association, Montreal, PQ.

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

SAS Institute (2004) *SAS Stat*. Cary, NC: SAS Institute, Inc.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Boca Raton, FL: Chapman and Hall/CRC.

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

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33*, 545–571.

Schafer, J. L., Khare, M., & Ezzati-Rice, T. (1993). Multiple imputation of missing data in NHANES III. Proceedings of the 1993 Annual Research Conference. Washington, DC: U. S. Bureau of the Census.

Sijtsma, K., & Van Der Ark, L. A. (2003). Investigation and treatment of missing item scores in test and questionnaire data. *Multivariate Behavioral Research, 38*, 505–528.

Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods*, *6*, 317–329.

Smits, N., Mellenbergh, G. J., & Vorst, H. C. M. (2002). Alternative missing data techniques to grade point average: Imputing unavailable grades. *Journal of Educational Measurement*, *39*, 187–206.

van Ginkel, J. R., & van der Ark, L. A. (2005). SPSS syntax for missing value imputation in test and questionnaire data. *Applied Psychological Measurement*, *29*, 152–153.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BIOLOGMG3* [Computer Software]. Lincolnwood, IL: Scientific Software International, Inc.

## Author

HOLMES FINCH is an Associate Professor in the Department of Educational Psychology at Ball State University, Muncie, IN 47306; whfinch@bsu.edu. His research interests include item response theory, dimensionality assessment, and DIF.