Modifying the M_2 Statistic to Handle Missing Data

Jeffrey C. Hoover[1] and W. Jake Thompson[1]

[1] University of Kansas

Author Note

Jeffrey C. Hoover ORCID: 0000-0002-0276-0308

W. Jake Thompson ORCID: 0000-0001-7339-0300

We have no known conflict of interest to disclose.

Correspondence concerning this article should be addressed to Jeffrey C. Hoover, ATLAS, University of Kansas, 1122 W. Campus Road, Lawrence, KS, 66045. Email: jhoover4@ku.edu

Abstract

To date, no research has addressed the issue of missing data when calculating the M_2 statistic. Given the ubiquity of missing data, modifying the M_2 statistic to account for missing data is needed. This simulation study evaluated the Type I error rates and statistical power of a modified M_2 statistic (M_2^*). Additionally, we examined the parameter recovery of the estimated models to contextualize the findings. The Type I error rates and statistical power for the M_2^* statistic were elevated in the presence of missing data. In contrast, the Type I error rates and statistical power of the models filtering out examinees with missing responses and recoding missing responses as incorrect were controlled. However, the generating parameters were not recovered as well for the filtered and recoded models as they were with the missing data models.

Modifying the M_2 Statistic to Handle Missing Data

Diagnostic classification models (DCMs) are latent trait models that categorically estimate test takers' proficiency in underlying latent traits (Rupp et al., 2010; Rupp & Templin, 2008). In DCMs, a Q-matrix links items to the assessed latent traits, where a value of one indicates the latent trait is being assessed and a value of zero indicates the latent trait is not being assessed (Tatsuoka, 1983). Items on DCM assessments can assess multiple underlying latent traits (Rupp & Templin, 2008), although the accuracy of DCM parameter estimations is influenced by the number of items measuring each underlying latent trait in isolation (i.e., items that only measure a single attribute; Madison & Bradshaw, 2015).

Model Fit

Model fit is used to provide empirical support for inferences made from the estimated DCM (Chen et al., 2013). Poor model fit typically degrades the accuracy of inferences made from the model (e.g., Ames & Penfield, 2015), because poor model fit indicates the model does not reflect the observed data well. Thus, adequate model fit is often a baseline requirement for supporting inferences made from the model.

Conceptually, model fit is defined as how well the model predicted values match the obtained data (Gu, 2011). Model fit has multiple subtypes (e.g., absolute model fit, relative model fit) that can be calculated at multiple levels (e.g., test-level, item-level, person-level; Han & Johnson, 2019). Absolute model fit assesses whether the parameters estimated by the model generally represent the observed data well, while relative model fit allows for the comparison of multiple models to determine which model fits the data better (Chen et al., 2013). Test-level model fit examines the consistency between model predicted and observed test scores (Gu, 2011; Sinharay & Almond, 2007); item-level model fit examines the consistency between model predicted and observed item responses for each item (Sinharay & Almond, 2007; Sorrel et al., 2017); and person-level model fit examines the consistency between model predicted and observed item response patterns for each test taker (Liu et al., 2009). This manuscript will focus on absolute test-level model fit. Test-level model fit is critically important to assessments, since inferences made from the model are dependent on the model representing the data well. For example, in DCMs, results are often reported as the dichotomous mastery status of each assessed skill, and the confidence in the accuracy of the reported skill mastery is contingent upon adequate test-level model fit.

Rupp et al. (2010) described goodness-of-fit statistics, resampling approaches, posterior predictive model checking (PPMC), and limited information fit statistics as the existing methods for evaluating model fit in DCMs. Goodness-of-fit statistics (e.g., G, χ^2) are perhaps the most common methods for assessing absolute test-level model fit, although these statistics are often problematic in DCMs because of sparsely filled contingency tables. The resampling approaches and PPMC are often problematic because they are time and computationally intensive. Rupp et al. (2010) described limited information fit statistics as promising methods for evaluating model fit in DCMs.

Limited Information Fit Statistics

Building on the work of Bartholomew and Leung (2002), Maydeu-Olivares and Joe (2005) introduced a family of limited information fit statistics, M_r, which use marginal proportions up to order *r* for multivariate binomial contingency tables. Maydeu-Olivares and Joe (2006) extended the M_r family of limited information fit statistics for application to multivariate multinomial contingency tables. To maximize power as well as to minimize computational issues, Maydeu-Olivares and Joe (2005, 2006) recommended using M_2 for assessing model fit with limited information fit statistics. Subsequent work further extended the applicability of the M_r family of limited information fit statistics by showing limited information statistics are as or more powerful than full information fit statistics even in situations where the contingency table is not sparse (Joe & Maydeu-Olivares, 2010). Maydeu-Olivares and Joe (2014) also developed a method for calculating the Root Mean Square Error of Approximation based on M_2 rather than χ^2.

The initial work on the M_2 statistic by Maydeu-Olivares and Joe has been applied to DCMs in many settings (e.g., Chen et al., 2018; Hansen et al., 2014; Jurich, 2014; Liu et al., 2016; Ma, 2019). Marginal proportions are easily calculated within DCMs, which facilitates the application of the M_2 statistic to DCMs. Further, software applications have also allowed for estimating the M_2 statistic quickly (Ma & de la Torre, 2020b), without the computationally intensive efforts required for assessing model fit using resampling techniques or PPMC (Rupp et al., 2010). However, applications of the M_2 statistic are limited by missing data.

Missing Data

Missing data is becoming increasingly examined in DCMs (e.g., Pan & Zhan, 2020; Shan & Wang, 2020; Sünbül, 2018), which is unsurprising since missing data is a common occurrence (Pan & Zhan, 2020) and even expected in some operational settings (e.g., Dynamic Learning Maps Consortium, 2020). Ultimately, the mechanism for generating the missing data is the impetus for the research examining missing data. For data missing completely at random or missing at random (Little & Rubin, 2020; Rubin, 1976), the parameters in DCMs can be estimated accurately with only the observed data (Rupp et al., 2010; Shan & Wang, 2020). For data missing not at random (Little & Rubin, 2020; Rubin, 1976), other procedures are needed to better estimate the parameters in DCMs (Ma et al., 2020). Regardless of the specific mechanism generating the missing data, the methods for estimating DCMs are increasingly acknowledging and adjusting for missing data when estimating model parameters and making mastery classifications.

Despite the growing number of applications of the M_2 statistic to DCMs, no research has addressed the issue of missing data when calculating the M_2 statistic. When calculating the M_2 statistic, the data are required to be full rank, but there is no mention of missing data (Joe & Maydeu-Olivares, 2010; Maydeu-Olivares & Garcia-Forero, 2010; Maydeu-Olivares & Joe, 2005, 2006, 2008, 2014). In the applications of the M_2 statistic to DCMs, researchers have focused on applying the M_2 statistic to correctly specified DCMs and misspecified DCMs, but missing data has not yet been included as a manipulated factor in the simulation studies (Chen et al., 2018; Hansen et al., 2014; Jurich, 2014; Liu et al., 2016; Ma, 2019). Given that missing data is almost unavoidable in real data analyses (Pan & Zhan, 2020), we assume methods such as coding missing data as incorrect responses or using listwise deletion are needed to avoid issues stemming from the presence of missing data in calculating the M_2 statistic, although it is possible that imputation methods could be used to address missing data (e.g., Sünbül, 2018).

Likely stemming from the relatively limited research on calculating the M_2 statistic in the presence of missing data, most existing software applications for estimating the M_2 statistic in DCMs are not capable of handling missing data (e.g., Ma & de la Torre, 2020a). Consequently, approaches such as coding missing data as incorrect responses, using listwise to remove examinees with missing data, or imputation are often required so that the M_2 statistic can be calculated. The marginal proportions may be skewed downward when coding missing data as incorrect responses, and valuable data is discarded when using listwise deletion. Either of these approaches to handling missing data may introduce bias in the M_2 statistic. Thus, there is a need to examine whether the M_2 statistic for DCMs can be modified to better handle missing data. The modified M_2 statistic will subsequently be referred to as M_2^*.

M_2 Limited Information Statistic for Missing Data

As presented in Liu et al. (2016), the M_2 statistic for DCMs is calculated using

$$M_2 = N\left(\boldsymbol{p}_2 - \hat{\boldsymbol{\pi}}_2\right)^{'} \hat{C}_2\left(\boldsymbol{p}_2 - \hat{\pi}_2\right), \tag{1}$$

where N is the number of examinees, $\boldsymbol{p}_2$ is the vector of observed marginal probabilities of correctly responding to each item or pair of items, $\hat{\boldsymbol{\pi}}_2$ is the vector of model predicted marginal probabilities of correctly responding to each item or pair of items, and $\hat{C}_2$ is the orthogonal complement to the Jacobian matrix up to the second order.

To better account for missing data, the $\boldsymbol{p}_2$ vector and N from Equation 1 must be modified. $\boldsymbol{p}_2$ is a vector composed of the first and second order marginal probabilities of correctly responding to each item and pair of items, respectively. While a single process could increase the efficiency of calculating (and modifying) the first and second order marginal probabilities, the process for first and second order marginal probabilities is presented separately to provide increased clarity of the modifications.

For the first order marginal probabilities, the proportion of correct responses is taken for each item such that missing responses are removed using casewise deletion. For example, consider the following five dichotomously scored responses to an item where 'NA' indicates a missing response: 1, 0, 1, 1, NA. The original method for calculating the M_2 statistic was not defined to accept missing responses, meaning the original method could not use the previous response pattern. Because of this, the original method would require filtering out this examinee, recoding the missing responses as incorrect, or recoding the missing responses using a form of imputation. In contrast, the proposed method would generate a marginal probability of .75 for this item (i.e., three of the four provided responses were correct), without requiring any of those approaches to replace the missing responses.

For the second order marginal probabilities, a three-step procedure is used to better account for missing data. As with the first order marginal probabilities, the original method for calculating the second order marginal probabilities for the M_2 statistic are not defined to accept missing data, so the modified procedure allows for incorporating missing data without using an approach to replace the missing responses. In step one of modifying the procedure for calculating the second order marginal probabilities, the number of examinees with responses to both items across all pairwise comparisons are summed. An example of the matrix resulting from step one using responses to four items from 1,000 randomly simulated examinees with approximately 3% of responses missing is presented in Table 1. The data presented in Table 1 can be read as indicating there were 955 examinees who responded to the first item, there were 920 examinees who responded to the first and second item, and so on. Additionally, each entry on the main diagonal must be greater than or equal to the other entries in that row and column because the number of examinees responding to two items (e.g., Item 1 and Item 2) is a subset of the number of examinees responding to either of the items individually (e.g., Item 1 or Item 2).

Table 1

The Matrix of Examinees Completing Both of the Items Across All Pairwise Comparisons.

Item	1	2	3	4
1	955			
2	920	963		
3	924	931	968	
4	918	922	926	958

In step two, missing data are scored as incorrect responses for calculating the crossproduct. This allows for the data to be utilized as if there were no missing data. As will be discussed in step three, coding missing data as incorrect allows for the crossproduct to be calculated without being biased by the missing data, because the results from step one allow us to adjust for missing data.

In step three, the crossproduct of the data with missing data scored as incorrect responses is taken, and the matrix resulting from the crossproduct operation is divided by the matrix from step one. As mentioned to in step two, coding the missing responses as incorrect responses (i.e., 0) does not affect the calculation of the crossproduct because the crossproduct is the number of examinees who correctly responded to both items in each pairwise comparison, which will be unaffected by coding examinees with missing data as responding incorrectly. Further, dividing by the matrix resulting from step one should allow for a more accurate estimation of the second order marginal probabilities than simply coding missing responses as incorrect responses or using listwise deletion to remove examinees with missing responses. Using the same randomly simulated data to four items from 1,000 randomly simulated examinees with approximately 3% of responses missing, the matrix resulting from the crossproduct operation is presented in Table 2, and the quotient of the crossproduct matrix (i.e., Table 2) divided by the matrix from step one (i.e., Table 1) is presented in Table 3.

Table 2

The Matrix of Examinees Responding Correctly to Both Items Across All Possible Pairwise Comparisons.

Item	**1**	**2**	**3**	**4**
1	579			
2	385	568		
3	470	404	585	
4	439	379	448	548

Regarding the second order marginal probabilities (Table 3), a few noteworthy observations should be made. First, the matrix presented in Table 3 can be interpreted as the proportion of examinees who responded correctly to both items out of the total number of examinees who responded to both items. For example, 42% of examinees who responded to both the first and second item responded correctly to both. Second, the main diagonal of the matrix in Table 3 is equivalent to the first order marginal probabilities. As mentioned previously, separately calculating the first order marginal probabilities is unnecessary, but we feel that presenting the calculations of the first and second order marginal probabilities separately allows for a greater understanding of what these estimates represent and how they were generated.

Table 3

The Matrix of the Probability of an Examinee Responding Correctly to Both Items Across All Pairwise Comparisons After Accounting for Missing Data.

Item	**1**	**2**	**3**	**4**
1	0.61			
2	0.42	0.59		
3	0.51	0.43	0.60	
4	0.48	0.41	0.48	0.47

In addition to modifying $\mathbf{p}_2$ to better account for missing data in the M_2 statistic, *N* should also be modified to reflect that sample size will vary for each of the marginal probabilities to reflect the varying number of responses that were provided to each item or pair of items. Thus, *N* can be modified to be **N**, a vector of the sample sizes for the first and second order marginal probabilities.

The remaining components in the definition of the M_2 statistic do not need to be modified to better account for missing data. Models that incorporate missing data are already accounting for missing data in the model predicted marginal probabilities, meaning that $\hat{\boldsymbol{\pi}}_2$ (i.e., the vector of model predicted marginal probabilities of correctly responding to each item or pair of items) already accounts for the missing data. Similarly, $\hat{C}_2$ relies on model predicted values rather than the observed data, meaning missing data is already accounted for when it is incorporated into the model and no further modifications are necessary.

Distribution of the M_2^* Statistic

Based on the work of Rao (1973), Bishop (1975), and Browne (1984), Maydeu-Olivares and Joe (2005) used Slutsky's theorem to demonstrate that the M_2 statistic is asymptotically distributed along a Chi-squared distribution with $s - q$ degrees of freedom, where *s* is the dimension of $\boldsymbol{\pi}_2$ and *q* is the number of estimated item and structural parameters minus one (i.e., the number of free parameters). Similarly, for DCMs, Hansen et al. (2016) demonstrated that the M_2 statistic is still asymptotically distributed along a Chi-squared distribution with $s - q$ degrees of freedom. By knowing the underlying distribution of the M_2 statistic, it is possible to conduct inferential statistical tests to evaluate model fit. In modifying the M_2 statistic, it is imperative to demonstrate that the M_2^* statistic is also distributed along a Chi-squared distribution with $s - q$ degrees of freedom so that model fit can be evaluated statistically.

Because *N* and $\mathbf{p}_2$ are the only components of the M_2 statistic that were altered in creating the M_2^* statistic, it must be shown that these alterations to *N* and $\mathbf{p}_2$ do not alter the underlying distribution.

Namely, following the work of Hansen et al. (2016), it must be shown that $\sqrt{\boldsymbol{N}}(\boldsymbol{p}_2 - \widehat{\boldsymbol{\pi}}_2)$ is still a vector that follows a normal distribution with a mean of zero and a covariance matrix of $\Xi_2 - \boldsymbol{\Delta}_{2*}\mathcal{F}^{-1}\boldsymbol{\Delta}'_{2*}$, where $\boldsymbol{N}$ is a vector of sample sizes for the first and second order marginal proportions, $\boldsymbol{p}_2$ is the vector of the observed second order marginal probabilities of correctly responding to each item and pair of items, $\widehat{\boldsymbol{\pi}}_2$ is the vector of the model predicted second order marginal probabilities of correctly responding to each item and pair of items, Ξ_2 is the multinomial covariance matrix, $\boldsymbol{\Delta}_{2*}$ is the second order Jacobian matrix containing the first order partial derivates of the response pattern probabilities with respect to the model parameters, $\mathcal{F}^{-1}$ is the inverse of the Fisher information matrix, and $\boldsymbol{\Delta}'_{2*}$ is the transposed second order Jacobian matrix containing the first order partial derivates of the response pattern probabilities with respect to the model parameters. In the work of Maydeu-Olivares and Joe (2005) as well as Hansen et al. (2016) where there were no missing data, *N* was a constant value for all marginal proportions, and $\sqrt{N}(\boldsymbol{p}_2 - \widehat{\boldsymbol{\pi}}_2)$ was shown to follow a normal distribution with a mean of zero and a covariance matrix of $\Xi_2 - \boldsymbol{\Delta}_{2*}\mathcal{F}^{-1}\boldsymbol{\Delta}'_{2*}$. Maydeu-Olivares and Joe (2005) noted that the M_2 statistic follows a Chi-square distribution as long as $\boldsymbol{p}_2$ is a $\sqrt{N}$-consistent estimator of $\widehat{\boldsymbol{\pi}}_2$, for any value of *N*. When data is missing completely at random or missing at random, $\boldsymbol{p}_2$ continue to be a $\sqrt{N}$-consistent estimator of $\widehat{\boldsymbol{\pi}}_2$, as DCMs can adequately estimate parameters in the presence of data missing completely at random or missing at random (Rupp et al., 2010; Shan & Wang, 2020). Further, there is no requirement for *N* to remain constant across the vector of the marginal proportions; thus, utilizing a vector of sample sizes that are specific to marginal proportions still allows $\sqrt{\boldsymbol{N}}(\boldsymbol{p}_2 - \widehat{\boldsymbol{\pi}}_2)$ to be a vector that follows a normal distribution with a mean of zero and a covariance matrix of $\Xi_2 - \boldsymbol{\Delta}_{2*}\mathcal{F}^{-1}\boldsymbol{\Delta}'_{2*}$.

Maydeu-Olivares and Joe (2005) demonstrated that the M_2 statistic has $s - q$ degrees of freedom based on results from Rao (1973) because $\hat{C}_2$ and $\Delta_2^{(c)}$ are of rank $s - q$. In modifying the M_2 statistic, $\hat{C}_2$ was not modified; hence, it is still of rank $s - q$. Similarly, $\Delta_2^{(c)}$ is based on $\boldsymbol{\pi}_2$, which was also not altered in modifying the M_2 statistic. Thus, the M_2^* statistic also has $s - q$ degrees of freedom based on the proofs presented by Maydeu-Olivares and Joe (2005).

Using these proofs building upon the work by Maydeu-Olivares and Joe (2005) as well as Hansen et al. (2016), we demonstrate that the M_2^* statistic is also distributed along a Chi-squared distribution with $s - q$ degrees of freedom. As such, the M_2^* statistic can be used to statistically evaluate model fit using inferential statistical tests, as the underlying distribution of the statistic is known.

Objective

The purpose of this study is to evaluate the performance of the M_2^* statistic. The performance of the M_2^* statistic can be evaluated across a variety of simulated conditions (e.g., sample sizes, between-attribute correlations, proportion of missing data). Further, the performance of the M_2^* statistic can be evaluated in comparison to alternative approaches to addressing missing data that are commonly used (e.g., filtering out examinees with missing data, recoding missing data as incorrect responses).

Simulation Framework

To evaluate the accuracy of the M_2^* statistic in the presence of missing data, we conducted a simulation study. We examined the effect of three factors on the M_2^* statistic. These factors and levels were:

- Sample size: To study the effects of missing data under varying sample sizes for DCMs, we examined DCMs with sample sizes of 1,000 and 5,000.
- Missing data proportions: To study how the extent of missing data influences calculation of the M_2^* statistic, we examined conditions where the proportion of missing data was 0%, 3%, and 5%.
- Between-attribute correlations: To analyze how the presence of missing data interacts with between-attribute correlations when calculating the M_2^* statistic, we examined between-attribute correlations of zero and 0.20.

This study used a full factorial simulation. In total, this study had 12 conditions with 100 replications per condition.

Data Simulation

Each simulated assessment measured three attributes with 12 items. For the Q-matrix (Tatsuoka, 1983), items were specified such that each of the items measured 2 attributes, with each attribute measured by a least 4 items. Table 4 presents a hypothetical Q-matrix for this simulation study.

When generating the true attribute mastery profiles for each of the examinees, we simulated examinees such that there was a 0.50 probability of test takers mastering each of the attributes. We also set the between-attribute correlation to be either zero or .20, depending on the simulation condition, which is representative of uncorrelated attributes and of weakly correlated attributes. The true attribute mastery patterns were simulated using a standard normal cumulative distribution function based on the prevalence of attribute mastery, the between-attribute correlation, and a random number. This approach is consistent with the simulation approach used by Johnson and Sinharay (2018).

Table 4

Hypothetical Q-Matrix

	Attribute		
Item	α_1	α_2	α_3
1	1	0	1
2	1	1	0
3	0	1	1
4	1	1	0
5	0	1	1
6	0	1	1
7	1	1	0
8	1	1	0
9	1	0	1
10	1	1	0
11	0	1	1
12	1	0	1

We simulated item parameters to generate the data using a log-linear cognitive diagnosis model (LCDM; Henson et al., 2009). The item parameters for an LCDM include the item intercepts, main effects, and interaction effects. The item intercepts were drawn from a uniform distribution ranging from .10 to .20 and were then converted to the logit scale. The item intercept translates to the probability that a non-master of the two assessed attributes provides a correct response. The item main effects were drawn from truncated $N(2,\ .5)$, where values were restricted to be positive. Constraining the main effects to non-negative numbers ensures that attribute mastery increases the probability of a test taker providing a correct response. The item interaction effects were drawn from $N(2, 1/6)$. As with the main effects, the interaction effects were drawn from a truncated normal distribution. However, the interaction effect for each item was constrained to be at least as large as negative one times the smallest main effect for that item. This ensures that masters of both assessed attributes have a probability of providing a correct response that is at least as large as the probability of masters of only one attribute providing a correct response.

In LCDMs, the probability of providing a correct response is calculated by summing the combination of item effects that reflect the examinee's attribute mastery profile. When just one of the two assessed attributes were mastered, only the main effect for the mastered attribute is added to the item intercept to produce the probability of providing a correct response. When both assessed attributes were mastered, the main effects and interaction effect are added to the item intercept to produce the probability of providing a correct response. Of note, the sum of the item effects is on the logit scale and must be converted back to the probability scale to be interpreted as the probability of providing a correct response.

Model Estimation

To introduce model misfit into the simulation, we estimated an LCDM and a deterministic-input, noisy-and-gate (DINA; de la Torre & Douglas, 2004; Haertel, 1989; Junker & Sijtsma, 2001) model to each generated dataset. In contrast to the LCDM, DINA models are non-compensatory models, meaning there are two probabilities that examinees will respond correctly: one for all examinees who have not mastered all the assessed attributes and one for examinees who have mastered all the assessed attributes (de la Torre & Douglas, 2004). In terms of the probability of responding correctly, there is no differentiation between examinees who have mastered some of the assessed attributes compared to examinees who have not mastered any of the assessed attributes. Because the simulated data was generated using an LCDM, where the probability of providing a correct response depends on the specific combination of attributes that have been mastered, it is expected that for any generated dataset, the LCDM should fit the data better than the DINA model.

To test the performance of the M_2^* statistic in assessing model fit in the presence of missing data, the LCDMs and DINA models will be estimated with missing data present based on the simulation condition. Because missing data is frequently addressed through data formatting approaches (e.g., recoding missing data as incorrect responses, filtering out examinees with missing data) rather than including the missing data in the estimated model, we also estimated LCDMs and DINA models when recoding missing data as incorrect responses and when filtering out examinees with missing data. Of note, neither the recoded data models nor the filtered data models include missing data; thus, the performance of the M_2^* statistic in the missing data models can be compared to alternative approaches to addressing missing data (e.g., recoding, filtering).

Performance Indices

We calculated the Type I error rates and statistical power to comprehensively evaluate the performance of the M_2^* statistic. To provide further context surrounding the performance of the M_2^* statistic, we also calculated the profile- and attribute-level classification accuracy, and the mean absolute deviation (MAD) of the estimated parameters.

Type I Error

Type I error rates were calculated as the proportion of LCDMs that are flagged for misfit within each condition. The LCDMs were flagged for misfit when the *p* value for the M_2^* statistic was less than .05. Because the true generating model in this simulation study was an LCDM, the estimated LCDMs should demonstrate adequate model fit; hence, LCDMs flagged for misfit are reflective of Type I errors (i.e., identifying misfit when none is present).

Power

Statistical power was calculated as the proportion of DINA models that are flagged for misfit within each condition. As with the Type I error rate calculations, the DINA models were flagged for misfit when the *p* value for the M_2^* statistic was less than .05. Because the true generating model in this simulation study was an LCDM, the estimated DINA models were expected to demonstrate misfit; thus, the power of the M_2^* statistic for detecting misfit is reflected by the proportion of truly misfitting models (i.e., the DINA models) that were correctly flagged as misfitting.

Parameter Recovery

The recovery of person, structural, and item parameters provides context for the performance of the M_2^* statistic. Recovery of the person parameters was measured through classification accuracy. Classification accuracy at the profile- and attribute-level were calculated as the proportion of examinees where the estimated model classified the examinees' latent classes and attribute mastery statuses consistently with the examinees' true latent classes and attribute mastery statuses, respectively. Polychoric correlations and Cohen's kappa were also used to examine profile-level classification accuracy. To quantify the absolute discrepancies between the true and estimated parameter values on the original scale for each parameter, parameter recovery was evaluated using MAD estimates. The MAD estimates were calculated using

$$MAD = \overline{|p_t - p_e|}\ , \tag{2}$$

where p_t is the true value for a generic parameter and p_e is the estimated value for the same generic parameter.

Results

The performance of our M_2^* statistic was primarily informed by Type I error rates and power. To provide additional context for our findings, we also examined classification accuracy and parameter recovery for the models estimated in this simulation study.

Type I Error

The Type I error rates for the missing data, recoded data, and filtered data LCDMs are presented in Table 5. We expected the Type I error rates to be approximately .05 since we used $\alpha = .05$ to flag missing models. In the conditions with no missing data, the Type I error rate of the models incorporating missing data was relatively well controlled. This was expected given that the M_2^* statistic replicates current operationalizations of the M_2 statistic when there is no missing data, and previous studies have demonstrated controlled Type I error rates for the currently operationalization of the M_2 statistic (e.g., Chen et al., 2018; Liu et al., 2016). When missing data was present, the M_2^* statistic demonstrated elevated Type I error rates, ranging from .17 to .31. For the recoded data and filtered data LCDMs, the Type I error rates were relatively well controlled.

Table 5

Type I Error Rates

n	Correlation	% Missing	Missing Data	Recoded Data	Filtered Data
1,000	0	0	.09	--	--
1,000	0	3	.17	.03	.06
1,000	0	5	.30	.07	.02
1,000	0.2	0	.05	--	--
1,000	0.2	3	.18	.04	.03
1,000	0.2	5	.23	.05	.05
5,000	0	0	.03	--	--
5,000	0	3	.20	.08	.10
5,000	0	5	.31	.04	.07
5,000	0.2	0	.05	--	--
5,000	0.2	3	.20	.04	.06
5,000	0.2	5	.22	.09	.08

Power

Across all the estimated DINA models, the M_2^* statistic demonstrated elevated statistical power, especially for the conditions with a sample size of 5,000. In these large sample size conditions across all models, the statistical power was .98 or greater, which suggests great sensitivity to model misfit. For the missing data models, the power ranged from .83 to 1.00, which again suggests great sensitivity to model misfit. However, only one condition demonstrated statistical power less than .91, which suggests the M_2^* statistic demonstrates near perfect identification of misfitting models when missing data is present. This suggests the elevated Type I error rates may be related to an over-flagging tendency when using the M_2^* statistic.

Table 6

Statistical Power

n	Correlation	% Missing	Missing Data	Recoded Data	Filtered Data
1,000	0	0	.95	--	--
1,000	0	3	.99	.91	.83
1,000	0	5	.97	.84	.65
1,000	0.2	0	.82	--	--
1,000	0.2	3	.91	.79	.70
1,000	0.2	5	.91	.74	.57
5,000	0	0	1.00	--	--
5,000	0	3	1.00	1.00	1.00
5,000	0	5	1.00	1.00	1.00
5,000	0.2	0	1.00	--	--
5,000	0.2	3	1.00	1.00	1.00
5,000	0.2	5	1.00	1.00	.98

Follow-Up Analyses

To better understand the elevated Type I error rates in the LCDMs when missing data is present and the broadly elevated statistical power, we explored classification accuracy and parameter recovery to contextualize these findings. It is possible that poor classification accuracy and/or poor parameter recovery could have led to model misfit. To examine this possibility, we compared model estimated profile- and attribute-level classification accuracy as well as the recovery of parameter estimates compared to the generating values.

Classification Accuracy

The profile-level classification accuracy for the estimated LCDMs and DINA models are presented in Table 7. The missing data and recoded data LCDMs demonstrated adequate profile-level classification accuracy. The remaining models, including the filtered data LCDM, demonstrated rather poor profile-level classification accuracy.

The missing data LCDMs demonstrated similar classification accuracy regardless of whether missing data was present, which is noteworthy in indicating that the missing data LCDMs were able to maintain adequate classification accuracy even in the presence of missing data. The recoded LCDMs, however, generally demonstrated slightly lower classification accuracy, which was as much as .10 lower than the classification accuracy of the corresponding missing data LCDMs. Further, increasing amounts of missing data tended to result in a greater decrease in classification accuracy for the recoded data LCDMs, while this trend was much less pronounced for the missing data LCDMs. Taken together, these findings suggest the missing data LCDMs were able to recover simulated examinees' true latent classes better than the recoded data LCDMs and tremendously better than the filtered data LCDMs.

Table 7

Profile-Level Classification Accuracy

			LCDM			DINA		
n	Correlation	% Missing	Missing Data	Recoded Data	Filtered Data	Missing Data	Recoded Data	Filtered Data
1,000	0	0	.64	--	--	.21	--	--
1,000	0	3	.70	.62	.13	.21	.21	.13
1,000	0	5	.66	.56	.13	.21	.20	.13
1,000	0.2	0	.69	--	--	.22	--	--
1,000	0.2	3	.66	.65	.15	.24	.22	.12
1,000	0.2	5	.66	.58	.15	.21	.22	.12
5,000	0	0	.73	--	--	.21	--	--
5,000	0	3	.72	.61	.12	.21	.21	.13
5,000	0	5	.71	.63	.13	.21	.20	.13
5,000	0.2	0	.73	--	--	.23	--	--
5,000	0.2	3	.70	.70	.15	.22	.22	.12
5,000	0.2	5	.73	.66	.15	.24	.23	.12

The polychoric correlations for the true and estimated attribute mastery profiles are presented in Table 8. The polychoric correlations for the missing data and recoded data models for both the LCDMs and the DINA models indicated consistent classifications. The polychoric correlations for the filtered data models for both the LCDMs and the DINA models indicated inconsistent classifications. Across all models, the polychoric correlations tended to be larger with correlated attributes, and the recoded data models tended to have slightly lower polychoric correlations as the proportion of missing data increased.

Table 8

Polychoric Correlations for the Attribute Mastery Profiles

			LCDM			DINA		
		%	Missing	Recoded	Filtered	Missing	Recoded	Filtered
n	Correlation	Missing	Data	Data	Data	Data	Data	Data
1,000	0	0	.871	--	--	.864	--	--
1,000	0	3	.889	.850	.010	.858	.857	.008
1,000	0	5	.874	.824	.005	.861	.857	.006
1,000	0.2	0	.910	--	--	.901	--	--
1,000	0.2	3	.903	.891	.000	.907	.903	.000
1,000	0.2	5	.902	.878	.002	.896	.893	.006
5,000	0	0	.895	--	--	.864	--	--
5,000	0	3	.890	.840	-.002	.863	.864	-.003
5,000	0	5	.891	.845	.006	.864	.862	.006
5,000	0.2	0	.915	--	--	.902	--	--
5,000	0.2	3	.908	.898	.001	.902	.898	.002
5,000	0.2	5	.914	.884	-.001	.899	.892	.001

The Cohen’s kappa estimates of agreement between the true and estimated attribute mastery profiles are presented in Table 9. The Cohen’s kappa interpretation guidelines from Landis and Koch (1977) indicate the agreement between the true and estimated attribute mastery profiles was moderate to good for the missing data and recoded data LCDMs. The agreement between the true and estimated attribute mastery profiles was “slight” for the filtered data LCDM and all the DINA models.

Attribute-level classification accuracy is presented in Table 10. The missing data and recoded data LCDMs demonstrated strong attribute-level classification accuracy. The remaining models demonstrated moderate attribute-level classification accuracy. The attribute-level classification accuracy was notably higher for the missing data and recoded data LCDMs compared to the missing data and recoded data DINA models. The filtered data models demonstrated consistently lower attribute-level classification accuracy across the LCDM and DINA models

Table 9

Cohen's Kappa for the Attribute Mastery Profiles

			LCDM			DINA		
		%	Missing	Recoded	Filtered	Missing	Recoded	Filtered
n	Correlation	Missing	Data	Data	Data	Data	Data	Data
1,000	0	0	.587	--	--	.101	--	--
1,000	0	3	.652	.560	.003	.097	.093	.002
1,000	0	5	.615	.498	.002	.098	.087	.000
1,000	0.2	0	.634	--	--	.108	--	--
1,000	0.2	3	.598	.589	.003	.130	.112	.000
1,000	0.2	5	.605	.512	.002	.099	.109	-.001
5,000	0	0	.690	--	--	.102	--	--
5,000	0	3	.677	.554	.000	.094	.098	.001
5,000	0	5	.673	.581	.002	.095	.091	.002
5,000	0.2	0	.684	--	--	.120	--	--
5,000	0.2	3	.653	.651	.000	.115	.115	.000
5,000	0.2	5	.681	.599	.001	.129	.119	.000

Table 10

Attribute-Level Classification Accuracy

			LCDM			DINA		
		%	Missing	Recoded	Filtered	Missing	Recoded	Filtered
n	Correlation	Missing	Data	Data	Data	Data	Data	Data
1,000	0	0	.857	--	--	.704	--	--
1,000	0	3	.871	.841	.502	.697	.692	.501
1,000	0	5	.862	.822	.502	.699	.690	.502
1,000	0.2	0	.878	--	--	.698	--	--
1,000	0.2	3	.867	.862	.502	.706	.695	.467
1,000	0.2	5	.867	.837	.501	.691	.688	.465
5,000	0	0	.885	--	--	.701	--	--
5,000	0	3	.881	.842	.500	.697	.695	.501
5,000	0	5	.878	.847	.502	.698	.692	.501
5,000	0.2	0	.891	--	--	.701	--	--
5,000	0.2	3	.881	.879	.507	.700	.697	.468
5,000	0.2	5	.890	.863	.504	.703	.692	.467

Structural Parameter Recovery

The MAD estimates for structural parameter recovery in the LCDMs are presented in Table 11. The MAD estimates for the missing data LCDMs were marginally smaller than the MAD estimates for the recoded data and filtered data LCDMs. For the missing data LCDMs, the MAD estimates were consistent across the different levels of missing data. In contrast, the MAD estimates for the recoded data and filtered data LCDMs showed small increases as the proportion of missing data increased, particularly in the small sample size conditions.

Table 11

Mean Absolute Difference of Structural Parameter Recovery

			LCDM		
n	Correlation	% Missing	Missing Data	Recoded Data	Filtered Data
1,000	0	0	.035	--	--
1,000	0	3	.027	.037	.037
1,000	0	5	.030	.045	.047
1,000	0.2	0	.034	--	--
1,000	0.2	3	.038	.036	.039
1,000	0.2	5	.037	.048	.045
5,000	0	0	.016	--	--
5,000	0	3	.017	.030	.017
5,000	0	5	.017	.023	.018
5,000	0.2	0	.022	--	--
5,000	0.2	3	.026	.021	.022
5,000	0.2	5	.022	.029	.032

Item Parameter Recovery

The MAD estimates for item parameter recovery in the small and large sample size conditions are presented in Figure 1 and Figure 2, respectively. The MAD estimates were calculated within each type of parameter (e.g., intercept, main effect, interaction effect) to provide a finer-grained analysis of how well each model recovered the different item parameters.

In the small sample size conditions, the intercept effect MAD estimates for all models were similar. The main effect and interaction effect MAD estimates for the missing data and filtered data

LCDMs were similar to one another. The main effect MAD estimates for the recoded data LCDMs were similar or slightly larger than the main effect MAD estimates for the missing data and filtered data LCDMs. The interaction effect MAD estimates for the recoded data LCDMs were considerably larger than the interaction effect MAD estimates for the missing data and filtered data LCDMs.

In the large sample size conditions, the intercept effect MAD estimates for all models were again similar. When the attributes were uncorrelated, the main effect MAD estimates for the recoded data and filtered data LCDMs were similar or slightly larger than the main effect MAD estimates for the missing data LCDMs. When the attributes were weakly correlated, the main effect MAD estimates for the recoded data and filtered data LCDMs were again similar; however, the main effect MAD estimates for the recoded data and filtered data LCDMs were slightly smaller than the main effect MAD estimates for the missing data LCDMs in the 3% missing data condition and slightly larger than the main effect MAD estimates for the missing data LCDMs in the 5% missing data condition. The interaction effect MAD estimates for the missing data and filtered data LCDMs were similar and were considerably lower than the interaction effect MAD estimates for the recoded data LCDMs.

Figure 1

Item Parameter Recovery Results (N = 1,000)

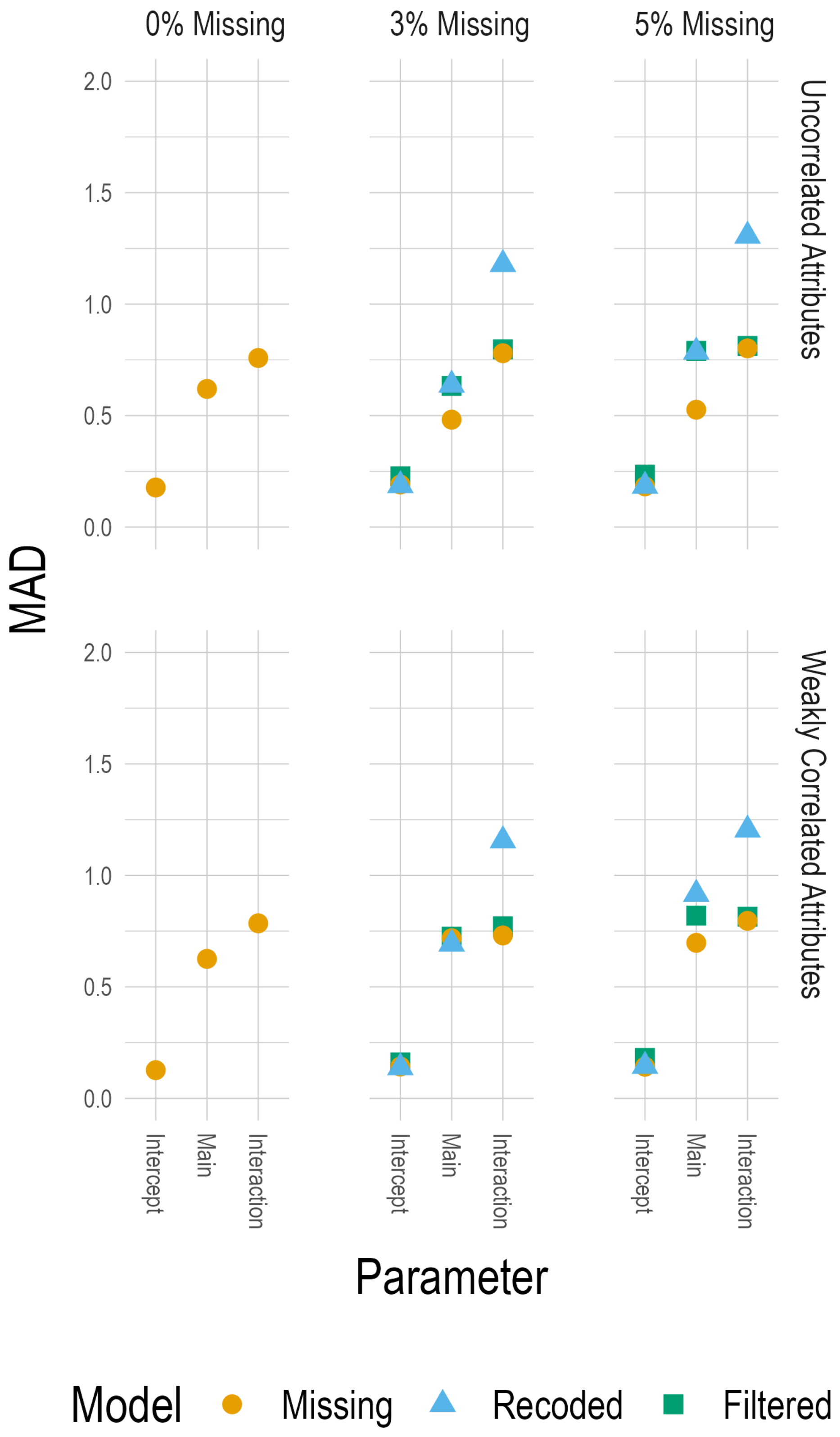
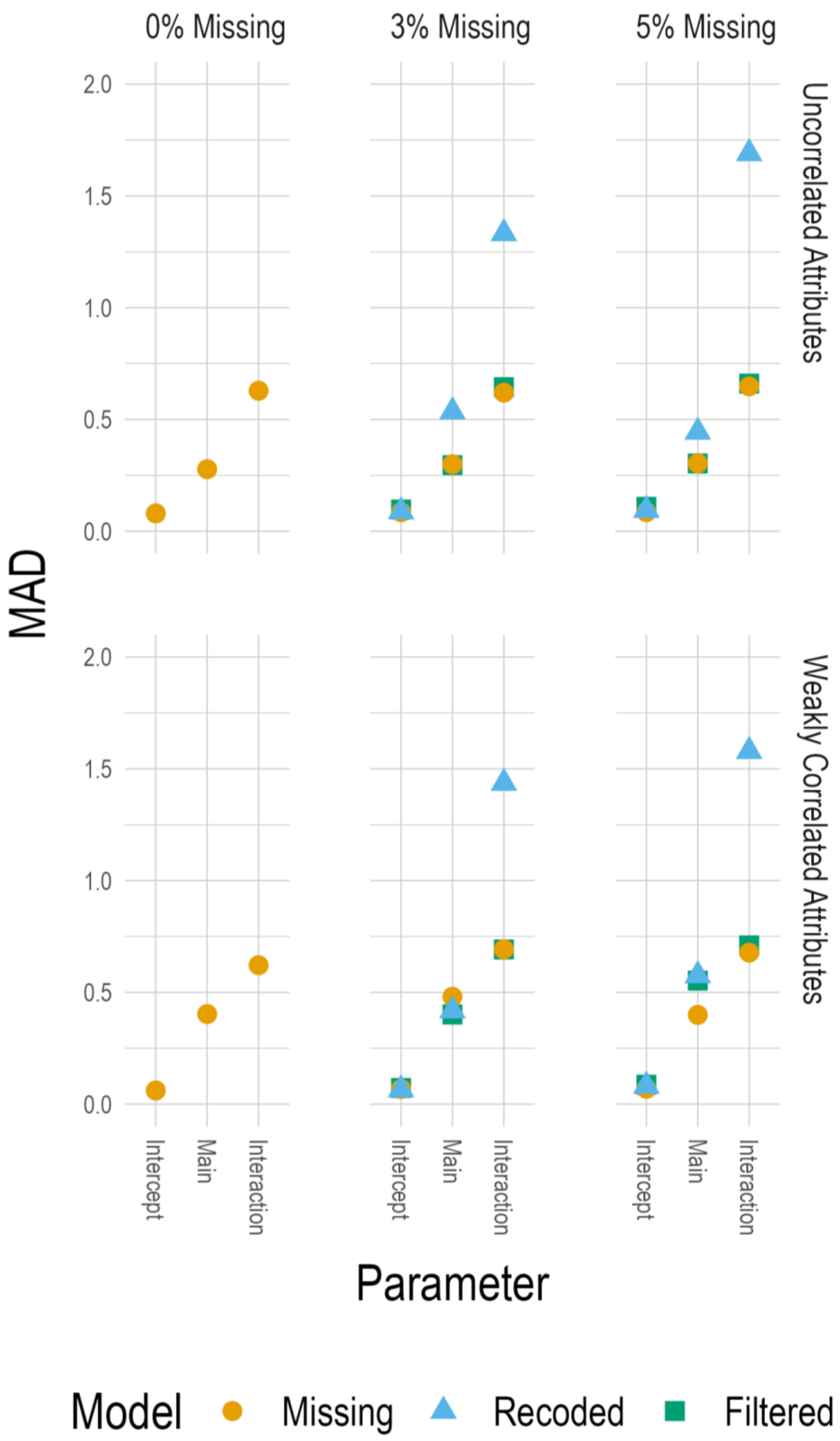


Figure 2

Item Parameter Recovery Results (N = 5,000)



Discussion

This simulation study examined the performance of our M_2^* statistic. Across all the studied conditions, our M_2^* statistic demonstrated elevated Type I error rates and statistical power. This suggests a sensitivity to model misfit, such that too many models were classified as misfitting.

To compare the performance of our M_2^* statistic to alternative methods for addressing missing data, we also estimated LCDMs when missing data was recoded as incorrect responses (i.e., recoded data LCDMs) and when examinees with missing data were filtered out of the data set (i.e., filtered data LCDMs). For the recoded data and filtered data LCDMs, the M_2 statistic indicated relatively controlled Type I error rates and acceptable statistical power, although the statistical power appeared to be slightly elevated in the large sample size conditions.

Purely examining the Type I error rates and statistical power, it appears that our M_2^* statistic was not successful; however, an underlying question is how well the estimated missing data, recoded data, and filtered data LCDMs recovered the person and item parameters of the generating LCDMs. Thus, classification accuracy and the MAD of the estimated parameters provide insight into how well the estimated models recovered the parameters of the generating models.

The estimated missing data and recoded data LCDMs demonstrated adequate attribute mastery profile classification accuracy, but the filtered data LCDM demonstrated attribute mastery profile classification accuracy that was clearly suboptimal. The poor performance of the filtered data LCDM is somewhat unsurprising given that a number of examinees are filtered out of the available data, leaving significantly fewer examinees to be included in the estimated model and thus significantly fewer item responses to use in estimating the model parameters. In the small sample size conditions, the average number of examinees in the filtered data LCDMs ranged from 539 to 694, and the average number of examinees in the filtered data LCDMs ranged from 2,700 to 3,466 in the large sample conditions. This translates to a 31-46% decrease of the sample sizes in both the 1,000 and 5,000 sample size conditions.

The polychoric correlations and Cohen's kappa estimates corroborate the findings of the attribute mastery profile classification accuracy. Both the polychoric correlations and Cohen's kappa estimates indicate strong agreement between the true and estimated latent classes in the missing data and recoded data models, while there is only marginal agreement in the filtered data model.

Given the strong profile-level classification accuracy in the missing data and recoded data models, it is unsurprising that these models also demonstrated strong attribute-level classification accuracy. After all, strong attribute-level classification accuracy is a prerequisite of strong profile-level classification accuracy. However, the moderate attribute-level classification accuracy in the filtered data models is a significant improvement over the marginal profile-level classification accuracy. When combined with the similar but slightly elevated MAD estimates for structural parameter recovery in the filtered data model, this suggests that the moderate error in the estimated structural parameters of the filtered data model led to estimated attribute mastery profiles that were similar to but not equal to the true attribute mastery profiles.

The estimated missing data and filtered data LCDMs demonstrated similar MAD estimates for the item parameters. The recoded data LCDM demonstrated similar MAD estimates for the intercept parameters, although there were minor differences for some of the main effect MAD estimates and significant differences for all the interaction effect MAD estimates. Theoretically, it is unsurprising that the main effect and interaction effect MAD estimates were elevated for the recoded data LCDMs. Because missing data are recoded as incorrect responses in the recoded data LCDMs, the practical impact on item parameter recovery is only actualized when the recoded response is different than the response that would have likely been provided. In other words, the practical impact on item parameter recovery occurs when the examinee would have likely responded correctly yet the missing data is recoded as incorrect. More specifically, recoding missing responses as incorrect from examinees who would have likely responded correctly makes the items appear to be more difficult than they truly are, which would inflate the main and interaction effect estimates. Thus, recoding missing data as incorrect responses likely explains the elevated MAD estimates for the main and interaction effects in the recoded LCDMs.

The parameter recovery results are largely consistent with other DCM simulations examining the impact of missing data on parameter recovery. Sünbül (2018) found that recoding data is incorrect led to decreased classification accuracy and increased error in the item parameter estimates when estimating DINA models. Further, Sünbül (2018) found that classification accuracy tended to decrease and item parameter error tended to increase with increasing amounts of missing data, which is consistent with the findings of this study. Shan and Wang (2020) found profile-level classification accuracy to be approximately .70, attribute-level classification accuracy to be above .90, and the item parameters to be recovery adequately when data were missing completely at random. Both findings were consistent with the findings of this study, although the Shan and Wang (2020) study utilized fewer examinees and more items, which increases the difficulty of making a direct comparison of those findings and the findings from this study.

Taking all these findings together, the results from this study are inconclusive regarding how to estimate model fit in the presence of missing data for DCMs. The Type I error rate was clearly suboptimal for the missing data LCDMs, but the Type I error rates were relatively well controlled for the recoded data and filtered data LCDMs. However, the classification accuracy and MAD estimates for the item parameters indicated the filtered data LCDM and the recoded data LCDM, respectively, were also suboptimal. Given the totality of these findings, recoding and filtering missing data appeared to allow an LCDM to be estimated that fits the resulting recoded or filtered data set, respectively, although these models were not adequate representations of the generating model. Conversely, incorporating missing data into the LCDM allowed for the best recovery of the person and item parameters; however, the M_2^* statistic did not recognize that model fit.

Because of the ubiquity of missing data as well as importance of appropriately addressing missing data, future work should continue modifying existing model fit indices to adequately estimate model fit in the presence of missing data. Additionally, future work may consider continued examination of the appropriateness of using recoding and filtering to address missing data. While the findings of this study are far from conclusive, the findings of this study cast doubt as to whether these approaches to addressing missing data allow for adequate parameter recovery.

Conclusion

The M_2^* statistic developed in this study performed sub-optimally in terms of Type I error rates and statistical power. However, the findings of this study indicated that the missing data LCDMs were able to adequately recover the person and item parameters, whereas the recoded and filtered data models were not able to adequately recover the person and item parameters. More work is needed to identify statistics that control Type I error rates in the presence of missing data while still adequately recovering the person and item parameters. Additional work may also be needed to explore implications of recoding and filtering missing data in terms of parameter recovery.

References

Ames, A. J., & Penfield, R. D. (2015). An NCME instructional module on item-fit statistics for item response theory models. *Educational Measurement: Issues and Practice*, *34*(3), 39–48.

Bartholomew, D. J., & Leung, S. O. (2002). A goodness of fit test for sparse 2^p contingency tables. *British Journal of Mathematical and Statistical Psychology*, *55*, 1–15.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis*. MIT Press.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37,* 62-83. https://doi.org/10.1111/j.2044-8317.1984.tb00789.x

Chen, F., Liu, Y., Xin, T., & Cui, Y. (2018). Applying the M2 statistic to evaluate the fit of diagnostic classification models in the presence of attribute hierarchies. *Frontiers in Psychology*, *9*, 1875.

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, *50*, 123–140. https://doi.org/10.1111/j.1745-3984.2012.00185.x

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333–353.

Dynamic Learning Maps Consortium. (2020). *2019-2020 Technical Manual Update—Instructionally Embedded Model.* [Technical Report].

Gu, Z. (2011). *Maximizing the Potential of Multiple-Choice Items for Cognitive Diagnostic Assessment* [Dissertation]. University of Toronto.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 301–323.

Han, Z., & Johnson, M. S. (2019). Global- and item-level model fit indices. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of Diagnostic Classification Models* (pp. 359–377). Springer Nature. https://doi.org/10.1007/978-3-030-05584-4_17

Hansen, M., Cai, L., Monroe, S., & Li, Z. (2014). Limited-Information Goodness-of-Fit Testing of Diagnostic Classification Item Response Theory Models. CRESST Report 840. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*.

Hansen, M., Cai, L., Monroe, S., & Li, Z. (2016). Limited-information goodness-of-fit testing of diagnostic classification item response models. *British Journal of Mathematical and Statistical Psychology*, *69*, 225–252.

Henson, R. A., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191–210.

Joe, H., & Maydeu-Olivares, A. (2010). A general family of limited information goodness-of-fit statistics for multinomial data. *Psychometrika*, *75*(3), 393–419. https://doi.org/10.1007/s11336-010-9165-5

Johnson, M. S., & Sinharay, S. (2018). Measures of agreement to assess attribute-level classification accuracy and consistency for cognitive diagnostic assessments. *Journal of Educational Measurement*, *55*(4), 635–664.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258–272.

Jurich, D. P. (2014). *Assessing model fit of multidimensional item response theory and diagnostic classification models using limited-information statistics* [PhD Thesis]. James Madison University.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Little, R. J., & Rubin, D. B. (2020). *Statistical analysis with missing data* (3rd edition). John Wiley & Sons.

Liu, Y., Douglas, J. A., & Henson, R. A. (2009). Testing person fit in cognitive diagnosis. *Applied Psychological Measurement*, *33*(8), 579–598. https://doi.org/10.1177/0146621609331960

Liu, Y., Tian, W., & Xin, T. (2016). An application of M2 statistic to evaluate the fit of cognitive diagnostic models. *Journal of Educational and Behavioral Statistics*, *41*(1), 3–26.

Ma, W. (2019). Evaluating the fit of sequential G-DINA model using limited-information measures. *Applied Psychological Measurement*, 1–15. https://doi.org/10.1177/0146621619843820

Ma, W., & de la Torre, J. (2020a). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software*, *93*(14), 1–26. https://doi.org/10.18637/jss.v093.i14

Ma, W., & de la Torre, J. (2020b). *GDINA: The generalized DINA model framework*.

Ma, W., Jiang, Z., & Schumacker, R. E. (2020). Modeling omitted items in cognitive diagnosis models. *AERA Annual Meeting*. AERA, San Francisco, CA.

Madison, M. J., & Bradshaw, L. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and Psychological Measurement*, *75*(3), 491–511.

Maydeu-Olivares, A., & Garcia-Forero, C. (2010). Goodness-of-fit testing. *International Encyclopedia of Education*, *7*(1), 190–196.

Maydeu-Olivares, A., & Joe, H. (2005). Limited- and Full-Information Estimation and Goodness-of-Fit Testing in 2^n Contingency Tables: A Unified Framework. *Journal of the American Statistical Association*, *100*(471), 1009–1020.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713–732. https://doi.org/10.1007/s11336-005-1295-9

Maydeu-Olivares, A., & Joe, H. (2008). An overview of limited information goodness-of-fit testing in multidimensional contingency tables. *New Trends in Psychometrics*, 253–262.

Maydeu-Olivares, A., & Joe, H. (2014). Assessing Approximate Fit in Categorical Data Analysis. *Multivariate Behavioral Research*, *49*(4), 305–328. https://doi.org/10.1080/00273171.2014.911075

Pan, Y., & Zhan, P. (2020). The impact of sample attrition on longitudinal learning diagnosis: A prolog. *Frontiers in Psychology*, *11*, 1051. https://doi.org/10.3389/fpsyg.2020.01051

Rao, C. R. (1973). *Linear statistical inference and its applications.* Wiley.

Rubin, D. B. (1976). Inference and missing data (with discussion). *Biometrika*, *63*, 581–592. https://doi.org/10.1093/biomet/63.3.581

Rupp, A. A., & Templin, J. (2008). Unique Characteristics of Diagnostic Classification Models: A Comprehensive Review of the Current State-of-the-Art. *Measurement: Interdisciplinary Research & Perspective*, *6*(4), 219–262. https://doi.org/10.1080/15366360802490866

Rupp, A. A., Templin, J., & Henson, R. A. (2010). Diagnostic measurement: Theory, methods, and applications. *New York: Guilford*.

Shan, N., & Wang, X. (2020). Cognitive diagnosis modeling incorporating item-level missing data mechanism. *Frontiers in Psychology*, *11*, 3231. https://doi.org/10.3389/fpsyg.2020.564707

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnostic models: A case study. *Educational and Psychological Measurement*, *67*(2), 239–257. https://doi.org/10.1177/0013164406292025

Sorrel, M. A., Abad, F. J., Olea, J., de la Torre, J., & Barrada, J. R. (2017). Inferential item-fit evaluation in cognitive diagnosis modeling. *Applied Psychological Measurement*, *41*(8), 614–631. https://doi.org/10.1177/0146621617707510

Sünbül, S. Ö. (2018). The impact of different missing data handling methods on DINA model. *International Journal of Evaluation and Research in Education (IJERE)*, *7*(1), 77–86.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*(4), 345–354. https://doi.org/10.1111/j.1745-3984.1983.tb00212.x