psych

MDPI

Article

Dealing with Missing Responses in Cognitive Diagnostic Modeling

Shenghai Dai [1,*] and Dubravka Svetina Valdivia [2]

[1] Department of Kinesiology and Educational Psychology, College of Education, Washington State University, Pullman, WA 99164, USA
[2] Department of Counseling and Educational Psychology, School of Education, Indiana University Bloomington, Bloomington, IN 47405, USA; dsvetina@indiana.edu
* Correspondence: s.dai@wsu.edu

Abstract: Missing data are a common problem in educational assessment settings. In the implementation of cognitive diagnostic models (CDMs), the presence and/or inappropriate treatment of missingness may yield biased parameter estimates and diagnostic information. Using simulated data, this study evaluates ten approaches for handling missing data in a commonly applied CDM (the deterministic inputs, noisy "and" gate (DINA) model): treating missing data as incorrect (IN), person mean (PM) imputation, item mean (IM) imputation, two-way (TW) imputation, response function (RF) imputation, logistic regression (LR), expectation-maximization (EM) imputation, full information maximum likelihood (FIML) estimation, predictive mean matching (PMM), and random imputation (RI). Specifically, the current study investigates how the estimation accuracy of item parameters and examinees' attribute profiles from DINA are impacted by the presence of missing data and the selection of missing data methods across conditions. While no single method was found to be superior to other methods across all conditions, the results suggest the use of FIML, PMM, LR, and EM in recovering item parameters. The selected methods, except for PM, performed similarly across conditions regarding attribute classification accuracy. Recommendations for the treatment of missing responses for CDMs are provided. Limitations and future directions are discussed.

Keywords: missing data; cognitive diagnostic modeling; DINA model; imputation; item parameter recovery; classification accuracy

Citation: Dai, S.; Svetina Valdivia, D. Dealing with Missing Responses in Cognitive Diagnostic Modeling. *Psych* **2022**, *4*, 318–342. https://doi.org/10.3390/psych4020028

Academic Editor: Alexander Robitzsch

Received: 14 May 2022
Accepted: 13 June 2022
Published: 14 June 2022

1. Introduction

Cognitive diagnostic models (CDMs; e.g., [1,2]), otherwise referred to as diagnostic classification models (DCMs; e.g., [3,4]), are a family of models that allow for the estimation of examinees' mastery status on a set of pre-specified attributes (or skills) using their cognitive item responses. In the context of CDMs, skills (e.g., [5]) and attributes (e.g., [6]) have been used interchangeably. We use attributes in the paper. Examples of CDMs include the deterministic inputs, noisy "and" gate (DINA) model [7], the reparametrized unified model (RUM, known as the fusion model, [8]), the log-linear cognitive diagnostic model (LCDM, [9]), and the general diagnostic model (GDM, [10]).

In practice, the attribute mastery information of examinees is known as their attribute profile [11]. Attribute profiles are usually in a binary format (e.g., mastery vs. non-mastery); they enable researchers to make inferences about whether an individual possesses or has mastered the hypothesized attributes, draw conclusions about students' instructional needs, and evaluate the effectiveness of current instructional programs [12–17].

The CDM family has been gaining attention following their introduction during the mid-1980s, when they were proposed by psychometricians to be appropriate for diagnostic purposes, and their applications have been widely discussed across many contexts (e.g., [10,14,16,18–26]). To date, both methodological advances and applications of CDMs

have been observed [15,27]. Examples of such methodological advances include the extension and development of the models, such as the generalized DINA model (GDINA, [28]) and the multistrategy DINA model (MS-DINA; [29]), Q-matrix validation and refinement methods (e.g., [5,30,31]), the investigation of differential item functioning in CDMs (e.g., [6,32,33]), and the existence of various software tools such as the **CDM** [34] and **GDINA** [35] packages in R [36]. Examples of CDM applications are evidenced by studies collected in the special issue of "Cognitive Diagnosis and Q-Matrices in Language Assessment" (e.g., [21,22,37,38]) in the journal *Language Assessment Quarterly*, as well as many others (e.g., [2,17,39–45]).

Despite these methodological advances and the increasing applications of CDMs, as indicated by Ravand and Baghaei [27] in their recent review article, there are issues that need to be addressed in order to move forward with the implementation of CDMs. One such issue is the presence of missing responses and how to handle them in CDM applications [27]. In an assessment, it is common that examinees may leave a few items unanswered for varying reasons, such as lack of confidence, test-taking strategy, lack of time, or metacognitive factors [46]. As CDMs are statistical models, it is very likely that the presence of missing data and its inappropriate treatment will lead to biased model estimates (e.g., examinees' estimated attribute profiles and item parameters) in the application of CDMs. Students may be misclassified (e.g., students are told they mastered a skill while they did not) if missing responses are ignored or improperly treated.

In CDM practices, missing data are usually ignored under a default likelihood-based estimator in the software, such as the full information maximum likelihood (FIML) estimation used by the R package **CDM [34]**. The robust performance of FIML has been supported across contexts, even when its assumption of MAR is violated [47–51]. Such a method, however, ignores potential reasons (e.g., lack of ability) for the presence of missing responses by using the available data only, which may encourage examinees to apply specific strategies when taking the test [52]. That is, an examinee can be categorized as a master of certain attributes by responding only to the items for which he or she can endorse a correct answer. Consequently, an imputation method is usually preferred to handle missing responses in practice, especially in obtaining individual ability estimates. For example, in many large-scale assessments such as the Trends in International Mathematics and Science Study (TIMSS), the omitted responses are treated as ignored in estimating item parameters of the selected item response theory (IRT) model and as incorrect in estimating individual-level ability scores or plausible values [53].

In the literature, research had been conducted to study the impact of missing responses; a variety of approaches had been developed to deal with this problem, most of which are within the framework of IRT. To the best of our knowledge, however, no study has yet systematically examined the phenomenon of missing data in the implementation of CDMs, with the exception of a few recent studies (e.g., [17,54–56]). Specifically, Sünbül [56] investigated the impact of four methods for treating missing data, including treating missing data as incorrect (IN), person mean (PM) imputation, two-way (TW) imputation, and expectation-maximization (EM) imputation, using the DINA model under both missing completely at random (MCAR) and missing at random (MAR) mechanisms. Dai et al. [55] examined the impact of missing responses on two Q-matrix validation methods. Shan and Wang [54] developed a joint CDM, the Higher-Order DINA (HO-DINA) model, together with the Bayesian Markov chain Monte Carlo (MCMC) estimation method, to handle item-level missing data. Xu and von Davier [17] explored the impact of missing responses on GDM (see the next section for a detailed review of the studies). While these studies shed light on the treatment of missing responses for CDMs, only certain missing data methods were included in the studies. Furthermore, none of these studies examined the necessity of an imputation method by comparing its performance to FIML. In light of this, the purpose of the present simulation study is to fill this gap in the literature by systematically investigating the impact of missing data on CDMs. We compare the performance of

different missing data handling approaches, including FIML, in CDM applications across a variety of data conditions.

This paper is organized as follows. The next section provides the background and literature review for this study, including an introduction to CDMs and previous research with regard to both CDMs and missing data. We center our discussion on the relevant CDM employed in the current study, namely, the DINA model, as well as the missing data mechanisms considered in the study. The third section describes the methods and design of the simulation study. The fourth section describes the results as they pertain to the previously-stated research purpose. Finally, we discuss the findings and implications for future research and point out the limitations of the present study.

2. Background and Literature

2.1. Cognitive Diagnostic Models (CDMs)

CDMs are probability-based statistical models that are used to analyze observed item response data, usually of a dichotomous and/or polytomous format, from cognitive assessments with the purpose of drawing diagnostic conclusions about students' performance on hypothesized attributes [4]. They are a family of statistical models of a multidimensional and confirmatory nature [15,57]. A common feature of CDMs is that the implementation of CDMs requires a pre-specified loading structure, known as the Q-matrix, that identifies the interaction between attributes and assessment items. It is defined as a matrix having rows as assessment items and columns as attributes, where '1' in a cell indicates that mastery of a certain attribute is required to respond to that item correctly, and '0' otherwise [58]. The Q-matrix is the key factor in determining the quality of diagnostic information for individual remediation in CDM applications [59]. It needs to be correctly specified in order to satisfy the purpose of the assessment while at the same time being theoretically defensible [12]. In the literature, various approaches to constructing a Q-matrix have been proposed and discussed, such as the interview and verbal (think-aloud) protocol, the digital eye-tracking method, the multiple-rater method, and several statistical methods (e.g., [57,60,61]).

Due to the popularity and usefulness of CDMs in providing fine-grained diagnostic information that goes beyond summative total scores, a large number of models have been developed over the past several decades [12,27]. Commonly used and widely mentioned CDMs include the DINA model [7] and its extensions, such as the HO-DINA and MS-DINA models, the deterministic input noisy-or-gate (DINO) model [2], the RUM or fusion model [8,11,62], the G-DINA model [28], the LCDM [9], and the GDM [10]. Despite the differences among CDMs, it is important to note that all models in the family serve the same purpose: to diagnose examinees' attribute profiles. A more comprehensive taxonomy of CDMs can be found in Rupp and Templin [15] (p. 239) and Rupp et al. [4] (p. 98).

The DINA Model

We consider the DINA model [7] in the current study for several reasons. First, as indicated previously, it is one of the most popular [15] and widely used CDMs in the literature, including both empirical and simulation studies. Second, its parsimonious nature allows for easier interpretation of the results compared to other CDMs [31,63] as well as its potential extension to more complex CDMs [64].

Assuming that a test consists of J items that aim to measure K attributes, the definition of the DINA model is illustrated below.

Let Y_{ij} (with possible values of $Y_{ij} = 1$ or 0) be the correct or incorrect item responses of examinee i to item j, q_{jk} (with possible values $q_{jk} = 1$ or 0) be elements of the Q-matrix that represent whether attribute k is required to respond to item j correctly, and α_{ik} (with possible values $\alpha_{ik} = 1$ or 0) be elements of the attribute profiles that represent whether examinee i is a master of attribute k. Parameter η_{ij} is thus formulated as the conjunctive

or non-compensatory attribute master status of examinee i on item j (i.e., whether the examinee is a master of all required attributes of the item or not), where

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}{}^{q_{jk}}. \quad (1)$$

In practice, it is possible for an examinee to respond to an item incorrectly even though the examinee possesses all required attributes, or vice versa. Consequently, two error probabilities, s_j and g_j, are modeled in the DINA model to represent the relationship between η_{ij} and Y_{ij}, as follows:

$$s_j = P(Y_{ij} = 0 | \eta_{ij} = 1) \quad (2)$$

and

$$g_j = P(Y_{ij} = 1 | \eta_{ij} = 0). \quad (3)$$

Specifically, s_j is the slipping parameter and represents the probability of an examinee responding to item j incorrectly despite possessing all the required attributes, while g_j is the guessing parameter that represents the probabilities of responding to item j correctly when at least one required attribute is not possessed by the examinee. Then, taking into account Equations (1) to (3), the probability of examinee i with attribute profile α_{ij} answering item j correctly is defined as

$$P(Y_{ij} = 1 | \alpha_{ij}) = (1 - s_j)^{\eta_{ij}} g_j{}^{(1-\eta_{ij})}. \quad (4)$$

2.2. Missing Responses in Psychometrics and CDMs

In this section, we first briefly introduce the missing data mechanisms and common methods used to treat missing responses in broad assessment contexts. We then turn our focus on the treatment of missingness in CDM research.

2.2.1. Missing Data Mechanisms

Rubin's [65] three mechanisms, namely, MCAR, MAR, and missing not at random (MNAR), are commonly used to distinguish different distributions of missing data. If the distribution of missing data does not have any systematic cause or depend on any other information, it is considered MCAR. With MCAR, missing responses can be treated as a random sample of the complete data in which no missing data is present [47,66]. When response data are MAR, the probability of a missing response depends on other measured characteristics of the individual, and not on the item itself. In addition, the characteristics of the individual should be observable and measurable [66,67]. For example, data could be considered MAR if missing data on a mathematics item are dependent on examinees' confidence, gender, or ethnicity [48,66]. When data are MNAR, the probability of a missing response on an item only depends on that item itself. For instance, an item with sensitive or biased content would likely result in data that are MNAR.

Researchers in the context of educational assessment have identified several reasons for item responses to be missing. Several scholars have suggested that missing responses can be assumed to be MAR because they are related to certain individual characteristics. For example, Lord [52] suggested that if an examinee knows how ability is estimated in IRT, he or she could obtain a high(er) score simply by responding to only those items that he or she has confidence in answering correctly. In addition, de Ayala et al. [46] stated that examinees with higher proficiency might have a greater tendency to omit items when they might not know the answer than those with lower proficiency. In a study by Finch [66] that used simulated assessment data, on the contrary, the probability of missingness was set to be related inversely to students' proficiency, which was measured using the observed total score of other variables. Others have suggested that the likelihood of an item being missing is more likely to be dependent on the characteristics of the item itself, in which case missing responses could be represented as MNAR [68–71]. In examining variation in missing responses on the National Assessment of Educational Progress (NAEP), for

instance, Brown et al. [69] showed that nearly all predictors, including student proficiency and item difficulty, explained negligible amounts of variance in omitted scores, with the exception of item format (i.e., multiple choice, short constructed response, or long constructed response), which accounted for 21.3 to 52.6% of the variance in omitted responses in 2009 grade 4 and grade 8 mathematics assessment data. Furthermore, researchers have suggested that omitted responses in a test were reported by students inadvertently, in which case the missingness would be treated as MCAR [72].

2.2.2. Missing Responses in Assessment Settings

Missing responses are inevitable in many assessments. In practice, missing responses in an assessment can fall into only one category (e.g., omitted responses, [73]). Alternatively, they can be categorized into multiple types of missing data, such as missing-by-design, omitted, and not-reached responses, based on whether items are presented to examinees or the position where the missingness happens. In certain assessments, particularly large-scale assessments such as TIMSS and NAEP, it is typical to utilize what is known as the booklet design, within which examinees are administered only a proportion/booklet of items rather than the entire item pool. Missing responses to items that are not administered to examinees are considered missing-by-design. For items that are presented to students, missing responses are further divided into omitted responses (missing data occurs within a block of responses that are presented to the student) and not-reached responses (missing data subsequent to the last question the student answered) in certain assessments. As missing-by-design is usually assumed to be MCAR and is often avoided in practice by researchers using only booklet-level response data (e.g., [31,41]), missing data in the current study only refer to omitted and not-reached responses. Furthermore, following previous studies that discussed missing data in the context of large-scale assessments (e.g., [17,51,74,75]), we did not distinguish the two categories of missing responses (i.e., omitted and not-reached) in the current study.

2.2.3. Treatment of Missing Responses in Assessment Data

Previous studies in educational measurement have suggested that both types of missing responses are nonignorable in statistical inferences, and may result in biased item and ability parameter estimates [46,68,76–78]. In light of this, a large number of different options for dealing with missing data have been proposed in the literature, which can be categorized into four broad groups [79]. For space reasons, we briefly introduce these methods in this section. For detailed introductions and comprehensive reviews of the methods, see Dai [79], Finch [66], Pohl and Becker [74], Robitzsch [80], or Schafer and Graham [47].

The first group of methods, namely, missing response ignoring methods, is to either remove all cases with missingness from the analysis (e.g., listwise deletion [LW]) or conduct the analysis on available cases using a likelihood-based estimator such as FIML.

The second group is the family of single imputation methods, through which each missing response is replaced with a single value. The value can be zero (i.e., IN; e.g., [41]), a fraction of $1/m$ where m is the number of response categories of a multiple-choice item (i.e., treat missing responses as fractionally correct (FR), e.g., [76]), PM, item mean (IM), or imputed from a specific method such as TW imputation [81], response function (RF) imputation [82], logistic regression (LR), EM imputation [48,49], or predictive mean matching (PMM) [83]. It should be noted that several of the single imputation methods (e.g., PM, IM, and TW) can be implemented either deterministically or stochastically [81,84]. The stochastic versions of the methods, denoted as PM-E, IM-E, and TW-E, are achieved by adding a random error with a specific distribution to the computed value [81].

The third group is the multiple imputation (MI) approach [85]. It replaces the missing data with multiple sets (e.g., five or ten) of values, resulting in multiple imputed data sets. Consequently, the same analysis needs to be conducted using each imputed data set. A pooling procedure such as Rubin's method is then used to combine all sets of the

results [47,83,86]. The MI is not a specific imputation method; rather, it is a multi-step approach to handling missing data by taking into account the uncertainty in the imputing process [47]. Thus, it can be implemented with any stochastic imputation method such as EM imputation, PMM, and the TW-E.

The fourth group is the family of model-based methods [51,74,75,80,87–91]. The model-based methods assume a missing tendency or response propensity that is dependent on the examinees' ability. To implement such a method, the missing tendency is usually modeled simultaneously with the ability that the assessment is designed to measure. Explicitly, the missing tendency can be a manifest variable that is achieved by counting the number of missing responses in the test. It can also be a latent variable and estimated using a specific measurement model (e.g., the Rasch model) with the missing response indicators that are transformed from the item response data (i.e., 1 = missing response, 0 = not a missing response). In addition to their use in directly estimating parameters, the model-based methods can be applied as an imputation method to handle missing responses (e.g., [92–95]).

The performance of the abovementioned methods has been investigated and compared by many studies using both empirical (e.g., [80,96]) and simulated data (e.g., [46,51,66,74,82,91,97–99], especially in the context of unidimensional IRT. While empirical studies have suggested the use of different missing response treatments would yield potentially "considerable" [96] (p. 629) differences in parameter estimations, simulation studies, however, have shown no evidence of a particular method being superior to others. Support has been found for the use of FR (e.g., [46,66]), TW and RF (e.g., [82]), FIML (e.g., [99]), MI ([66,97]), and model-based methods (e.g., [74]) across various studies. Bernaards and Sijtsma [81,100] have suggested that EM algorithm and PM-based approaches (e.g., PM, TW) yield more accurate estimates when data are multidimensional.

2.2.4. Treatment of Missing Responses in CDMs

As indicated above, the issue of missing responses has not been systematically investigated in the context of CDMs. In the limited practical research on CDMs where missing responses are reported, researchers have usually treated omissions as incorrect (e.g., [37,41,101]). Moreover, to the best of our knowledge, only four recent studies (i.e., [17,54–56]) have investigated the impact of missing responses in CDMs. We briefly describe the main aspects of these studies in order to further situate our current work.

To investigate the parameter recovery of CDMs in the presence of a sparse data matrix, Xu and von Davier [17] conducted a simulation study using GDM. Specifically, the authors first simulated 40 datasets involving 2880 examinees, 36 items that measured four attributes, and three proportions of missing data (10%, 25%, and 50%). GDM was then used to analyze the sparse data sets without handling the missing data. After the GDM analysis, Xu and von Davier examined the recovery of both item and person parameters and reached the conclusion that the missing data did not have a significant effect on either item or person parameters, even in the 50% missing condition. This study, however, has several limitations. First, it only considered the MCAR missing mechanism, as "missing responses were randomly assigned to items for each examinee using the corresponding proportion." [17] (p. 3). Other mechanisms of missing data were not considered. Second, no missing data handling methods were used.

Dai et al. [55] examined the impact of missing responses on two Q-matrix validation methods, the EM-based δ-method [31] and the nonparametric refinement method [30]. In their study, the performance of four missing data methods (EM, IN, LR, and LW) was compared in terms of missing mechanism (MAR and MNAR), missing rate (10%, 20%, and 30%), test length (J = 20 and 40 items), number of attributes in the Q-matrix (K = 3 and 5), sample size (N = 500; 1000; and 2000), and Q-matrix misspecification rate (5%, 10%, and 20%). The results of the study suggested that EM and LR outperformed IN and LW in Q-matrix validation. This study, however, did not investigate the impact of missingness on specific CDMs.

Sünbül [56] compared the performance of four missing data methods (IN, PM, TW, and EM) on the DINA model under both MCAR and MAR. The data conditions used in the study were missing rate (5%, 10%, and 15%), test length (J = 15 and 30), and sample size (N = 1000; 2000; and 3000). The results of the study suggested a noticeable impact of the missing rate on both item parameter recovery and classification accuracy. The effects of sample size and test length, however, were small across the conditions except in that a larger number of items yielded larger model estimation biases under MCAR. The results revealed similar performance of TW, PM, and EM, and they all performed superior to IN across conditions. Limitations of this study lie in that: (1) only two missing mechanisms (i.e., MCAR and MAR) with small to moderate missing proportions (5~15%) were considered; and (2) both the number of attributes in the Q-matrix (i.e., K = 4) and the item quality (i.e., both item parameters were drawn from a uniform distribution of U (0.10, 0.30)) were fixed at only one level.

A recent study by Shan and Wang [54] explored the use of a model-based approach in handling missing responses for CDMs. Specifically, the authors proposed a categorical missingness propensity for the item-level missing response. The propensity model was then modeled jointly with a HO-DINA model for the item responses, resulting in the use of a joint CDM. The estimation of the joint CDM was accomplished by a Bayesian MCMC method. The performance of the joint CDM was investigated in terms of missing mechanism (MCAR and MNAR), sample size (N = 500 and 1000), test length (15 and 30 items), and missing rate (high and low) in the first simulation study. A second simulation study with N = 500 and J = 15 was further conducted to examine the sensitivity of the model fit for the incorrect use of MNAR. For both studies, a Q-matrix with five attributes was used for the study. The results revealed that the joint modeling approach was able to recover the parameters well when the missing data mechanism was correctly specified.

While these four studies have shed light on the treatment of missing data in CDMs, a clear guideline for an optimal approach to handling missing responses in the context of CDMs remains elusive. Further research is needed to systematically investigate this issue and guide the implementation of CDMs in the presence of missingness.

3. Methods

To investigate the performance of methods for handling missing data in the implementation of CDMs, we conducted a Monte Carlo simulation study with selected design factors. Levels of the factors were specified following the literature on both missing data in psychometrics and on CDMs.

3.1. Simulation Design

3.1.1. Factors Related to Missing Data

Missing data mechanisms. We considered two levels of missing data mechanisms: MAR and MNAR. We did not consider MCAR. The reason was twofold: (1) the previous literature suggested that MCAR did not have a significant impact on the estimation of either person or item parameters in CDM (e.g., [17]); and (2) we hypothesized that there was always a reason for an examinee's skipping behavior on a test. In addition to MAR and MNAR, we followed de Ayala et al. [46] and generated the missing responses based on a real dataset, which we refer to here as the MIXED type of missing mechanism. The real dataset was obtained from the 2011 Grade 8 NAEP mathematics assessment.

Missing rate (MR). Five missing rates (MR = 0%, 5%, 10%, 20%, and 30%) were specified for the two missing mechanisms of MAR and MNAR. For the MIXED mechanism, there were only two levels of missing rate, including MR = 0% (i.e., complete data) and the actual missing rate that was modeled from the real dataset (i.e., the 2011 Grade 8 NAEP mathematics assessment).

Missing data methods. Considering the purpose of our study, we considered eight major missing data treatment methods that are commonly studied and/or applied in assessment settings: IN, PM, IM, TW, RF, LR, EM, and FIML. We included PMM as well,

even though its application in assessment settings remains nominal to date; however, its robustness in handling missing data, including those in categorical data, has been supported in the literature (e.g., [83]). Additionally, random imputation (RI, i.e., replacing the missing responses with random values) was applied as the baseline of comparison. The complete study design resulted in an evaluation of ten missing data methods.

3.1.2. Factors Related to CDMs

We used the DINA model, for the reasons stated in the previous section, as both the data generation and analysis model for the current study. To make the results comparable across missing data mechanisms, we assumed a fixed test length of 35 items (J = 35), which was the number of the items in the real dataset that we used to model the MIXED mechanism. The sample size was fixed at N = 1000. Following the literature (e.g., [102,103]), the examinees' attribute profiles followed a dichotomized multivariate normal distribution, with means being zero and covariances between the specified attributes being 0.5. The number of attributes (K) and the item discrimination power were specified to vary across different levels.

Number of attributes (K). Three levels of attribute numbers (K = 3, 5, and 8) were used, resulting in three Q-matrices for data generation and CDM analysis. Additionally, the three matrices were generated by taking into account the identifiability and estimability of the DINA model (see [104,105] for details).

Item discrimination power. Two levels of item discrimination power (high and low) were considered. Specifically, under the high discrimination power condition, the item parameters (i.e., guessing and slipping) were randomly generated from a uniform distribution U (0.05, 0.25), whereas under the low discrimination power condition, the item parameters followed a uniform distribution of U (0.25, 0.45).

A fully crossed design of the study yielded a total number of 546 conditions, six conditions of which used the complete item response data while the other 540 conditions used data with missing responses. Each condition was replicated 500 times.

3.2. Data Generation

3.2.1. Generate Complete Data

For each of the six baseline conditions, a total of 500 data sets with complete dichotomous item responses were generated (N = 1000, J = 35).

3.2.2. Generate Missing Responses of MAR

Following de Ayala et al. [46] and Finch [66], the probability of examinees' missing response under MAR on a specific item was assumed to be inversely related to their observed total scores on all other items on the test. Specifically, the possible values of the observed total score (0 to 34) were first divided into seven fractiles, each with five possible scores (0–4, 5–9, . . . , 30–34). The examinees in a specific fractile were then assigned a corresponding probability of missingness. In the process, the fractiles with higher scores were assigned a lower probability, while at the same time the overall missing rates were manipulated at the pre-specified levels (i.e., MR = 5%, 10%, 20%, and 30%). Furthermore, the actual missing rates on the generated datasets were examined. Across all 500 replications, the actual missing rates ranged from 4.993% to 5.003% (SD = 0.001) for the MR = 5% condition, from 9.987% to 10.008% (SD = 0.002) for MR = 10%, from 19.993% to 20.014% (SD = 0.002) for MR = 20%, and from 29.986% to 30.007% (SD = 0.002) for MR = 30%.

3.2.3. Generate Missing Responses of MNAR

Following Finch [66], the probability of a missing response under MNAR on a specific item was assumed to be directly related to whether or not the examinee answered the item correctly in the original complete dataset. That is, those who responded to the item correctly were assigned a lower probability of missingness than those who answered the item incorrectly, while at the same time the overall missing rates were controlled. The

examination of the actual missing rates revealed that, across all replications, the actual missing rates ranged from 4.996% to 5.009% (SD = 0.001) for the MR = 5% condition, from 9.988% to 10.007% (SD = 0.002) for MR = 10%, from 19.990% to 20.002% (SD = 0.002) for MR = 20%, and from 29.990% to 30.016% (SD = 0.002) for MR = 30%.

3.2.4. Generate Missing Responses of the MIXED Mechanism

Following de Ayala et al. [46], the contingency table approach was used to generate missing responses under the MIXED type of mechanism. Specifically, data from the 2011 Grade 8 NAEP mathematics assessment booklet were used in this study. The booklet contained item responses from 4450 students on 35 items. An initial screening of the data revealed 90 examinees who did not respond to any of the items (i.e., 100% MR); these were removed from the analysis. Among the remaining responses, 34.5% did not report any missingness, 39.2% missed fewer than five items, 11.4% missed six to ten items, and 7.5% missed more than ten items. Among the 35 items, there was one item that had 0% MR, eighteen items with MR from 1% to 5%, five items from 6% to 10%, six items from 11% to 20%, and another five with MR from 20% to 50%.

The contingency table was then obtained for each item and was used to generate the missing responses. Each of the tables consisted of the information and features of the missingness, such as the number of missing responses and relative frequency of missingness compared to the number of incorrect and correct item responses (see Table S1 in the Supplementary Materials for an example of the contingency table and de Ayala et al. [46] for details of the approach). The examination of the actual missing rates after the generation process revealed that the actual missing rates of the MIXED mechanism ranged from 24.61% to 28.43% (SD = 0.002).

3.3. Analyses and Outcome

3.3.1. Analyses

The data with missing responses were treated first with selected approaches (i.e., IN, PM, IM, TW, RF, LR, EM, PMM, and RI) prior to analysis, except for FIML, which was accomplished directly in the DINA model analysis. The **TestDataImputation [106]** package in R [36] was used for the missing data treatment methods of IN, PM, IM, TW, RF, LR, EM, and PMM. The RI method was achieved by randomly replacing the missing responses from a binomial distribution with a 0.5 probability.

After the missing responses were handled, the DINA model was applied to analyze the data using the *din()* function in the R package **CDM [34]** with default settings. Both the item parameters (i.e., guessing and slipping) and individual attribute mastery profiles and patterns were obtained in order to evaluate the performance of the missing data methods. Model convergence was monitored and collected for the analyses.

3.3.2. Outcome Variables

Three outcomes, namely, model convergence, item parameter recovery, and the classification accuracy of individual attribute mastery profiles and patterns, were calculated and evaluated. The model convergence was evaluated using the rate of convergence across conditions and replications.

The item parameter recovery was investigated using both the mean bias and root mean squared error (RMSE, defined below) across replications. Specifically, for each replication,

$$bias = \frac{\sum_{j=1}^{J}(\hat{\theta}_j - \theta_j)}{J} \tag{5}$$

and

$$RMSE = \sqrt{\frac{\sum_{j=1}^{J}(\hat{\theta}_j - \theta_j)^2}{J}}. \tag{6}$$

where $\hat{\theta}_j$ and θ_j refer to the estimated and true item parameters, respectively, for item j, while J is the total number of items (J = 35). Both indices were averaged across the 500 replications and used for results reporting.

The classification accuracy was examined using both the attribute-wise classification accuracy (ACA) and pattern-wise classification accuracy (PCA, defined below):

$$ACA = \frac{1}{N \times K} \sum_{i=1}^{N} \sum_{k=1}^{K} I[\hat{\alpha}_{ik} = \alpha_{ik}]. \quad (7)$$

where $\hat{\alpha}_{ik}$ and α_{ik} refer to the estimated and true attribute profile, respectively, for examinee i on attribute k, and

$$PCA = \frac{1}{N} \sum_{i=1}^{N} I[\hat{\alpha}_i = \alpha_i]. \quad (8)$$

where $\hat{\alpha}_i$ and α_i are the estimated and true profiles for examinee i on all attributes. Similarly, both ACA and PCA were averaged across the replications and used for results reporting.

4. Results

Results are presented visually through profile plots across the manipulated factors, with three main sections: (a) model convergence; (b) item parameter recovery; and (c) attribute profile classification accuracy. Due to space limits, the complete results of bias for item parameter estimations are included in the Supplementary Materials (Figures S1–S3).

4.1. Model Convergence

The results of model convergence suggested no issues with the DINA model analyses. A convergence rate of almost 100% was accomplished across all conditions and replications. Throughout all analyses (546 conditions × 500 replications = 273,000 datasets), only twelve datasets failed to converge. Ten of these twelve datasets were associated with conditions where the missing data handling methods were RI and IM (see Table 1). Furthermore, we observed that the conditions of 30% MR and a Q-matrix with eight attributes yielded most of the non-convergence (eleven out of twelve cases); however, the overall number of models that did not converge was minimal.

Table 1. Results of model convergence.

Conditions				
Missing Mechanism	**Missing Rate**	**No. of Attributes in Q-Matrix**	**Missing Response Treatment**	**Model Convergence Rate**
MAR	20%	8	IM	0.998
	30%	8	RI	0.998
MNAR	20%	5	IM	0.998
	20%	8	IM	0.998
	30%	8	IM	0.996
	30%	8	IN	0.998
	30%	8	RI	0.992
	30%	8	PMM	0.998

Note: Number of replications = 500; MAR = missing at random, MNAR = missing not at random, IM = item mean imputation, RI = random imputation, IN = treating missing responses as incorrect, PMM = predictive mean matching.

4.2. Item Parameter Recovery

4.2.1. Item Parameter Recovery under MAR

The RMSEs of the item parameter estimations under the MAR mechanism are presented in Figure 1. Each graph in the figure represents the mean RMSEs across levels of the column factor of the number of attributes (K = 3, 5, and 8) × the row factor of item discrimination power (high and low) for both the guessing (first two rows) and slipping (last two rows) parameters. The ten lines within each graph represent the missing data handling methods. Specifically, the IN, IM, PM, TW, and RF methods are represented

by dotted lines (with different shapes), while the EM, LR, PMM, FIML, and RI methods are represented by solid lines. The two baseline-of-comparison methods, RI (solid red lines with crosses) and FIML (solid blue lines with diamonds), are color-coded for better visualization of the results.

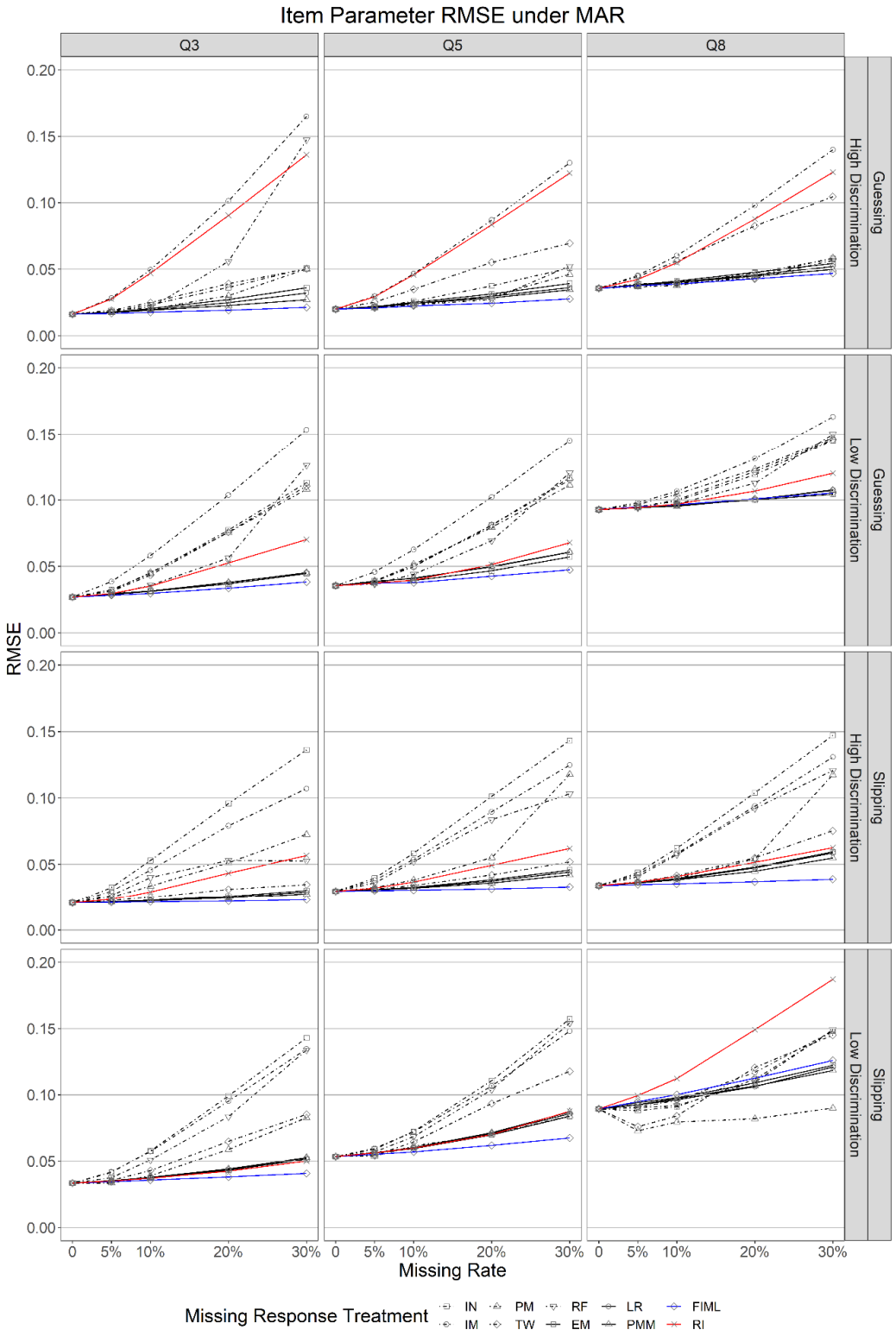

Figure 1. RMSE of item parameter estimations under the MAR mechanism. Columns Q3, Q5, and Q8 present results for Q-matrices with three, five, and eight attributes, respectively, while the rows refer to results across high and low discriminations for both guessing and slipping parameters.

The RMSEs of the guessing parameters revealed a larger impact of missing data on IN and the four mean-based methods (i.e., IM, PM, TW, and RF) as the missing rate increased (MR $\geq$ 10%), especially in the presence of low item discrimination. When the item discrimination was high (first row), IM, PM, and TW resulted in larger RMSEs than other methods. IM yielded the largest RMSEs among the methods across all conditions, and was the only method that performed worse than RI except for PM under the condition MR = 30% and K = 3. The performance of PM was not superior to other methods (except for IM) in the presence of relatively large missing rates (e.g., MR $\geq$ 20%) and K = 3. It yielded better performance as the number of attributes increased, as evidenced by RMSEs similar to other methods, especially when K = 8. TW, however, showed an opposite pattern to PM. While its performance was similar to other methods, it yielded larger RMSEs as K increased. Other methods showed similar performance, with only small RMSE discrepancies observed in the condition MR = 30% and K = 3, under which the methods FIML, PMM, LR, and EM (solid lines) performed slightly better than IN and RF (dotted lines). When item discriminations were low (second row), we noticed a clear pattern in that IN and the mean-based methods (IM, PM, TW, RF) yielded larger RMSEs than RI, especially when MR $\geq$ 20%. Other methods, including FIML, LR, PMM, and EM, showed similar performance that was stable across MR and superior to RI. Across all conditions, FIML was superior to the other methods.

The RMSEs of the slipping parameters favored FIML, PMM, LR, and EM more than IN and the mean-based methods in general. When the item discriminations were high, IN, IM, PM, TW, and RF yielded larger RMSEs than RI, except for TW in the conditions of high discrimination and $K \leq 5$. FIML, PMM, LR, and EM showed similar results, with the performance of FIML being slightly better across conditions. When item discriminations were low, FIML was the only method that yielded smaller RMSEs than RI when $K \leq 5$. When K = 8, PM yielded the smallest RMSEs, and all methods showed superior performance to RI.

4.2.2. Item Parameter Recovery under MNAR

The RMSEs of the item parameter estimations under the MNAR mechanism are presented in Figure 2. We noticed a similar pattern of IM compared to MAR conditions, when investigating the RMSEs of the guessing parameters. It remained the only method with worse performance than RI across all conditions. When item discriminations were high, TW performed similarly as under MAR; that is, while its performance was similar to other methods, it resulted in larger RMSEs as K increased, especially when K = 8. While RMSE differences were small among the rest of the methods, IN, PM, and RF performed slightly better than FIML, LR, PMM, and EM. When item discriminations were low, all methods except for IM showed very similar RMSEs that were smaller than RI across all conditions.

The results of the slipping parameters under MNAR were similar to those under MAR. When item discriminations were high, IN and all mean-based methods showed worse performance than RI, except for TW in conditions $K \leq 5$ and PM in conditions K = 3 and MR $\leq$ 20%. When item discriminations were low, RI yielded the smallest RMSEs among the methods when $K \leq 5$. When K = 8, all methods except IM showed superior performance to RI.

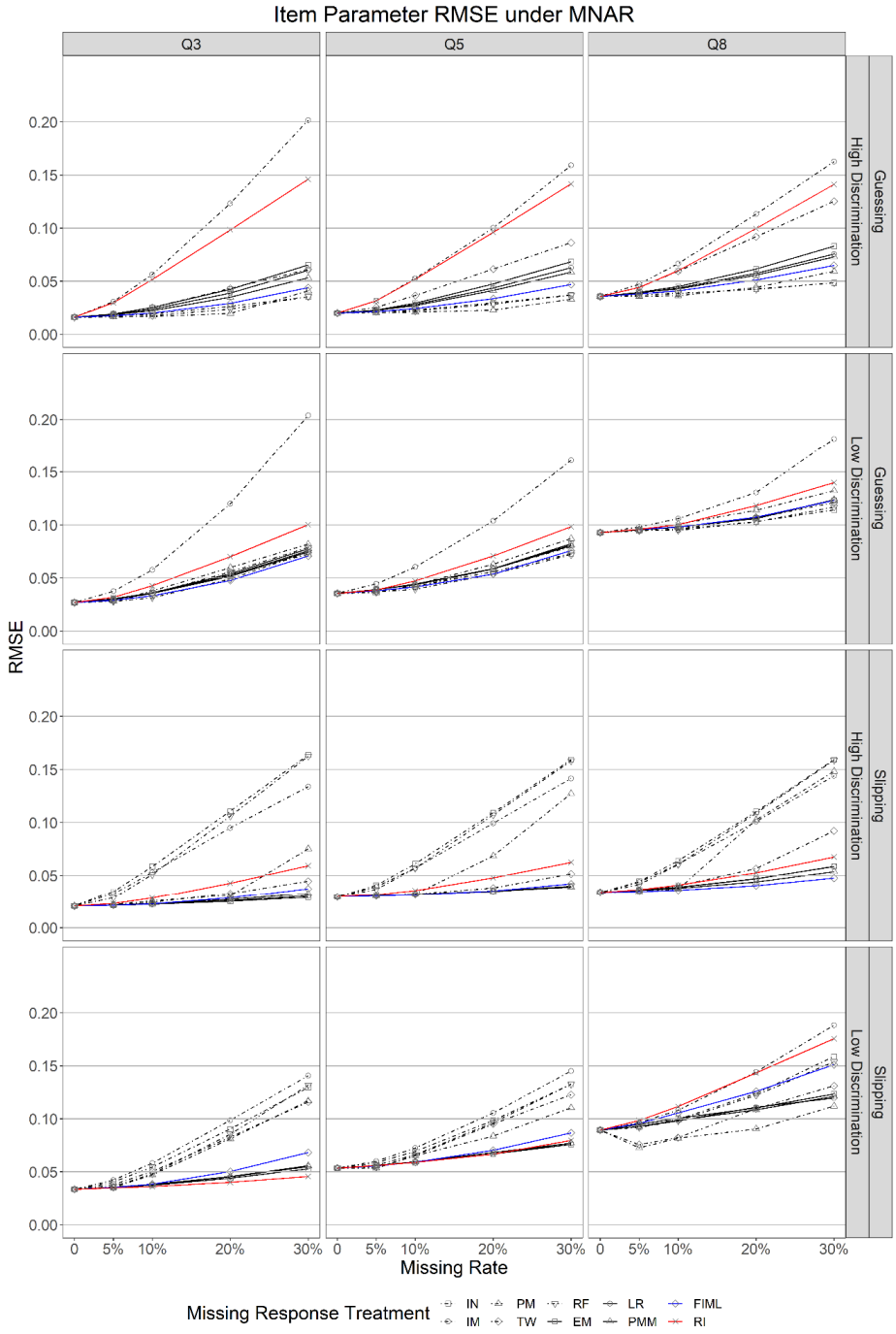

Figure 2. RMSE of item parameter estimations under the MNAR mechanism. Columns Q3, Q5, and Q8 present results for Q-matrices with three, five, and eight attributes, respectively, while rows refer to results across high and low discriminations for both guessing and slipping parameters.

4.2.3. Item Parameter Recovery under the MIXED Mechanism

The RMSEs of the item parameter estimations under the MIXED mechanism are shown in Figure 3. A clear pattern was observed for the slipping and guessing parameters. Namely, random imputation was superior to other methods when item discriminations were low. For items with high discrimination, the results differed for the guessing and slipping parameters. Generally, LR, FIML, PMM, and EM yielded better performance than other methods in recovering the guessing parameters except for RF. RF yielded larger RMSEs than LR, FIML, PMM, and EM when $K = 3$, while showing better performance than most other methods (all except LR), with smaller RMSEs when $K \geq 5$. Additionally, IM

performed worse than RI when $K \leq 5$, whereas when $K = 8$, there were four methods (IM, TW, PM, and IN) that yielded larger RMSEs than RI. Across all conditions, LR showed slightly better performance than the other methods. Regarding the recovery of the slipping parameters when item discriminations were high, there were only two treatment methods (LR and FIML) that showed consistently acceptable performance (i.e., smaller RMSEs than RI) across all of the conditions.

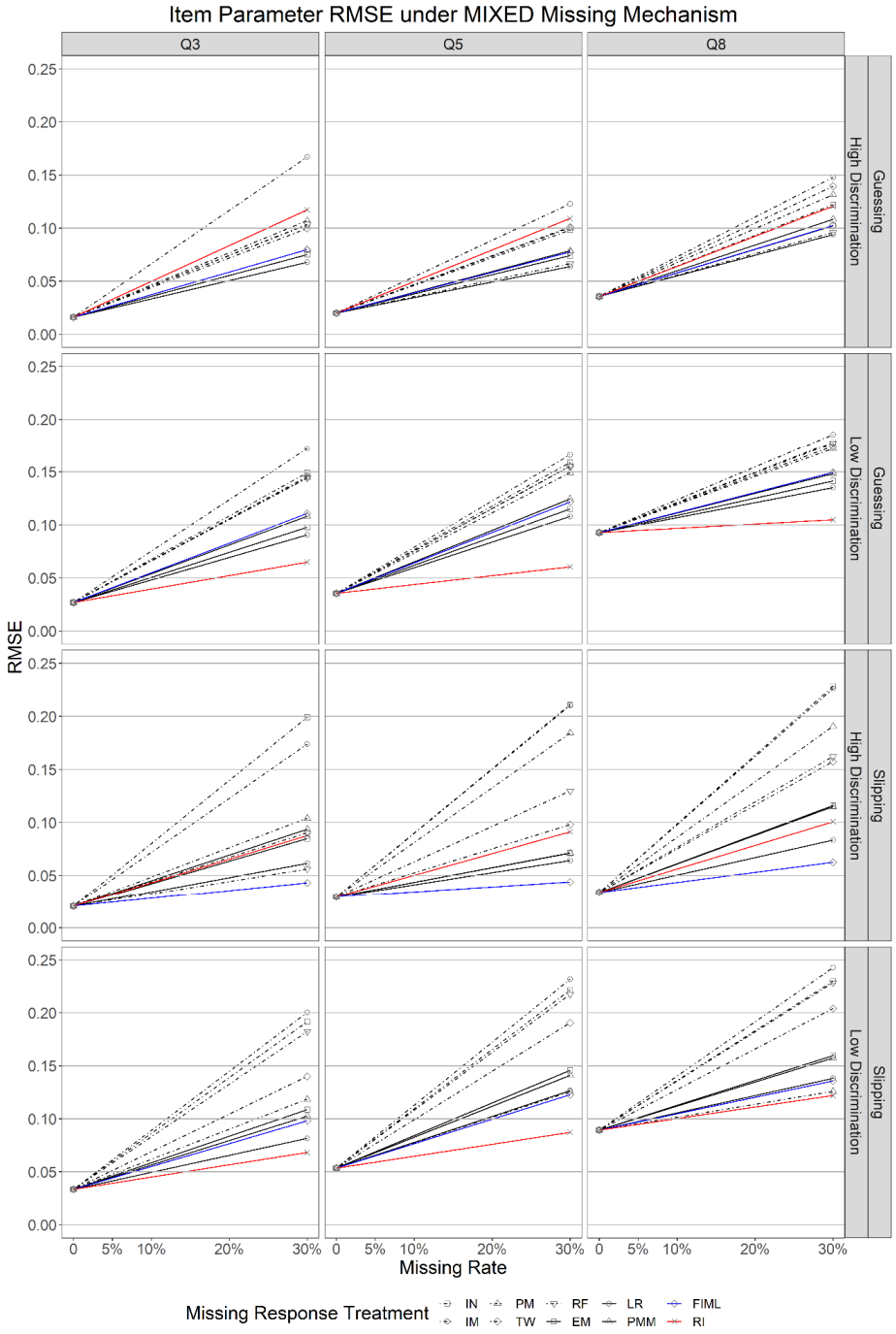

Figure 3. RMSE of item parameter estimations under the MIXED missing mechanism. Columns Q3, Q5, and Q8 present results for Q-matrices with three, five, and eight attributes, respectively, while rows refer to results across high and low discriminations for both guessing and slipping parameters.

4.3. Attribute Classification Accuracy

4.3.1. Classification Accuracy under MAR

The results for the mean ACAs (first two rows) and PCAs (last two rows) under the MAR mechanism are presented in Figure 4. We observed that the impact of missing response proportions (i.e., MR) on classification accuracy remained very small across all conditions, especially for ACAs with MR $\leq$ 10%. The ACA discrepancies became observable as MR increased ($\geq$ 20%), especially with K = 3 and high item discrimination. Specifically, EM showed slightly better performance, while PM yielded the lowest accuracy rates. Such differences, however, were very small. All selected methods produced higher ACAs than did RI across all conditions, except for PM when K = 3 and MR = 30%. The missingness showed a larger impact on PCAs, especially when item discrimination was high. FIML, PMM, LR, and EM, especially FIML, showed better performance than other methods across all conditions. The PCAs of all selected methods, however, were larger than those of RI, except for PM. PM yielded worse performance than RI across most of the conditions, especially in the presence of low item discrimination and a 30% MR.

4.3.2. Classification Accuracy under MNAR

The results of both mean ACAs and PCAs under MNAR are presented in Figure 5. We observed almost identical patterns from the results across the two missing mechanisms of MAR and MNAR. While both ACAs and PCAs decreased as the MR increased, the changes in ACAs were more stable across MR, and ACA differences remained small across the methods. In general, FIML performed slightly better than other methods, whereas PM yielded the lowest accuracy rates across most of the conditions.

4.3.3. Classification Accuracy under the MIXED Mechanism

The results of the attribute classification accuracy analysis revealed a different pattern under the MIXED missing mechanism (see Figure 6). The ACA results suggested that FIML was the only missing data approach superior to RI when K = 3, regardless of item discrimination level. As the number of attributes increased (i.e., $K \geq 5$), however, all methods yielded very similar ACAs that were larger than those generated by RI. An exception was PM, which showed lower ACAs than RI when K = 8. Differences in PCAs were more obvious across the methods. The patterns of the results, however, were similar across the two measures of classification accuracy.

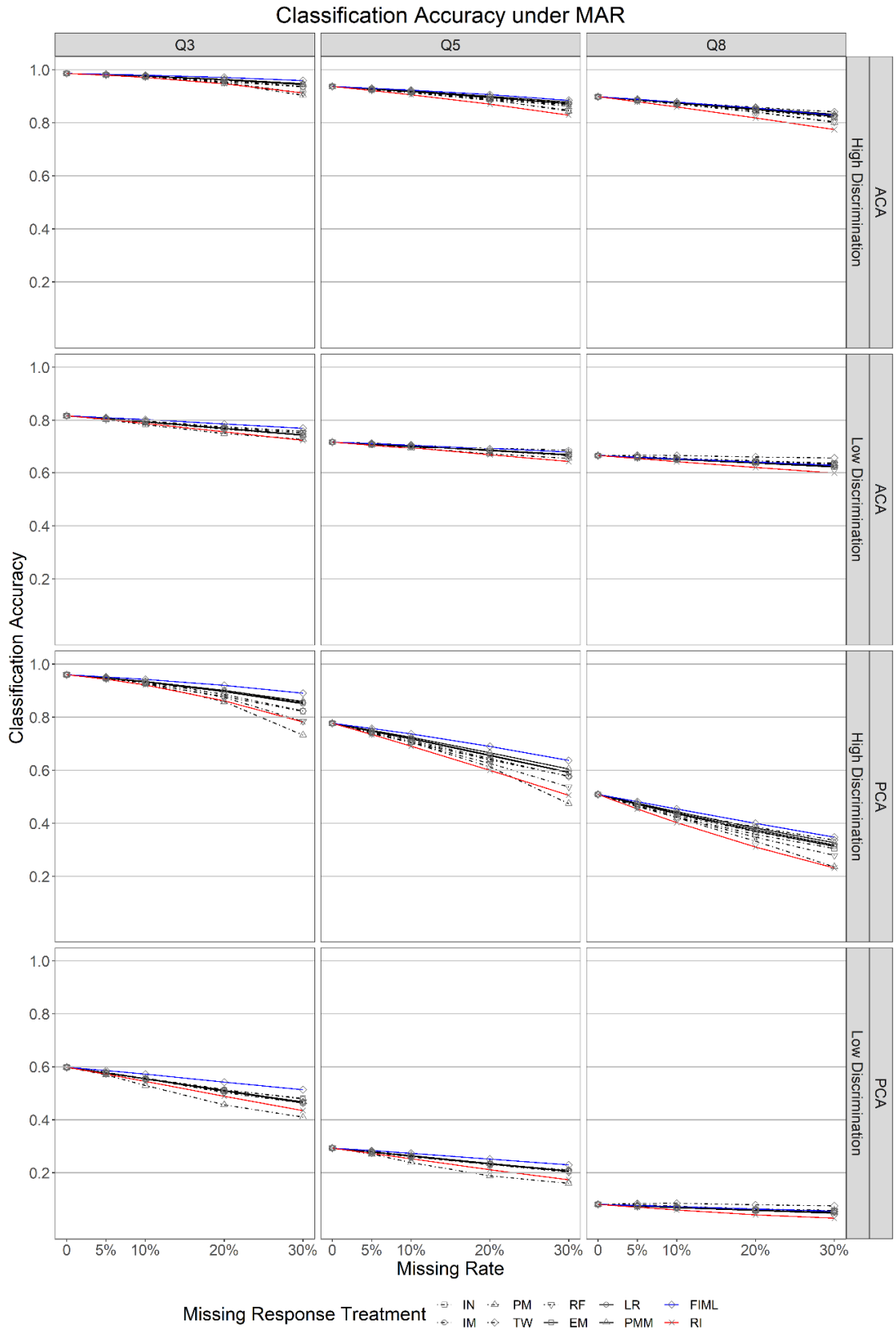

Figure 4. Classification accuracy under the MAR mechanism. Columns Q3, Q5, and Q8 present results for Q-matrices with three, five, and eight attributes, respectively, while rows refer to results across high and low discriminations for both guessing and slipping parameters.

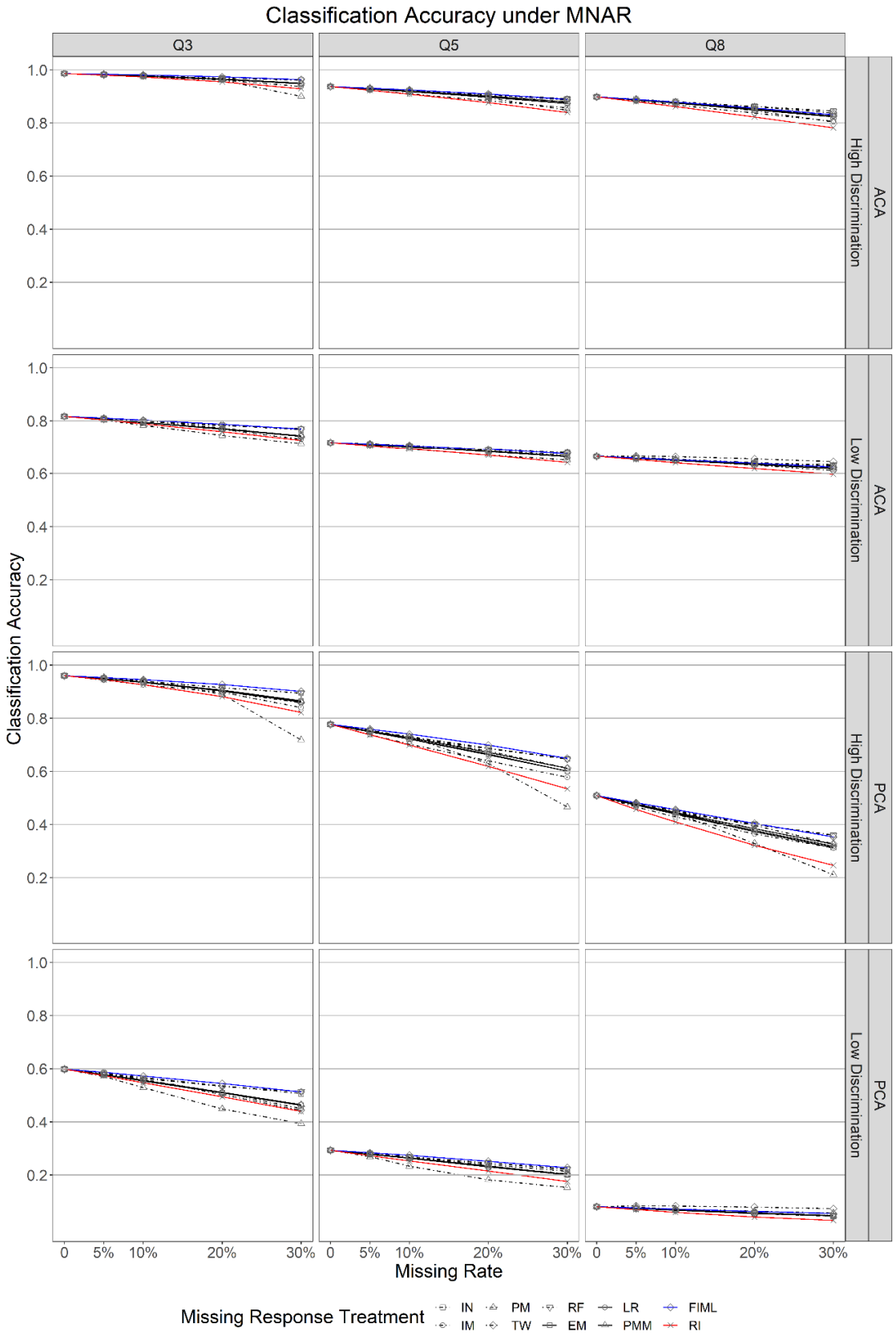

Figure 5. Classification accuracy under the MNAR mechanism. Columns Q3, Q5, and Q8 present results for Q-matrices with three, five, and eight attributes, respectively, while rows refer to results across high and low discriminations for both guessing and slipping parameters.

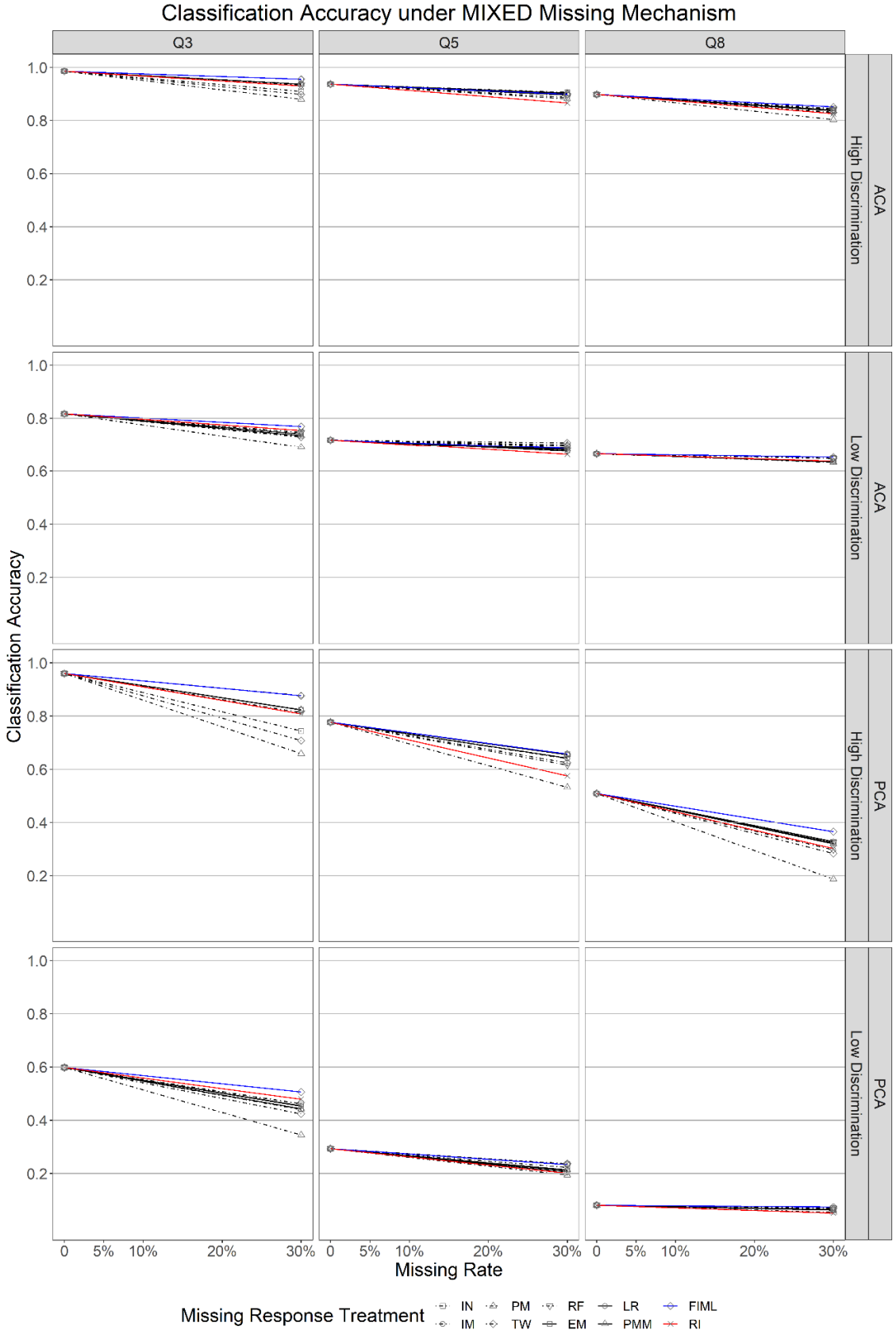

Figure 6. Classification accuracy under the MIXED missing mechanism. Columns Q3, Q5, and Q8 present results for Q-matrices with three, five, and eight attributes, respectively, while rows refer to results across high and low discriminations for both guessing and slipping parameters.

5. Discussion

5.1. Findings

In this simulation study, we investigated the impact of missing responses and evaluated ten missing data handling methods, namely, IN, PM, IM, TW, RF, LR, EM, FIML, PMM, and RI, in the implementation of CDMs across various conditions. In particular, we evaluated the performance of the selected methods for the DINA model under three different missing data mechanisms (MAR, MNAR, and MIXED), five missing rates (MR = 0%, 5%, 10%, 20%, and 30%), three Q-matrix sizes (K = 3, 5, and 8), and two item discrimination classes (high and low).

The results suggest that the presence of missing data, even when the missing rate was only 5%, affected the estimation of both item parameters and classification of attribute profiles. Larger missing rates yielded higher RMSEs for both the guessing and slipping parameters, along with smaller ACAs and PCAs for estimating attribute profiles. The performance of different missing data treatment methods was impacted by the proportion of missingness. The differences among the methods became larger as the missing rate increased, especially when the miss rate was 20% or higher.

A closer look at the results suggested that the magnitude of the RMSEs in most conditions was relatively small for item parameter estimates, especially in conditions with a smaller missing rate, high item discrimination, and a smaller number of attributes in the Q-matrix. The RMSEs ranged from 0.016 (MR = 0%, K = 3, high item discrimination, EM imputation) to 0.243 (MR = 30%, K = 8, low item discrimination, IM imputation) across conditions. The relatively small magnitude of the RMSEs for the item parameters under the DINA model indicated that the impact of missing data and the performance of the different methods handling it appeared to be small when taken at face value. The consequence of these values, however, can be rather large, for several reasons. First, both the guessing and slipping parameters are defined as error probabilities in the DINA model. In particular, the guessing parameter of an item denotes the probability of a student answering an item correctly when at least one required attribute is not present, while the slipping parameter denotes the probability of a student answering an item incorrectly when they possess all of the required attributes (see Section 2.1 for more details about the DINA model). In addition, there is an order constraint assumption applied to both item parameters in the DINA model [4]. The order constraint, which is analogous to the assumption of monotonicity in other latent variable models, assumes that $1 - s_j > g_j$ must hold for the DINA model. This ensures that a student who possesses all required attributes has a higher probability of answering an item correctly than a student who has not mastered all required attributes. Compared to the usually unbounded parameters (e.g., those following a normal distribution) in other latent variable models, the smaller ranges of both the guessing and slipping parameters make it possible, along with the order constraint, to obtain relatively small bias and RMSEs, especially when item parameters are large (i.e., when item discrimination is low). This explains why the magnitude of the RMSEs was larger in a few conditions with high item discrimination than in conditions with low item discrimination.

Our results revealed that the various methods performed inconsistently across the different levels of design factors and under different missing data mechanisms, especially IN and the mean-based methods (i.e., PM, IM, TW, and RF), in recovering the item parameters. No single method was found to be superior to other methods across all conditions, which is somewhat similar to previous results in the IRT literature, e.g., in [66].

Overall, the item parameter recovery results favored FIML, PMM, EM, and LR. These methods showed stable performance and yielded relatively lower RMSEs than other methods and random imputation across most conditions, except for the MIXED missing mechanism and low item discrimination. IM performed worse than random imputation in nearly all conditions. The remaining methods, including IN, PM, TW, and RF, showed acceptable performance (i.e., they yielded smaller RMSEs than random imputation) in recovering the guessing parameters under MAR and MNAR, as well as when the item discrimination

level was high. They were able to outperform FIML, PMM, EM, and LR under MNAR, although only under high item discrimination conditions. Additionally, the performance of TW decreased as the number of attributes in the Q-matrix increased, whereas PM showed the opposite trend.

While the impact of the missing rate was visually obvious from the figures, the results revealed the influence of both item quality (i.e., item discrimination level) and missing mechanisms on item parameter estimations, especially under the mixed type of missing mechanism. For instance, with the mixed mechanism and low item discrimination, none of the methods showed superior performance to random imputation.

The performance of the different methods for attribute profile estimation was evaluated using both ACA and PCA. ACA denotes the agreement of the estimated and the true profiles at the attribute level (i.e., elementwise), whereas PCA indicates the agreement at the pattern/profile level (i.e., row-wise). Values of PCA were largely affected by low item discrimination and large numbers of attributes in the Q-matrix. Extremely small PCAs were obtained in conditions with low item discrimination and $K = 8$. For example, PCAs ranged from 3% to 8% across all MR in conditions with low item discrimination and $K = 8$ under MAR. The trivial values of PCA in such conditions made it less meaningful to compare the performance of the methods. Furthermore, in conditions where PCA had acceptable values, similar patterns in the performance of the various methods were detected for both ACA and PCA; therefore, our summarization and discussion of their performance is mainly based on the results of ACA.

Our investigation of the results reveals that ACA values decreased by various degrees as MR increased across nearly all conditions, especially when MR $\geq$ 20%. The presence of low item discrimination and a larger number of attributes in the Q-matrix resulted in lower ACAs and larger differences among the methods in terms of ACAs, especially when item discrimination was low. Nevertheless, similar patterns of performance were found across all levels of the three design factors and missing data mechanisms; that is, the performance of the methods was very similar and all were acceptable as compared to random imputation, except for PM, which performed slightly worse than random imputation in several conditions (e.g., MAR, $K = 3$, MR = 30%). There were only very small differences among the methods across all conditions, with FIML, PMM, LR, and EM performing slightly better than other methods in several conditions.

5.2. Recommendations

In considering which of the methods should be used when dealing with missing data, it is important to note that the methods performed differently with respect to the recovery of item parameter estimates and classification of attribute profiles. Although complex, it is suggested that different methods be used to achieve more accurate item parameters and attribute profiles, respectively. This is reasonable in the implementation of CDMs, as item parameters and attribute profiles usually serve different purposes. Researchers and practitioners who intend to provide diagnostic information for students and teachers might impute the missing data using a method that favors attribute profile estimations, whereas those tasked with item design and test development may select a method that yields more accurate item parameters.

Regarding the performance of the methods for item parameters, there is no definite evidence that a certain method is always superior to others. Considering that the two item parameters are usually estimated simultaneously in the DINA model, the performance of the methods on both parameters across design factors should be considered when selecting a method to handle missing data. Although there is not an optimal method that can be applied to all conditions, the results of the present study yield several general suggestions. First, IM is not optimal across any conditions and is never recommended, even in the presence of a relatively small missing rate (e.g., MR $\leq$ 5%). Second, with MR $\leq$ 5% (or even MR $\leq$ 10%), selection among the remaining methods makes little difference. Third, when MR $\geq$ 10%, the choice of a method from among FIML, PMM, LR, and EM is recommended.

With respect to the performance of the methods for attribute profiles, any of the methods except for PM is acceptable, regardless of the missing rate. PM is not recommended across any conditions and missing data mechanisms, especially with a higher MR.

5.3. Limitation and Future Directions

As with all simulation studies, generalizations of the findings in the current research are limited due to design choices. First, only the DINA model was considered in this study. Although it has been widely used in the literature, there are many other CDMs that are available, such as the GDM and LCDM models. Future research should carry this line of inquiry further and examine the impact of missing data in the context of CDMs using more generalized models.

Second, in the study we assumed that all Q-matrices were correctly specified. In practice, however, it is possible that a Q-matrix might be mis-specified. A mis-specified Q-matrix may lead to invalid classification of students' attribute mastery status and/or biased item parameters [5,7,44]. Future studies should take this into account.

Third, we only considered the overall level outcomes to evaluate the performance of selected methods for handling missing data. For example, overall ACAs were used to compare the performance of the methods in terms of attribute profile estimation. As a consequence, our conclusions and recommendations are made at the overall level. Further research is needed to examine the impact of missing responses on diagnostic inferences at the individual level.

Fourth, this study investigated the missing data problem with a fixed sample size (i.e., $N = 1000$) and test length (i.e., $J = 35$). Our intention was to place this study in a context that is similar to many current assessment forms, especially major large-scale assessments, in which students are usually administered a booklet of items that are of a similar size. In the future, the missing data issue should be addressed in wider assessment settings as well.

Additionally, we only examined the performance of single imputation methods. Considering the increasing attention on MI and model-based approaches for addressing missing data issues in psychometrics, the performance of such methods should be investigated in the context of CDMs.

Finally, the generalizability of the current simulation study is limited by the fact that the data were generated based on the specified design factors and their corresponding levels. In empirical settings, however, it is challenging to find a dataset that meets the specified assumptions. Although we included three types of missing mechanisms (i.e., MAR, MNAR, and MIXED) in this study, the way the missing responses were generated imposes limitations. For instance, the MAR data were generated based on examinees' total scores on all other items on the test, not on other variables, the results of which may favor methods that assume dependency between missing propensity and ability (e.g., IN and model-based methods). Similarly, the MNAR mechanism was generated based on item responses, which may introduce bias.

Furthermore, simulation studies usually only consider general features of the assessment, such as test length and the number of attributes in the Q-matrix, instead of specific features such as assessment types (e.g., digital vs. paper-based assessments), item formats (multiple choice vs. constructed response items) and content or knowledge domains (e.g., mathematics vs. reading literacy). Therefore, investigations into missing data issues using empirical data should be considered in future research.

6. Conclusions

The findings of this study provide guidelines for considering which of the imputation approaches to missing responses is best for use in the framework of CDMs across multiple assessment settings. CDMs are popular models thanks to their advantages in supporting construct validation, informing diagnostic test development, and providing detailed diagnostic information that assists in learning and instruction in the framework of cognitive diagnostic assessments (CDAs). In practice, due to the lack of CDAs, most applications

of these models involve retrofitting existing assessments, and are primarily focused on diagnostic score reporting rather than construct validation and test development. Both types of applications rely on accurate estimation of parameters from CDMs. With missing responses treated in a more appropriate way, more accurate diagnostic information from CDMs can serve as a better tool to help teachers, and ultimately their students, to use diagnostic information more effectively for future learning outcomes, monitor, evaluate, and improve educational systems, and inform reforms and policy changes in education, particularly in the area of curriculum development and/or instructional sensitivity.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/psych4020028/s1, Table S1: Example Contingency Table of an Item for Generating the MIXED Type of Missingness; Figure S1: Bias of item parameter estimations under the MAR mechanism; Figure S2: Bias of item parameter estimations under the MNAR mechanism; Figure S3: Bias of item parameter estimations under the MIXED missing mechanism.

Author Contributions: Conceptualization, S.D. and D.S.V.; methodology, S.D. and D.S.V.; formal analysis, S.D.; writing—original draft preparation, S.D.; writing—review and editing, S.D. and D.S.V.; visualization, S.D.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data is contained within this article and its supporting information. The data that support the findings of the simulation study are available from the authors upon reasonable request. Analysis using NAEP data was conducted under a restricted data license granted by the National Center of Educational Statistics of the United States Institute of Education Sciences (IES), and the disclosure of the data information in the paper has been approved by the IES Data Security Office.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. De la Torre, J. DINA model and parameter estimation: A didactic. *J. Educ. Behav. Stat.* **2009**, *34*, 115–130. [CrossRef]
2. Templin, J.L.; Henson, R.A. Measurement of psychological disorders using cognitive diagnosis models. *Psychol. Methods* **2006**, *11*, 287–305. [CrossRef] [PubMed]
3. Bradshaw, L.; Izsák, A.; Templin, J.; Jacobson, E. Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educ. Meas. Issues Pract.* **2014**, *33*, 2–14. [CrossRef]
4. Rupp, A.A.; Templin, J.; Henson, R.A. *Diagnostic Measurement: Theory, Methods, and Applications*; Guilford Press: New York, NY, USA, 2010.
5. Rupp, A.A.; Templin, J. The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educ. Psychol. Meas.* **2008**, *68*, 78–96. [CrossRef]
6. Li, X.; Wang, W.-C. Assessment of differential item functioning under cognitive diagnosis models: The DINA model example. *J. Educ. Meas.* **2015**, *52*, 28–54. [CrossRef]
7. Junker, B.W.; Sijtsma, K. Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Meas.* **2001**, *25*, 258–272. [CrossRef]
8. DiBello, L.V.; Stout, W.F.; Roussos, L.A. Unified cognitive/psychometric diagnostic assessment Likelihood-based classification techniques. In *Cognitively Diagnostic Assessment*; Routledge: London, UK, 1995; pp. 361–389.
9. Henson, R.A.; Templin, J.L.; Willse, J.T. Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika* **2009**, *74*, 191–210. [CrossRef]
10. Von Davier, M. A General diagnostic model applied to language testing data. *ETS Res. Rep. Ser.* **2005**, *2005*, i-35. Available online: https://files.eric.ed.gov/fulltext/EJ1111422.pdf (accessed on 23 March 2016). [CrossRef]
11. Roussos, L.A.; DiBello, L.V.; Stout, W.; Hartz, S.M.; Henson, R.A.; Templin, J.L. The fusion model skills diagnosis system. In *Cognitive Diagnostic Assessment for Education: Theory and Applications*; Leighton, J.P., Gierl, M.J., Eds.; Cambridge University Press: Cambridge, UK, 2007; pp. 275–318.
12. DiBello, L.V.; Roussos, L.A.; Stout, W. Review of cognitively diagnostic assessment and a summary of psychometric models. *Handb. Stat. Psychom.* **2007**, *26*, 979–1030.
13. Leighton, J.; Gierl, M. *Cognitive Diagnostic Assessment for Education: Theory and Applications*; Cambridge University Press: Cambridge, UK, 2007.

14. Nichols, P.D. A framework for developing cognitively diagnostic assessments. *Rev. Educ. Res.* **1994**, *64*, 575–603. [CrossRef]
15. Rupp, A.A.; Templin, J.L. Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement* **2008**, *6*, 219–262. [CrossRef]
16. Snow, R.E.; Lohman, D.F. Implications of cognitive psychology for educational measurement. In *Educational Measurement*; Linn, R., Ed.; American Council on Education/Macmillan: New York, NY, USA, 1989; pp. 263–331.
17. Xu, X.; von Davier, M. Cognitive diagnosis for NAEP proficiency data. *ETS Res. Rep. Ser.* **2006**, *2006*, i-25. Available online: https://files.eric.ed.gov/fulltext/EJ1111414.pdf (accessed on 23 March 2016). [CrossRef]
18. Embretson, S. Applications of cognitive design systems to test development. In *Cognitive Assessment: A Multidisciplinary Perspective*; Reynolds, C.R., Ed.; Springer: New York, NY, USA, 1994; pp. 107–135.
19. Embretson Whitely, S.E. Construct validity: Construct representation versus nomothetic span. *Psychol. Bull.* **1983**, *93*, 179–197. [CrossRef]
20. Huff, K.; Goodman, D.P. The demand for cognitive diagnostic assessment. In *Cognitive Diagnostic Assessment for Education: Theory and Applications*; Leighton, J.P., Gierl, M.J., Eds.; Cambridge University Press: Cambridge, UK, 2007; pp. 19–60.
21. Lee, Y.-W.; Sawaki, Y. Application of three cognitive diagnosis models to ESL reading and listening assessments. *Lang. Assess. Q.* **2009**, *6*, 239–263. [CrossRef]
22. Lee, Y.-W.; Sawaki, Y. Cognitive diagnosis and Q-matrices in language assessment. *Lang. Assess. Q.* **2009**, *6*, 169–171. [CrossRef]
23. Lee, Y.-W.; Sawaki, Y. Cognitive diagnosis approaches to language assessment: An overview. *Lang. Assess. Q.* **2009**, *6*, 172–189. [CrossRef]
24. Leighton, J.P.; Gierl, M.J. Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educ. Meas. Issues Pract.* **2007**, *26*, 3–16. [CrossRef]
25. Mislevy, R.J.; Steinberg, L.S.; Breyer, F.J.; Almond, R.G.; Johnson, L. A cognitive task analysis, with implications for designing a simulation-based performance assessment. *Comput. Hum. Behav.* **1998**, *15*, 335–374. [CrossRef]
26. Tatsuoka, K.K. Analysis of errors in fraction addition and subtraction problems. In *Computer-Based Education Research Laboratory Report*; University of Illinois: Urbana, IL, USA, 1984. Available online: https://files.eric.ed.gov/fulltext/ED257665 (accessed on 23 March 2016).
27. Ravand, H.; Baghaei, P. Diagnostic classification models: Recent developments, practical issues, and prospects. *Int. J. Test.* **2020**, *20*, 24–56. [CrossRef]
28. De la Torre, J. The generalized DINA model framework. *Psychometrika* **2011**, *76*, 179–199. [CrossRef]
29. Huo, Y.; de la Torre, J. Estimating a cognitive diagnostic model for multiple strategies via the EM algorithm. *Appl. Psychol. Meas.* **2014**, *38*, 464–485. [CrossRef]
30. Chiu, C.-Y. Statistical refinement of the Q-matrix in cognitive diagnosis. *Appl. Psychol. Meas.* **2013**, *37*, 598–618. [CrossRef]
31. De la Torre, J. An empirically based method of Q-matrix validation for the DINA model: Development and applications. *J. Educ. Meas.* **2008**, *45*, 343–362. [CrossRef]
32. Hou, L.; la Torre, J.D.; Nandakumar, R. Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate dif in the DINA model. *J. Educ. Meas.* **2014**, *51*, 98–125. [CrossRef]
33. Svetina, D.; Feng, Y.; Paulsen, J.; Valdivia, M.; Valdivia, A.; Dai, S. Examining DIF in the context of CDMs when the Q-matrix is misspecified. *Front. Psychol.* **2018**, *696*, 1–15. [CrossRef] [PubMed]
34. Robitzsch, A.; Kiefer, T.; George, A.C.; Uenlue, A. CDM: Cognitive Diagnosis Modeling. R Package Version 7.5-15. 2022. Available online: https://CRAN.R-project.org/package=CDM (accessed on 8 May 2022).
35. Ma, W.; de la Torre, J. GDINA: The Generalized DINA Model. Framework. 2022. Available online: https://cran.r-project.org/package=GDINA (accessed on 8 May 2022).
36. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2022.
37. Jang, E.E. Demystifying a Q-Matrix for making diagnostic inferences about L2 reading skills. *Lang. Assess. Q.* **2009**, *6*, 210–238. [CrossRef]
38. Sawaki, Y.; Kim, H.-J.; Gentile, C. Q-matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Lang. Assess. Q.* **2009**, *6*, 190–209. [CrossRef]
39. Effatpanah, F.; Baghaei, P.; Boori, A.A. Diagnosing EFL learners' writing ability: A diagnostic classification modeling analysis. *Lang. Test. Asia* **2019**, *9*, 12. [CrossRef]
40. Jurich, D.; Bradshaw, L. An illustration of diagnostic classification modeling in student learning outcomes assessment. *Int. J. Test.* **2014**, *14*, 49–72. [CrossRef]
41. Lee, Y.-S.; Park, Y.S.; Taylan, D. A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. *Int. J. Test.* **2011**, *11*, 144–177. [CrossRef]
42. Mei, H.; Chen, H. Assessing students' translation competence: Integrating China's standards of English with cognitive diagnostic assessment approaches. *Front. Psychol.* **2022**, *13*, 872025. [CrossRef] [PubMed]
43. Park, Y.S.; Lee, Y.-S. An extension of the DINA model using covariates examining factors affecting response probability and latent classification. *Appl. Psychol. Meas.* **2014**, *38*, 376–390. [CrossRef]
44. Svetina, D.; Gorin, J.S.; Tatsuoka, K.K. Defining and comparing the reading comprehension construct: A cognitive-psychometric modeling approach. *Int. J. Test.* **2011**, *11*, 1–23. [CrossRef]

45. Birenbaum, M.; Tatsuoka, C.; Yamada, T. Diagnostic assessment in TIMSS-R: Between-countries and within-country comparisons of eighth graders' mathematics performance. *Stud. Educ. Eval.* **2004**, *30*, 151–173. [CrossRef]
46. De Ayala, R.J.; Plake, B.S.; Impara, J.C. The impact of omitted responses on the accuracy of ability estimation in item response theory. *J. Educ. Meas.* **2001**, *38*, 213–234. [CrossRef]
47. Schafer, J.L.; Graham, J.W. Missing data: Our view of the state of the art. *Psychol. Methods* **2002**, *7*, 147–177. [CrossRef]
48. Little, R.J.; Rubin, D.B. The analysis of social science data with missing values. *Sociol. Methods Res.* **1989**, *18*, 292–326. [CrossRef]
49. Little, R.J.; Rubin, D.B. *Statistical Analysis with Missing Data*, 3rd ed.; Wiley Series in Probability and Statistics; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2020.
50. Collins, L.M.; Schafer, J.L.; Kam, C.-M. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol. Methods* **2001**, *6*, 330–351. [CrossRef]
51. Pohl, S.; Gräfe, L.; Rose, N. Dealing with omitted and not-reached items in competence tests evaluating approaches accounting for missing responses in item response theory models. *Educ. Psychol. Meas.* **2014**, *74*, 423–452. [CrossRef]
52. Lord, F.M. *Applications of Item Response Theory to Practical Testing Problems*; Routledge: New York, NY, USA, 1980.
53. IEA. *Methods and Procedures: TIMSS 2019 Technical Report*; Martin, M.O., von Davier, M., Mullis, I.V.S., Eds.; TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA): Boston, MA, USA, 2020.
54. Shan, N.; Wang, X. Cognitive diagnosis modeling incorporating item-level missing data mechanism. *Front. Psychol.* **2020**, *11*, 1–11. [CrossRef]
55. Dai, S.; Svetina, D.; Chen, C. Investigation of missing responses in Q-matrix validation. *Appl. Psychol. Meas.* **2018**, *42*, 660–676. [CrossRef] [PubMed]
56. Sünbül, S.Ö. The impact of different missing data handling methods on DINA model. *Int. J. Eval. Res. Educ.* **2018**, *7*, 77–86.
57. Gorin, J.S. Test construction and diagnostic testing. In *Cognitive Diagnostic Assessment for Education: Theory and Applications*; Leighton, J.P., Gierl, M.J., Eds.; Cambridge University Press: New York, NY, USA, 2007; pp. 173–201.
58. Henson, R.; Douglas, J. Test construction for cognitive diagnosis. *Appl. Psychol. Meas.* **2005**, *29*, 262–277. [CrossRef]
59. Jang, E.E. Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for fusion model application to language assessment. *Lang. Test.* **2009**, *26*, 31–73. [CrossRef]
60. Embretson, S.; Gorin, J. Improving Construct Validity with Cognitive Psychology Principles. *J. Educ. Meas.* **2001**, *38*, 343–368. [CrossRef]
61. Chen, Y.; Liu, J.; Xu, G.; Ying, Z. Statistical analysis of Q-matrix based diagnostic classification models. *J. Am. Stat. Assoc.* **2015**, *110*, 850–866. [CrossRef]
62. Hartz, S.M. A Bayesian Framework for the Unified Model for Assessing Cognitive Abilities: Blending Theory with Practicality. Ph.D. Thesis, The University of Illinois at Urbana-Champaign, Champaign, IL, USA, 2002.
63. Huebner, A. An overview of recent developments in cognitive diagnostic computer adaptive assessments. *Pract. Assess. Res. Eval.* **2010**, *15*, 3.
64. Von Davier, M. The DINA model as a constrained general diagnostic model: Two variants of a model equivalency. *Br. J. Math. Stat. Psychol.* **2014**, *67*, 49–71. [CrossRef]
65. Rubin, D.B. Inference and missing data. *Biometrika* **1976**, *63*, 581–592. [CrossRef]
66. Finch, H. Estimation of item response theory parameters in the presence of missing data. *J. Educ. Meas.* **2008**, *45*, 225–245. [CrossRef]
67. Cheema, J.R. Some general guidelines for choosing missing data handling methods in educational research. *J. Mod. Appl. Stat. Methods* **2014**, *13*, 53–75. [CrossRef]
68. Mislevy, R.J.; Wu, P.K. Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing. *ETS Res. Rep. Ser.* **1996**, *1996*, i-36. [CrossRef]
69. Brown, N.J.S.; Dai, S.; Svetina, D. Predictors of omitted responses on the 2009 National Assessment of Educational Progress (NAEP) mathematics assessment. In Proceedings of the Annual Meeting of the American Educational Research Association, Philadelphia, PA, USA, 3–7 April 2014.
70. Sportisse, A.; Boyer, C.; Josse, J. Imputation and low-rank estimation with missing not at random data. *Stat. Comput.* **2020**, *30*, 1629–1643. [CrossRef]
71. Robitzsch, A. On the treatment of missing item responses in educational large-scale assessment data: An illustrative simulation study and a case study using PISA 2018 mathematics data. *Eur. J. Investig. Health Psychol. Educ.* **2021**, *11*, 1653–1687. [CrossRef]
72. Huisman, M.; Molenaar, I.W. Imputation of missing scale data with item response models. In *Essays on Item Response Theory*; Boomsma, A., Duijn, M.A.J., Snijders, T.A.B., Eds.; Springer: New York, NY, USA, 2001; pp. 221–244.
73. College Board Understanding Your Score Report. Available online: https://satsuite.collegeboard.org/media/pdf/understanding-your-sat-score-report.pdf (accessed on 4 June 2022).
74. Pohl, S.; Becker, B. Performance of missing data approaches under nonignorable missing data conditions. *Methodology* **2020**, *16*, 147–165. [CrossRef]
75. Rose, N.; Davier, M.; Xu, X. Modeling nonignorable missing data with item response theory (IRT). *ETS Res. Rep. Ser.* **2010**, *2010*, i-53. Available online: https://files.eric.ed.gov/fulltext/ED523925.pdf (accessed on 24 March 2016). [CrossRef]

76. Lord, F.M. Quick estimates of the relative efficiency of two tests as a function of ability level. *J. Educ. Meas.* **1974**, *11*, 247–254. [CrossRef]
77. Lord, F.M. Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika* **1983**, *48*, 233–245. [CrossRef]
78. Mislevy, R.J.; Wu, P.K. Inferring examinee ability when some item responses are missing. *ETS Res. Rep. Ser.* **1988**, *1988*, i-75. Available online: https://files.eric.ed.gov/fulltext/ED395017.pdf (accessed on 23 March 2016).
79. Dai, S. Handling missing responses in psychometrics: Methods and software. *Psych* **2021**, *3*, 673–693. [CrossRef]
80. Robitzsch, A. About Still Nonignorable Consequences of (Partially) Ignoring Missing Item Responses in Large-Scale Assessment. Osfpreprints. 2020. Available online: https://osf.io/hmy45 (accessed on 23 August 2021).
81. Bernaards, C.A.; Sijtsma, K. Influence of imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. *Multivar. Behav. Res.* **2000**, *35*, 321–364. [CrossRef] [PubMed]
82. Sijtsma, K.; van der Ark, L.A. Investigation and treatment of missing item scores in test and questionnaire data. *Multivar. Behav. Res.* **2003**, *38*, 505–528. [CrossRef]
83. Van Buuren, S. *Flexible Imputation of Missing Data*; CRC Press: Boca Raton, FL, USA, 2018.
84. Van Ginkel, J.R.; Van der Ark, L.A.; Sijtsma, K.; Vermunt, J.K. Two-way imputation: A bayesian method for estimating missing scores in tests and questionnaires, and an accurate approximation. *Comput. Stat. Data Anal.* **2007**, *51*, 4013–4027. [CrossRef]
85. Rubin, D.B. The calculation of posterior distributions by data augmentation: Comment on a noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. *J. Am. Stat. Assoc.* **1987**, *82*, 543–546.
86. Enders, C.K. *Applied Missing Data Analysis*; Guilford Press: New York, NY, USA, 2010.
87. Glas, C.A.; Pimentel, J.L. Modeling nonignorable missing data in speeded tests. *Educ. Psychol. Meas.* **2008**, *68*, 907–922. [CrossRef]
88. Glas, C.A.W.; Pimentel, J.L.; Lamers, S.M.A. Nonignorable data in IRT models: Polytomous responses and response propensity models with covariates. *Psychol. Test Assess. Model.* **2015**, *57*, 523–541.
89. Moustaki, I.; Knott, M. Weighting for item non-response in attitude scales by using latent variable models with covariates. *J. R. Stat. Soc. Ser. A Stat. Soc.* **2000**, *163*, 445–459. [CrossRef]
90. O'muircheartaigh, C.; Moustaki, I. Symmetric pattern models: A latent variable approach to item non-response in attitude scales. *J. R. Stat. Soc. Ser. A Stat. Soc.* **1999**, *162*, 177–194. [CrossRef]
91. Rose, N.; von Davier, M.; Nagengast, B. Modeling omitted and not-reached items in IRT models. *Psychometrika* **2017**, *82*, 795–819. [CrossRef]
92. Choi, J.; Dekkers, O.M.; le Cessie, S. A Comparison of different methods to handle missing data in the context of propensity score analysis. *Eur. J. Epidemiol.* **2019**, *34*, 23–36. [CrossRef]
93. Sperrin, M.; Martin, G.P. Multiple imputation with missing indicators as proxies for unmeasured variables: A simulation study. *BMC Med. Res. Methodol.* **2020**, *20*, 185. [CrossRef] [PubMed]
94. Groenwold, R.H.; White, I.R.; Donders, A.R.T.; Carpenter, J.R.; Altman, D.G.; Moons, K.G. Missing covariate data in clinical research: When and when not to use the missing-indicator method for analysis. *Cmaj* **2012**, *184*, 1265–1269. [CrossRef] [PubMed]
95. Sinharay, S. Reporting proficiency levels for examinees with incomplete data. *J. Educ. Behav. Stat.* **2022**, *47*, 263–296. [CrossRef]
96. Ludlow, L.H.; O'Leary, M. Scoring omitted and not-reached items: Practical data analysis implications. *Educ. Psychol. Meas.* **1999**, *59*, 615–630. [CrossRef]
97. Edwards, J.M.; Finch, W.H. Recursive partitioning methods for data imputation in the context of item response theory: A Monte Carlo simulation. *Psicológica* **2018**, *39*, 88–117. [CrossRef]
98. Sulis, I.; Porcu, M. Handling missing data in item response theory: Assessing the accuracy of a multiple imputation procedure based on latent class analysis. *J. Classif.* **2017**, *34*, 327–359. [CrossRef]
99. Xiao, J.; Bulut, O. Evaluating the performances of missing data handling methods in ability estimation from sparse data. *Educ. Psychol. Meas.* **2020**, *80*, 932–954. [CrossRef]
100. Bernaards, C.A.; Sijtsma, K. Factor analysis of multidimensional polytomous item response data suffering from ignorable item nonresponse. *Multivar. Behav. Res.* **1999**, *34*, 277–313. [CrossRef]
101. Aryadoust, V.; Goh, C.; Galaczi, E.D.; Weir, C.J. Exploring the Relative Merits of Cognitive Diagnostic Models and Confirmatory Factor Analysis for Assessing Listening Comprehension. In Proceedings of the Studies in Language Testing, Volume of Proceedings from the ALTE Kraków Conference, Kraków, Poland, 7–9 July 2011.
102. Cui, Y.; Gierl, M.J.; Chang, H.-H. Estimating classification consistency and accuracy for cognitive diagnostic assessment. *J. Educ. Meas.* **2012**, *49*, 19–38. [CrossRef]
103. Templin, J.; Henson, R.A.; Templin, S.E.; Roussos, L. Robustness of hierarchical modeling of skill association in cognitive diagnosis models. *Appl. Psychol. Meas.* **2008**, *32*, 559–574. [CrossRef]
104. Gu, Y.; Xu, G. The sufficient and necessary condition for the identifiability and estimability of the DINA model. *Psychometrika* **2019**, *84*, 468–483. [CrossRef] [PubMed]
105. Xu, G.; Zhang, S. Identifiability of diagnostic classification models. *Psychometrika* **2015**, *89*, 625–649. [CrossRef] [PubMed]
106. Dai, S.; Wang, X.; Svetina, D. TestDataImputation: Missing Item Responses Imputation for Test and Assessment Data. R Package Version 2.3. Available online: https://CRAN.R-project.org/package=TestDataImputation (accessed on 18 October 2021).