psych

Article

Handling Missing Responses in Psychometrics: Methods and Software

Shenghai Dai

Educational Psychology, Washington State University, Pullman, WA 99164, USA; s.dai@wsu.edu

Abstract: The presence of missing responses in assessment settings is inevitable and may yield biased parameter estimates in psychometric modeling if ignored or handled improperly. Many methods have been proposed to handle missing responses in assessment data that are often dichotomous or polytomous. Their applications remain nominal, however, partly due to that (1) there is no sufficient support in the literature for an optimal method; (2) many practitioners and researchers are not familiar with these methods; and (3) these methods are usually not employed by psychometric software and missing responses need to be handled separately. This article introduces and reviews the commonly used missing response handling methods in psychometrics, along with the literature that examines and compares the performance of these methods. Further, the use of the TestDataImputation package in R is introduced and illustrated with an example data set and a simulation study. Corresponding R codes are provided.

Keywords: missing responses; assessment data; psychometrics; TestDataImputation; R

Citation: Dai, S. Handling Missing Responses in Psychometrics: Methods and Software. *Psych* **2021**, *3*, 673–693. https://doi.org/10.3390/psych3040043

Academic Editor: Gongjun Xu

Received: 23 October 2021
Accepted: 16 November 2021
Published: 19 November 2021

1. Introduction

Missing item responses are a common issue in psychological and educational assessment settings in which their occurrence is inevitable because of various reasons. In this paper, we focus on the missingness in item responses that are of dichotomous or polytomous formats and are usually analyzed using psychometric models such as item response theory (IRT). In general, three common types of missing responses—missing-by-design, omitted, and not-reached responses—are categorized according to whether items are administered to the examinees or the positions where the missingness occurs on an examinee's response sheet.

Missing-by-design happens when only a fraction of the items are presented to the participants due to the booklet design that is usually employed in large-scale assessments, such as the Programme for International Student Assessment (PISA) and the National Assessment of Educational Progress (NAEP). In practice, the missing-by-design only applies to the assessments with a booklet design, and its impact is usually treated as ignorable when the plausible values approach is implemented [1,2] (In the literature, there are studies showed that the missing-by-design might still impose a non-ignorable impact on psychometric procedures. For instance, Goodman et al. [3] examined the influence of the booklet design on the Mantel-Haenszel (MH) method in differential item function (DIF) detection. Their results revealed that the missing-by-design might impact DIF detection when the sample size was small.). In this paper, the focus is on the other two categories and the term missing responses (or missing data) refers to omitted and not-reached responses, only.

For items that are administered to examinees, missing responses can be either omitted or not-reached. Omitted responses refer to the missingness identified before the last valid response observed on an examinee's response sheet. The definition of an omitted response assumes that an examinee is presented with an item but skips it over either purposefully or accidentally. The not-reached responses refer to the consecutive missingness at the end of an examinee's response sheet, as it would occur if the examinee showed no attempt to answer an item and any of the subsequent items in an assessment due to lack of time. In practice, the definition of the two types of responses may vary slightly across assessment programs, mostly lying in the coding of the missing response right after the last answer responded by the examinee. For instance, such a response is categorized as omitted in TIMSS [3], whereas in NAEP, it is usually coded as not-reached unless (1) the item is the last one in a block and of the extended constructed-response (CR) format; and (2) the examinee responded to the preceding item [4].

It is typical that an assessment yields a substantial amount of omitted and/or not-reached responses [4–7]. According to the statistics of NAEP on its 2013 fourth grade mathematics assessment, for instance, the average rates of missing responses, including both omitted and not-reached responses, were 2.65% for multiple-choice (MC) items (ranged from 0.51% to 11.45%), 6.83% for dichotomously scored CR items (ranged from 0.99% to 38.06%), and 4.21% for polytomous scored CR items (ranged from 4.21% to 15.85%) [4]. Similarly, PISA 2018 reported average proportions of missing responses that ranged from 3.57% (design B core items) to 7.42% (design B stage 1 items) for omitted, and from 0.09 (design A core items) to 12.57% (design B stage 1 items) for not-reached responses [6] (The PISA 2018 was administered using a multistage adaptive testing design [5]).

Previous research suggested that the existence of missing responses could yield biased estimations for both item parameters and ability scores, and they should be handled with caution (e.g., [8–13]). To date, the common practices to handle missing responses are to treat omitted responses as incorrect or fractionally correct and not-reached responses as incorrect or not administered. Such practices may vary across assessment programs (e.g., TIMSS vs. NAEP) and also within the same assessment for different purposes (e.g., item parameter calibration vs. ability estimation). In TIMSS 2019, for example, while the omitted responses were always treated as incorrect, the not-reached responses were treated differently. They were treated as not-administered when calibrating item parameters and as incorrect when estimating student plausible values [3]. In NAEP, however, the not-reached are always treated as not-administered while the omitted responses are usually treated with different strategies. Specifically, the omitted responses are replaced with the reciprocal value of the number of responses for MC items (e.g., $\frac{1}{4}$ if the item has four response options) and 0 for non-MC items [4].

A long-lasting debate on such practices (i.e., treating missing responses as incorrect, fractionally correct, or not-administered) in the literature has revealed that their performance might not be optimal and other approaches should be considered in handling missingness in psychometrics (e.g., [6–9,14–18]). In consequence, many methods have been proposed and discussed to handle missing responses in the literature, such as the expectation-maximization (EM) imputation [19–21], the two-way (TW) and response function (RF) imputation [22], the multiple imputation (MI) [19,20], and, more recently, the model-based approaches [23]. Applications of these approaches, however, remain nominal in assessment practices. Part of the reason may be that the practitioners tasked with psychometric analysis are not familiar with the newly proposed methods in handling missing responses, while the corresponding tools and software in implementation of such methods are not available, and the training of the specialized methods and tools are usually not part of most educational graduate programs.

In light of this, the purpose of this paper is to introduce the methods that are proposed to handle missing responses in the context of psychometrics, and further, how such methods can be implemented in practice with software packages. After a brief introduction of mechanisms and reasons for missing responses, a detailed introduction of the methods is provided as well as a synthesis of previous studies that examined the performance of the methods. With an example data set and a simulation study, the application of selected methods is then illustrated and evaluated using the TestDataImputation package [24] in R [25].

2. Mechanisms and Reasons of Missing Responses

Following Rubin [26], three categories are generally used to distinguish different mechanisms of missing responses in the context of psychometrics. The three mechanisms are missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR; also known as not missing at random or NMAR).

The MCAR is defined as the distribution of missingness on a variable that does not depend on both the values of the variable itself and other variables, observed or missing [20,27,28]. That is, there is no systematic cause for missing responses and the probability of reporting a missing response is the same for all individuals. Thus, if missing responses are MCAR, they can be treated as a random sample of the original complete observed data (i.e., no missing data are present) and ignored without introducing bias to parameter estimates [20,21,29].

MAR is a less restrictive assumption of missing data than MCAR. When missingness follows MAR, the distribution of missing responses on a variable is dependent on other observed variables but not the values of the variable itself. That is, the probability of missingness is solely a function of these observed variables [27]. There is no residual relationship between the missing and remaining components of the data after controlling for these variables. In assessment settings, for instance, MAR can be assumed if missing responses of an item on a mathematics test are only related to individuals' self-concept on mathematics, ethnicity, or social-economic status that are observed as part of the data [19,29]. In the literature, MAR is usually referred to as ignorable missingness [21]. As Little and Rubin stated, "MAR is a sufficient condition for pure likelihood and Bayesian inferences to be valid without modeling the missing mechanism" [20] (p. 14). In psychometric modeling when a likelihood-based approach is used for estimation, MAR is usually treated as ignorable missingness [16,19,30].

The definition of MAR requires that all the missingness-dependent variables should be observed [20,21,28]. In psychometric analysis, and more broadly, latent variable modeling, it is also possible that the missing responses are dependent on the latent variable (e.g., the ability or theta variable in IRT) that cannot be observed directly. This is referred to as latent ignorability when the MAR holds with the latent variable [31,32], as compared to manifest ignorability in which the MAR is dependent on observed variables.

Under MNAR, the probability of a missing response on a variable depends solely on the values of the variable itself. For instance, an item with biased or sensitive content would likely yield the mechanism of MNAR.

Rubin's work and definition on missing mechanisms provide theoretical support on research that proposed methods to handle missing responses and evaluated the performance of the methods. The assumption for an underlying mechanism, such as MAR, usually needs to be held when implementing a specific method. For example, parametric methods, such as the FIML estimation and EM imputation [19,21,30], usually assume MAR. Further, in simulation studies that examined and compared the performance of the missing data methods, data are usually generated following one or more of the mechanisms (e.g., [14]).

In practice, the causes and distribution of missing responses are usually beyond our control and the underlying mechanisms cannot be assumed nor tested empirically [17,21]. Research has been conducted to explore the complex nature of missing responses and the reasons why examinees leave items unanswered. Results revealed various associations between missing responses and the characteristics of items and/or examinees. The impact of item characteristics on the occurrence of missingness mainly lies in item format (e.g., [33–36]) and difficulty (e.g., [8]), i.e., items with a generally more complex format or of a greater difficulty were skipped more frequently by examinees. There are also studies suggesting the connection between missing responses and the ability that the test measures (e.g., [7,9,23,37]). It is noted that, however, the impact of ability on the presence of missing responses varied in terms of directions and magnitudes in the literature. As Pohl et al. [7] suggested, while examinees with a lower ability tended to omit more items, they showed opposite tendencies for not-reached responses across content domains of reading (more not-reached responses from highly able examinees) and mathematics (more not-reached responses from examinees with lower ability). Similarly, de Ayala [9] also showed that examinees with higher ability might have a greater probability to leave an item unanswered when they did not know the answer than their peers with lower ability. Further, the results of Köhler et al. [35] suggested that the relationship between ability and the occurrence of missing responses, including both omitted and not-reached responses, varied across examinees' background variables, such as immigration status and school type. In addition to ability, previous studies also revealed other person characteristics that could lead to missing responses such as reading speed [35], test-taking strategy [7], lack of confidence, and metacognitive factors [9].

3. Missing Response Handling Methods

In this section, different methods to deal with missing responses in assessment data are introduced and discussed in psychometrics. Explicitly, these methods are grouped into four categories, including (1) methods ignoring missing responses; (2) single imputation methods; (3) multiple imputation; and (4) model-based methods.

3.1. Methods Ignoring Missing Responses

3.1.1. Listwise Deletion (LW) or Complete Case Analysis

This method assumes missing responses are MCAR to produce unbiased estimates. Examinees who reported any missing values are removed from the data before analysis. Given that the assumption of underlying missing mechanisms cannot be examined empirically, it is usually not recommended to use this method to handle missingness. It might result in a loss of both data information and statistical power (e.g., [38]).

In practice, the LW is commonly used, especially for descriptive statistics and classical test theory (CTT) output, such as coefficient α, proportion correct, and CTT item discrimination index. This is partly because it is the default method in many tools and software, such as SPSS (version 28), flexMIRT (version 3.6; [39]), and IRTPRO (version 5.0), for such statistics. For the analysis where a modern psychometric model, such as an IRT model, is applied, however, the LW method is rarely used.

3.1.2. Pairwise Deletion

Known as an available-case analysis, the pairwise deletion approach uses the available data for the analysis. It is often paired up with limited information estimation methods such as the diagonally weighted least squares (DWLS) when the analysis is conducted under the ordinal factor analytic framework. For instance, the pairwise deletion is the default option in *M*plus in computing the elements of the polychoric correlation matrix when the DWLS (i.e., the WLSMV estimator) is used for the analysis [40]. The implementation of pairwise deletion assumes MCAR. If the assumption is violated, it may result in biased parameter estimates and, sometimes, a correlation or covariance matrix that is not positive definite [28,41,42].

3.1.3. Full Information Maximum Likelihood (FIML) Estimation

In psychometric modeling, the FIML is also referred to as the missing data ignoring process when a likelihood-based estimation approach is used [19,30]. When the MAR assumption holds, it yields unbiased parameters [20,27]. In practice, the FIML is usually used as a default option in the calibration of psychometric models by many software and packages such as flexMIRT (FIML; [39]) and the ltm package (available cases; [43]) in R [25].

The robustness of ML in handling missing responses has been supported by many studies, especially when there is a large sample size [21]. According to Collins et al. [44], the impact of a violation of the MAR assumption on model parameter estimations and standard errors might be only minor. Pohl et al. [7] also showed that ignoring missing responses did not introduce more bias for IRT models than the model-based missing data approaches even when the correlation between missing responses and ability was nonnegligible. (It is to be noted that Pohl et al. [7] only studied weak MNAR mechanisms in which the correlations between the missing propensity and other characteristics such as ability and item difficulty. For instance, in their study, the correlation between the missing rate and ability in reading was assumed to be −0.203 for omitted responses and 0.102 for not-reach items.)

Despite the robust performance of FIML in psychometric model estimations, some scholars expressed their concerns about its use in estimating individual ability scores, especially in high-stake settings, because it may encourage specific test-taking strategies [7,45]. As Lord [45] stated, if examinees know how missing responses are treated in estimating their ability with IRT, they could achieve a high(er) ability estimate simply by only answering items that they are confident in answering correctly. Regarding this, imputation is usually preferred over FIML when obtaining individual scores in practice. In TIMSS 2019, for instance, the omitted responses were treated as incorrect when estimating student plausible values while they were handled as ignored in the item parameter calibration process [3].

3.2. Single Imputation Methods

3.2.1. Treating Missing Responses as Incorrect (IN)

This method is also referred to as zero imputation (e.g., [16]) with which missing responses are replaced with zero before analysis. It assumes that an examinee leaves an item unanswered because they make an accurate appraisal of their ability to endorse the right option, and concludes they would not get the item correct [12,13]. In empirical settings, it is one of the standard practices to treat omitted responses (e.g., PISA, NAEP). In the research literature, a large number of studies showed that the use of IN could lead to both biased item parameters and ability estimates (e.g., [7,9,10,16]) and poor model fit (e.g., [18]) in IRT. A major criticism over this method lies in that it ignores the fact that, for an examinee with a specific ability level, their probability to endorse a correct answer is always positive [7]. There are different perspectives against such criticisms. For instance, Robitzsch [17] suggested that the conclusion of "never treat missing responses as incorrect" was doubtful in large-scale assessments because: (1) the adverse impact of IN on model parameter estimates was concluded from simulation studies in which the missing responses were generated to be dependent on ability or person covariates; and (2) the criticism again IN was based on test-theoretical arguments that "utilizes an intraindividual interpretation of item response probabilities" (p. 10).

3.2.2. Treating Missing Data as Fractionally Correct (FR)

Based on the work of Lord, this method replaces missing responses with a value of *1/m*, where *m* is the number of response options of an item [10]. It is the standard practice in NAEP to treat omitted responses on MC items. Previous studies showed that this method could perform as well as other methods (e.g., MI; [46,47]), and outperform IN and LW (e.g., [9]). Further, de Ayala et al. [9] suggested that replacing omitted responses by 0.5 (i.e., assuming two response options only) instead of *1/m* could yield a smaller bias in the estimation of ability with expected a posteriori (EAP) and maximum likelihood estimator (MLE). In practice, the application of this method is often hindered by the decimal values that it yields in the assessment data and the fact that many software and tools can only handle whole numbers (i.e., dichotomous or polytomous). One package that allows for the implementation of FR is the sirt [48] package in R. Its *rasch.mml ()* function is capable of conducting Rasch analysis with fractional item responses using the pseudo-likelihood estimation.

The FR method can also be used in a stochastic way when the *1/m* is treated as the probability endorsing the correct response of an item in such a way that it can be implemented with the multiple imputation (MI) procedure. Random draws of item responses are generated with the probability of *1/m*, resulting in multiple data sets for consecutive analysis (see the MI section below for more details).

3.2.3. Mean Imputation

Known as mean substitution, this method usually includes the person mean (PM) imputation and the item mean (IM) imputation. PM replaces missing responses for an examinee with the average of their available responses (i.e., row mean or available mean), while IM replaces missing responses with the average of available items (i.e., column mean). The purpose of the mean imputations, according to de Ayala et al. [9], is to minimize the maximum error that would occur from a wrong guess in handling missing responses. Both PM and IM are easy to implement and each has its unique advantages. According to Bernaards and Sijtsma [49], an advantage of PM is that each induvial is treated as unique and the value is imputed taking into account correlated items. The PM can also be computed based on classes defined by covariates (i.e., conditional mean imputation). Its performance might be impacted, however, when the test is multidimensional. The IM imputation is capable of correcting for multidimensional data but it cannot be conducted for covariate classes. Given their unique advantages, the two mean imputations are often used collectively in handling missing responses, resulting in various imputation methods, such as the corrected mean substitution, two-way imputation, and response function imputation (see sections below for details).

3.2.4. Corrected Mean (CM) Substitution

This method is also known as the corrected item mean (CIM) substitution [50]. It is another alternative method to PM proposed by Huisman and it imputes missing responses using the adjusted item and person means [50]. It employs the unique features of both PM and IM to correct for ability per individual and multidimensionality, respectively [49]. Let CM_{ij} be the imputed response for examinee i on item j; $\overline{y}_{i.}$ and $\overline{y}_{.j}$ be the person and item means, respectively; and $n.obs_i$ be the total number of available item responses for examinee i. Then CM_{ij} can be obtained as

$$CM_{ij} = \left[\frac{\overline{y}_{i.}}{\frac{1}{n.obs_i} \sum_j \overline{y}_{.j}} \right] \overline{y}_{.j}$$

3.2.5. Two-Way (TW) Imputation

This method imputes for the missing responses by taking into account the person mean, the item mean, and the overall mean based on available data. Let $\overline{y}_{i.}$ be the person mean or the row mean for examinee i, $\overline{y}_{.j}$ be the item mean or the column mean for item j, $\overline{y}_{..}$ be the overall mean of the available data, and TW_{ij} be the imputed response for examinee i on item j. Then, with this method, $TW_{ij} = \overline{y}_{i.} + \overline{y}_{.j} - \overline{y}_{..}$. It was proposed by Bernaards and Sijtsma [49] as an extension of PM. Their results showed that the TW performed better than PM because it added information from the item mean and the grand mean. This method was adopted by Robitzsch and Rupp in examining the impact of missing data on DIF detection [16]. In their study, the TW method was implemented within each of the examinee groups and was referred to as TW-adj.

In addition to the abovementioned deterministic use of the mean-based imputation methods (i.e., PM, IM, CM, and TW), these methods can also be employed stochastically [49,51]. The stochastic versions of the methods are accomplished by adding a random error that follows a normal distribution with the mean being zero and the variance being the residual variance and, thus, are denoted as PM-E, IM-E, CM-E, and TW-E [49]. With a comprehensive study comparing the performance of both versions of these methods, Bernaards and Sijtsma [49] suggested that: (1) a PM-based method (e.g., PM, CM, TW) was superior to those that were not PM-based (e.g., IM); (2) in the presence of unidimensionality, the stochastic version of a PM-based method outperformed its deterministic version, except for PM-E; and (3) for multidimensional data, however, the deterministic versions, especially TW, performed better. Further, a Bayesian-based proper two-way imputation with data augmentation (TW-DA) was proposed by van Ginkel et al. [51] to address the Bayesian improperness of TW-E. Their results supported the superior performance of TW-DA but also suggested that the TW-E could be used as an "accurate approximation" (p. 4013) of TW-DA. Additionally, the stochastic nature of the methods also allows for their implementations with the MI procedure (e.g., [49]).

3.2.6. Response Function (RF) Imputation

Proposed by Sijtsma and van der Ark [22], this method imputes missing responses under the nonparametric IRT framework. It assumes an item response function (IRF) for the ability variable (i.e., θ) that varies across items. What makes it different than parametric IRT models is that no item parameters are assumed and estimated with a likelihood function. Explicitly, the RF method imputes missing responses for examinee i on item j via the following steps:

- Let $\hat{R}_{(-j)i}$ be the rest score of examinee i on all available items except j, and J be the total number of items on a test. Then, $\hat{R}_{(-j)i} = \overline{y}_{i.}\ (J-1)$.
- Define $\hat{P}_j\left[\hat{R}_{(-j)i}\right]$ as the probability of endorsing a correct response for examinee i on item j based on the integer value of $\hat{R}_{(-j)i}$. Thus, if $\hat{R}_{(-j)i}$ is an integer, $\hat{P}_j\left[\hat{R}_{(-j)i}\right]$ is the fraction of examinees with $\hat{R}_{(-j)i}$ who answer item j correctly. If $\hat{R}_{(-j)i}$ is not an integer, $\hat{P}_j\left[\hat{R}_{(-j)i}\right]$ is computed by using its left and right neighbors.
- Impute the missing response with a random draw from the Bernoulli distribution defined by $\hat{P}_j\left[\hat{R}_{(-j)i}\right]$.

3.2.7. EM Imputation

In addition to its use as an algorithm in estimations, EM can also be used to impute missing responses [19,30]. It is an iterative process and each iteration consists of an E (expectation) step and an M (maximization) step. In the E-step, the expectation of the log-likelihood of the available data is obtained using the observed responses (i.e., means and covariances from available examinees). In the M-step, the maximized expectation of the log-likelihood is obtained. When the assumptions of MAR and a multivariate normal distribution underlying target variables hold, the EM algorithm imputes missing data via an iterative process until the convergence criterion is achieved (e.g., the maximum change in the parameter estimates between iterations is less than a specific value, such as 0.0001). An issue that lies in the use of EM imputation is that the imputed values are not integers and may also be out of bounds (i.e., extremely large or small). A common way to address the out-of-bounds issue in practice is to set boundaries for the values (e.g., [−10, 10]). Further, to handle missing item responses that are dichotomous and polytomous, specific procedures, such as rounding, are needed to turn the imputed values into integers.

The use of EM as an imputation method has been studied in the literature of psychometrics (e.g., [47,52,53]). Results suggested that the EM imputation might work well for polytomous data (e.g., [53]), whereas it yielded a greater bias in IRT parameter estimations than other methods (e.g., MI, [47,52]) when the data were dichotomous. Part of the reason, according to Finch [47], might lie in that the multivariate normality assumption usually does not hold for dichotomous data. In practice, the EM imputation can be accomplished with various software, such as the SAS PROC MI (e.g., [47]), SPSS (e.g., [52]), and the R package Amelia [54]. Additionally, EM imputation is a stochastic procedure and can be implemented with the MI procedure through the Bayesian bootstrap (e.g., Amelia [54]).

3.3. Multiple Imputation (MI)

Proposed by Rubin [55], the MI approach replaces missing responses with multiple sets of values. It addresses the uncertainty in the imputing process and has shown to be a robust method to handle missing data across contexts [21]. The MI procedure usually consists of three steps: imputation, analysis, and pooling [27,28].

In the first step, the missing responses are imputed m times (e.g., m = 5) through a stochastic imputation method, resulting in m complete data sets. The imputation method used in MI for the imputation task may vary and the commonly used ones include the Markov chain Monte Carlo (MCMC), predictive mean matching (PMM), and the EM algorithm [56] which are equipped in common statistical software, such as SPSS [57], SAS [58], and R [25] (e.g., mice [59]; Amelia [54]). Other stochastic imputation methods, such as RF, TW-E, and PM-E, which are mentioned above, can also be implemented with MI. Consequently, MI is a broad term in the literature of missing data as it encompasses a variety of imputation methods. It is to be noted that model convergence issues may apply in this step when the data are multidimensional [28,60,61]. To address this issue, a dimension reduction procedure, such as the principal component analysis, can be applied [28,60].

In psychometrics, while the software usually does not provide MI as an option to handle missing responses, this method is included in common statistical software, such as SPSS [57] and SAS [58]. The number of imputations in the application of MI usually ranges from 5 to 20. For example, the default number in SPSS is 5 for MI [57].

In the second step, the same analysis is conducted with each of the imputed data sets. Then, in the final step of pooling, a specific procedure (e.g., Rubin's rules) is used to combine the results by taking into account both the between and within-imputation variance (see Schafer and Graham [21]).

When multiple variables are involved in the imputation process, which is usually the case with assessment data, MI can usually be accomplished through two approaches, namely joint modeling (JM) and fully conditional specification (FCS) [28,62]. The JM assumes that the data follow a joint probability distribution (e.g., the multivariate normal distribution) and the imputed values are drawn under the distribution. The FCS is also referred to as the multivariate imputation by chained equations (MICE) or sequential regression multiple imputation [63,64]. It does not assume a joint distribution but specifies a separate model for each variable and imputes for the missing values variable by variable. That is, different imputation models are specified based on the distribution and scale of the variables (e.g., logistic regression [LR] for dichotomous variables and predictive mean matching [PMM] for continuous variables). In practice, the selection of an imputation model for the variable may vary given the software used. For instance, the SAS PROC MI procedure uses regression for continuous variables and a discriminant function for classification variables as default [65], while the R mice package uses PMM, LR, and polytomous regression for continuous, binary, and polytomous variables, respectively [59] (Both packages also allow users to specify the models used for each variable. For instance, with the SAS PROC MI, users can select from regression, LR, PMM, or discriminant function for each of the target variables in the FCS statement.). To date, the robust performance of MI, including both MIJM and MICE, in handling missing data has been widely accepted [27,28]. In practice, however, the use of MICE is more advised because of its greater flexibility, especially when the variables are of various scales and the assumption of a joint distribution is challenging [28].

The effectiveness of MI in handling missing responses for the implementation of psychometric models, such as IRT, has also been investigated and supported by various studies (e.g., [47,52,63,66,67]). For assessment data consisting of dichotomous and/or polytomous responses, the logistic and polytomous regressions are often the default choices with MI. According to van Buuren [62], the performance of PMM can be robust and stable, too, especially when the sample size is large. In addition to the commonly used methods with MI (e.g., LR, PMM, discriminant function analysis), researchers have also been exploring the potential of other methods with MI to handle missing responses in psychometric modeling, especially IRT. Examples include the MICE with classification and regression trees (MICE-CART) and the random forest imputation (MICE-RFI) [63,67], and the MI with latent class analysis (MILCA) [68–70].

In the context of psychometrics, a large number of studies in the literature have been conducted to evaluate and compare the performance of various methods using both empirical and simulated assessment data. Evidence from empirical studies suggested a notable impact of the selection of a specific method on the results of psychometric modeling at both individual and group levels (e.g., [15,17]). Ludlow and O'Leary [15] examined the impact of different approaches in dealing with omitted and not-reached responses on the results of the Rasch model. The data were from 116 seventh-graders on a 40-item numerical reasoning test. In the analysis, four methods were applied to handle missing responses: (1) treating both omitted and not-reached responses as ignored; (2) treating both as incorrect; (3) treating omitted responses as incorrect and not-reached as ignored; and (4) treating them separately (the same as the third method) in item calibration then treating both as incorrect for ability estimation. Their results revealed a "considerable impact on the estimation and interpretation of person and item statistics" (p. 629). Similarly, using PIRLS 2011 data, Robitzsch [17] found that the average achievement scores estimated with IRT varied notably when different methods were used to deal with missing responses.

The discrepancies in the results caused by different treatments of missing responses inspired research to evaluate the effectiveness of these methods in psychometrics using simulated data (e.g., [9,22,47]). de Ayala et al. [9] examined the impact of omitted responses on the performance of 3PL IRT in terms of ability estimates. Three methods were used, including IN, LW, and replacing missing responses by 0.5. They noticed that the results were most accurate when omitted responses were replaced by 0.5 and least accurate when IN was applied. Sijtsma and van der Ark [22] evaluated the performance of four methods, including PM, TW, RF, and mean RF, in dealing with missing responses. The proportions of missingness in the study were 5% and 10%, respectively. Results suggested the usefulness of both TW and RF in test and questionnaire data when missing responses were both ignorable and nonignorable. Finch [47] examined and compared seven methods, including CIM, RF, LW, IN, FR, EM, and MI on their performance of item parameter estimation in IRT. The missing responses were generated following MAR and MNAR, and at proportions of 5%, 15%, and 30%, respectively. Results revealed that MI, LW, and FR yielded lower biased item parameter estimates but no method was superior to others in terms of parameter recovery rates. Edward and Finch [63] explored the performance of MICE-LR, MICE-CART, and MICE-RFI as compared to FIML and JM in the context of both 2- and 3PL IRT. In their study, both MAR and MNAR were simulated with proportions of 5%, 10%, 15%, and 30%. Results of the study suggested superior performance of both MICE-CART and MICE-RFI to other methods across conditions. Xiao and Bulut [67] compared the effectiveness of four methods for 3PL IRT, including FIML, IN, MICE-CART, and MICE-RFI. In the study, all three missing mechanisms were included and the missing proportions were set at 10%, 20%, 30%, and 40%. Their results, however, suggested that FIML was superior to other methods while IN yielded accurate ability estimates in the presence of large missing proportions. The performance of MILCA was evaluated with the graded response model (GRM) by Sulis and Porcu [68,69] across all missing mechanisms and six missing proportions (i.e., 5%, 10%, 15%, 20%, 20%, 25%, and 30%). Other methods, including JM, MICE with polytomous regression, MI with stochastic regression, and relative mean substitution, were also included in the study for the purpose of comparison. Results of the study supported the effectiveness of the MILCA in handling missing responses for GRM, especially in the presence of MNAR and a large proportion of missingness.

Some scholars have also investigated the performance of some missing data treatment methods when the data were multidimensional. Bernaards and Sijtsma [71] investigated the use of seven methods, including LW, EM, random imputation, overall mean imputation, PM, IM, and corrected IM, to impute missing responses in simulated multidimensional latent trait data that followed MCAR and MAR. Their results showed that both EM and PM could be considered to handle missing responses in multidimensional data, as the authors stated that "... when dealing with missing item scores from multidimensional questionnaires: (a) use the EM algorithm if possible; otherwise (b) impute the Person Mean" (p. 310). The authors extended their previous study and investigated the performance of 14 missing data methods on factor analysis using questionnaire data [49]. Results suggested that the methods that used person means such as PM and TW yielded better estimates when data are multidimensional.

3.4. Model-Based Methods

The use of model-based methods in handling missing responses in psychometrics, especially IRT, has been gaining attention in recent years (e.g., [7,8,17,23,37,72–77]). In such methods, examinees' missing tendency is treated as nonignorable and included in the analysis model (e.g., Rasch, [8]; latent class IRT, [78]) when estimating the item parameters and the ability. The missing tendency is also referred to as response propensity (e.g., [75]), missing propensity (e.g., [7]), or propensity to respond/answer [78,79], and it can be modeled as a latent construct (e.g., [23]) or a manifest variable (e.g., [8]). The underlying assumption of the methods is that examinees' missing tendencies are dependent on their ability. They are not distinct, thus the missing responses are not ignorable [7].

To implement the model-based approach, the missing tendency is modeled by first transforming the assessment data into missing response indicators (i.e., 1 = a response is missing, 0 = a response is present). Then, it is modeled simultaneously with examinees' ability with a joint distribution using select models. The latent missing propensity is usually modeled using a separate unidimensional IRT model (e.g., Rasch or 2PL IRT), resulting in two measurement models, one with observed item responses and the other with the transformed missing indicators. The manifest missing tendency for an examinee can be computed as the relative number of missing responses across items. Once obtained, it can be incorporated into the analysis with a multiple-group IRT model or latent regression [7,8]. The manifest approach is usually not recommended as it may yield distorted correlations and introduce attenuated bias [7].

The (latent) model-based methods are capable of handling both omitted (e.g., [23]) and not-reach responses (e.g., [37]) and their effectiveness has been supported by many studies abovementioned. As Pohl et al. [7] indicated, such methods could not only reduce bias but also increase the accuracy for parameter estimations in IRT. The validity for the use of a model-based method, however, depends on two assumptions: (1) dimensionality underlying the missing indicators, and (2) specific types of missing response mechanisms [7,77].

The missing propensity, as measured by the missing indicators, is assumed to be unidimensional under the model-based approaches. When a unidimensional IRT such as the Rasch model is used, this assumption needs to be examined and justified. In practice, however, this is not always the case. The missing indicators may represent a multidimensional structure. For instance, in the analysis with the *National Educational Panel Study* data, Pohl et al. [7] noticed a poor model fit of three missing indicators with the unidimensional Rasch model on the reading test.

Another assumption lies in that this approach is only capable of handling specific types of nonignorable missing mechanisms. As Pohl and Becker [77] indicated, the model-based approach "performs well when the missing mechanism is MCAR, MAR, or when the missing mechanism is generated according to the model for nonignorable missing values." (p. 2). That is, the approach may not be helpful if the nonignorablility of the missingness does not rely on the missing tendency. For instance, the recent research of Pohl and Becker [77] investigated the performance of the model-based method across three nonignorable mechanisms (probability of missingness depended on the function of item responses, a latent missing propensity, or both). Results of the study revealed that the model-based method yielded unbiased estimates only when the missing responses relied on the missing propensity.

Besides the measurement model of missing indicators, it is possible for the model-based approach to include covariates in the analysis (e.g., [78,79]). For instance, Bacci and Bartolucci [79] proposed a multidimensional latent class IRT model to handle missing responses in dichotomous data. Specifically, in the proposed framework, individual covariates, in addition to the latent missing propensity and the latent ability variables, can be included through the multinomial logistic parameterization.

While less discussed in the context of psychometrics, the model-based approach can also be used to impute missing responses (i.e., model-based imputation). When the measurement model for the missing indicators is fitted, imputed values can be drawn directly from the posterior distribution for consecutive analysis. This procedure is usually implemented together with an MI approach (e.g., MICE) and referred to as a multiple imputation with missing indicators (MIMI) method [80–82]. The model-based MIMI relies on the correct specification of the imputation model that meets the assumptions (e.g., dimensionality, missing mechanisms). If the imputation model is correctly specified and the same model is used for analysis, it is equivalent to the model-based approach with which the missing indicators and item responses are modeled concurrently. Once the assumptions do not hold, according to van Buuren [28], it will "amplify aspects assumed in the model that are not supported by the data" and a data-based imputation can be used (p. 125). The data-based MIMI does not assume a missing propensity but uses the missing indicators directly in the analysis. An example is the regularized iterative multiple correspondence analysis (MCA) proposed by Josse and Husson [83] to handle missing responses in MCA with categorical data.

In this section, different treatments of missing responses in psychometrics are described. Despite the large number of methods proposed to handle missing responses and studies conducted to evaluate and compare the effectiveness of such methods, there is no evidence and support for an optimal method in practice. Despite the superior performance of MI and the model-based approach, they are still seldom used to handle missing responses in operational assessment settings [7,77]. Part of the reason may lie in the complexity in implementing such methods, potential estimation problems, lack of specific software, and the concerns in the nature and dimensional structure of the missing tendency. For instance, a complex nested MI is required when the MI is employed to handle missing responses in a large-scale assessment (e.g., PISA) that uses plausible values as ability estimations [7]. To date, set rules and guidelines remain unclear on the selection of an appropriate imputation method for missing responses. In practice, as indicated in Table 1, missing responses are treated as ignored by default with likelihood-based estimation in popular psychometric software and packages. While studies suggested that the missing responses should be handled with a specific imputation or modeling approach, the decision should be made with caution.

Table 1. Missing response handling methods in psychometric software.

Software	Version	Missing Responses Handling Methods	Source
Commercial			
ConQuest [84]	5.0	a. Missing responses can be treated in different ways (e.g., system missing, as incorrect) if different codes are used in the data (e.g., ".", M, R). b. Test reliability and standard error of measurement will not be computed if missing data are more than 10%. c. If the "regression" statement is specified, then listwise deletion is used to handle missing data. d. Joint maximum likelihood (JML) cannot be used in any cases that have missing data for all of the items on a dimension.	ACER ConQuest Manual: https://conquestmanual.acer.org
flexMIRT [39]	3.6	a. Listwise deletion is used for descriptive statistics and CTT output. b. FIML is used for model estimations.[1]	flexMIRT user's manual version 3.6 [85]
IRTPRO [86]	5.0	a. Coefficient Alpha is calculated using listwise deletion (if there are missing values in the data).	IRTPRO User Guide: https://vpgcentral.com/wp-content/uploads/2020/06/IRTPROGuide.pdf
R Packages			
CDM [87]	7.5-15	a. "NA" values are allowed in data.	https://CRAN.R-project.org/package=CDM
ltm [43]	1.1-1	a. Uses MML under MAR. b. With "na.action" argument, the analysis uses available cases by default. Users can use listwise deletion (or complete case analysis) by specifying "na.action = na.exclude".	https://CRAN.R-project.org/package=ltm
mirt [88]	1.34	a. "NA" values are allowed in data. b. Function *imputeMissing ()* allows for imputation of plausible data for missing responses. c. S_ χ^2 and Zh are not available for item fit if missing data are in presence.	https://CRAN.R-project.org/package=mirt
sirt [48]	3.10-118	a. "NA" values are allowed in data. b. FIML. c. Function *rasch.mml2 ()* allows for Rasch modeling with fractional item responses via pseudo-likelihood estimation.	https://cran.r-project.org/package=sirt
TAM [89]	3.7-16	a. "NA" values are allowed in data. b. Function *tam_remove_missings* () allows for removing rows and columns with complete missingness.	https://cran.r-project.org/package=TAM

Note. The R packages are selected from the CRAN Task View: Psychometric Models and Methods [90] and are based on their cumulative number of downloads from 2011 to 2020. CTT = classical test theory; FIML = full information maximum likelihood estimation; MML = marginal maximum likelihood; MAR = missing at random. [1] Based on "Frequently Asked Questions" on the flexMIRT website (https://vpgcentral.com/software/flexmirt).

4. Handling Missing Responses with TestDataImputation

In this section, the R package TestDataImputation [24] is introduced and its use in handling missing responses in assessment data is illustrated. The purpose of the package is to provide functions that allow for the imputation of missing responses in both dichotomous and polytomous assessment data. The current version (v2.3) is equipped with 11 functions. In addition to the main function `ImputeTestData()` and an example data set `test.data()`, it also contains eight functions for each of the imputation methods, including LW, IN, PM, IM, TW, RF, LR, PMM, and EM. Additionally, the implementation of LR and PMM depends on the mice package [59], while the EM imputation uses the Amelia package [54] with an additional rounding procedure.

4.1. Package and Data Preparation

To use the package, it needs to be installed and loaded first.

```
#Install and load the package.
install.packages('TestDataImputation')
library(TestDataImputation)
```

The package contains an example data set which consists of hypothetical dichotomous item responses from 775 examinees to 20 items. The missing proportion is 15% overall and it ranges from 13% to 17% across items.

```
#Read the example data set from the package.
attach(test.data)
#use the describe() function in the psych package for descriptive statistics.
psych::describe(test.data)
# vars        n mean sd median trimmed mad min max range skew kurtosis  se
# Item_1    1 657 0.51 0.50      1    0.51  0   0  1    1 -0.03    -2.00 0.02
# Item_2    2 642 0.59 0.49      1    0.61  0   0  1    1 -0.37    -1.87 0.02
# Item_3    3 644 0.49 0.50      0    0.49  0   0  1    1 0.02    -2.00 0.02
# ...
# Item_18  18 647 0.23 0.42      0    0.16  0   0  1    1 1.29    -0.34 0.02
# Item_19  19 662 0.19 0.40      0    0.12  0   0  1    1 1.55    0.40 0.02
# Item_20  20 661 0.18 0.39      0    0.11  0   0  1    1 1.62    0.63 0.02
```

4.2. Handling Missing Responses

4.2.1. Listwise Deletion (LW)

The LW can be conducted with the main function and the `Listwise()` function. It simply removes the examinees who showed any missing responses. The argument `Mvalue="NA"` in both functions specifies how the missing values are coded in the data. By default, all missing values are coded as "NA". Other values can be used. For example, if the number 8 is used for missing data, then `Mvalue=8` should be used. If the main function is used, then we also need to specify the method to be used. For LW, the argument is specified as `method="LW"`. The two other arguments in the main function, `max.score=1` and `round.decimal=0`, do not apply in LW and can be ignored. Once the procedure is accomplished, the complete data are stored in the object `data.imputed` and can be saved with the function `write.csv()`. After LW, only responses of 85 examinees are retained in the data.

```
#Use the main ImputeTestData () function.
data.imputed<-ImputeTestData(test.data, Mvalue = "NA", max.score = 1,
              method = "LW", round.decimal = 0)
#Use the Listwise() function.
data.imputed<-Listwise(test.data, Mvalue = "NA")
#Save the imputed data set as a .csv file.
write.csv(data.imputed, file = "data.imputed.csv")
```

4.2.2. Treating Missing Responses as Incorrect (IN)

The IN can be conducted with the main function or the `TreatIncorrect()` function. If the main function is used, the argument is specified as `method="IN"`. Similar to LW, the two other arguments in the main function, `max.score=1` and `round.decimal=0`, can be ignored for IN. Because the function assigns the value of 0 as incorrect, it is suggested that the incorrect responses are coded as 0 in the original data set too. Once the procedure is accomplished, the complete data are stored in the object `data.imputed` and can be saved with the function `write.csv()`.

```
#Use the main ImputeTestData () function
data.imputed<-ImputeTestData(test.data, Mvalue = "NA", max.score = 1,
                              method = "IN", round.decimal = 0)
#Use the TreatIncorrect () function
data.imputed<-TreatIncorrect(test.data, Mvalue="NA")
#Save the imputed data set as a .csv file.
write.csv(data.imputed, file = "data.imputed.csv")
```

4.2.3. Person mean (PM) Imputation

The PM method replaces missing responses for examinees with their averaged response across available items (i.e., row means). It may yield decimal values. In such a case, `round.decimal=0` can be used to round the imputed data into integers. Other values can also be specified, such as `round.decimal=2`, to retain the corresponding decimal places. The default `max.score=1` argument implies that the data are dichotomous (0 = incorrect, 1 = correct). When the data consists of polytomous responses, the maximum response value needs to be specified with this argument. Once the procedure is accomplished, the complete data are stored in the object `data.imputed` and can be saved with the function `write.csv()`.

```
#Use the main ImputeTestData () function
data.imputed<-ImputeTestData(test.data, Mvalue = "NA", max.score = 1,
                              method = "PM", round.decimal = 0)
#Use the PersonMean () function
data.imputed<-PersonMean(test.data, Mvalue = "NA",
                          max.score = 1, round.decimal = 0)
#Save the imputed data set as a .csv file.
write.csv(data.imputed, file = "data.imputed.csv")
```

4.2.4. Item Mean (IM) Imputation

The IM method replaces missing responses with average item responses (i.e., column means). By default, all the values are rounded to be integers unless specified otherwise. The specification of the arguments is similar to that for PM. Once the procedure is accomplished, the complete data are stored in the object `data.imputed` and can be saved with the function `write.csv()`.

```
#Use the main ImputeTestData () function
data.imputed<-ImputeTestData(test.data, Mvalue = "NA", max.score = 1,
                              method = "IM", round.decimal = 0)
#Use the ItemMean () function
data.imputed<-ItemMean(test.data, Mvalue = "NA",
                              max.score = 1, round.decimal = 0)
#Save the imputed data set as a .csv file.
write.csv(data.imputed, file = "data.imputed.csv")
```

4.2.5. Two-Way (TW) Imputation

TW replaces missing responses by taking into account PM, IM, and the grand mean of the available responses. In the calculation of the TW values, we noticed that it could be out of bounds of 0 and the specified maximum score. Thus in the function, the out-of-bounds values are coded to the corresponding bound (i.e., 0 or the specified maximum score). By default, all the values are rounded to be integers unless specified otherwise. Once the procedure is accomplished, the complete data are stored in the object `data.imputed` and can be saved with the function `write.csv()`.

```
#Use the main ImputeTestData () function
data.imputed<-ImputeTestData(test.data, Mvalue = "NA", max.score = 1,
                              method = "TW", round.decimal = 0)
#Use the Twoway () function
data.imputed<-Twoway(test.data, Mvalue = "NA", max.score = 1,
                      round.decimal = 0)
#Save the imputed data set as a .csv file.
write.csv(data.imputed, file = "data.imputed.csv")
```

4.2.6. Response Function (RF) Imputation

This method imputes missing responses using the non-parametric IRT response functions. The probability of endorsing a correct response (i.e., $\hat{P}_j\left[\hat{R}_{(-j)i}\right]$) is first calculated. The missing response is then replaced with a random value drawn from the Bernoulli distribution with the parameter $\hat{P}_j\left[\hat{R}_{(-j)i}\right]$. Once the procedure is accomplished, the complete data are stored in the object `data.imputed` and can be saved with the function `write.csv()`.

```
#Use the main ImputeTestData () function.
data.imputed<-ImputeTestData(test.data, Mvalue = "NA", max.score = 1,
                              method = "RF", round.decimal = 0)
#Use the ResponseFun () function.
data.imputed<-ResponseFun(test.data, Mvalue = "NA", max.score = 1,
                           round.decimal = 0)
#Save the imputed data set as a .csv file.
write.csv(data.imputed, file = "data.imputed.csv")
```

4.2.7. Logistic Regression (LR) Imputation

The LR method imputes missing responses using binary (dichotomous responses) or polytomous (polytomous responses) logistic regression. The function uses the package mice [59] to conduct the logistic regressions. The choice of binary or polytomous logistic regressions is made based on the specification of the `max.score` argument. Once the procedure is accomplished, the complete data are stored in the object `data.imputed` and can be saved with the function `write.csv()`.

```
#Use the main ImputeTestData () function.
data.imputed<-ImputeTestData(test.data, Mvalue = "NA", max.score = 1,
                              method = "LR", round.decimal = 0)
#Use the LogisticReg () function.
data.imputed<-LogisticReg(test.data, Mvalue = "NA", max.score = 1)
#Save the imputed data set as a .csv file.
write.csv(data.imputed, file = "data.imputed.csv")
```

4.2.8. Predictive Mean Matching (PMM)

The PMM method imputes missing responses using the predictive mean matching method. The function uses the package mice [59] to conduct the PMM. Once the procedure is accomplished, the complete data are stored in the object `data.imputed` and can be saved with the function `write.csv()`.

```
#Use the main ImputeTestData () function.
data.imputed<-ImputeTestData(test.data, Mvalue = "NA", max.score = 1,
                              method = "PMM", round.decimal = 0)
```

```
#Use the micePMM () function.
data.imputed<-micePMM(test.data, Mvalue = "NA")
#Save the imputed data set as a .csv file.
write.csv(data.imputed, file = "data.imputed.csv")
```

4.2.9. EM Imputation

This method imputes missing responses using the EM algorithm. The function uses the package Amelia [54] to implement the EM algorithm. When the EM algorithm is accomplished, the `EMimpute()` or the main function will screen the imputed data and check for out-of-bounds values. Furthermore, if there are any, the out-of-bounds values are coded to the specified response boundaries (i.e., 0 or the specified maximum score). The functions also round the imputed values to integers unless specified otherwise. Once the procedure is accomplished, the complete data are stored in the object `data.imputed` and can be saved with the function `write.csv()`.

```
#Use the main ImputeTestData () function.
data.imputed<-ImputeTestData(test.data, Mvalue = "NA", max.score = 1,
                              method = "EM", round.decimal = 0)
#Use the EMimpute () function.
data.imputed<-EMimpute(test.data, Mvalue = "NA", max.score = 1,
                        round.decimal = 0)
#Save the imputed data set as a .csv file.
write.csv(data.imputed, file = "data.imputed.csv")
```

4.3. A Simulation Study

4.3.1. Simulation Design and Data Generation

A Monte Carlo simulation study was conducted with the main purpose to evaluate the validity for the use of the TestDataImputation package in handling missing responses. Firstly, both dichotomous and polytomous (five categories) item responses of 1000 individuals (N = 1000) and 20 items (J = 20) were generated using the *simdat ()* function from the R package irtplay [91], resulting in two baseline conditions. Explicitly, the dichotomous responses were generated using 2PL IRT while the polytomous responses were simulated using the graded response model (GRM). The item difficulty parameters for the 2PL model were random draws from a standard normal distribution, N (0, 1), and the four item thresholds for the GRM were from the uniform distributions of U (−2, −1), U (−1, 0), U (0, 1), and U (1, 2). For both models, the item discrimination parameters were generated from a uniform distribution of U (0.75, 1.33) and the theta values from N (0, 1). The generated item parameters for both models are provided in the Supplementary File.

Following de Ayala et al. [9] and Finch [29], missing responses were then generated, with the rates of 15% and 30%, respectively. Specifically, the probability of missing responses for an individual was assumed to be inversely related to his or her observed total score. To accomplish the process, the total scores were first divided into seven fractiles (0–2, 3–5, 6–8, . . . , 18–20). The individuals in a higher score fractile were then assigned a lower probability of missingness and the total missing rates were controlled at the designed levels (i.e., 15% and 30%). Further, all items were prone to missingness in the process.

When the incomplete data sets were ready, each of the methods in the TestDataImputation package (i.e., LW, IN, PM, IM, TW, LR, EM, RF, and PMM) were applied to impute the missing responses. Given that the mirt package [88] in R was used to analyze the data, the two missing data handling options in the package were also included in the study. One was to ignore the missingness under the likelihood-based estimation (i.e., FIML) while the other was the use of the *imputeMissing ()* function to impute the missing responses. Explicitly, the model estimates and ability scores yielded from FIML were used when implementing the *imputeMissing ()* function.

The simulation design yielded a total number of 46 conditions, consisting of 2 baseline and 44 missing data conditions (2 response formats × 2 missing rates × 11 missing response handling methods). Each condition was replicated 500 times.

4.3.2. Analysis and Outcome

The function *mirt ()* with default settings from the R package mirt [88] was used to analyze the simulated data sets. The dichotomous item responses were analyzed using the 2PL IRT while the polytomous data were analyzed with the GRM. To achieve the purpose of the simulation study, the item parameter recovery rates were examined for each condition using both the average mean absolute difference (MAD) and root mean squared error (RMSE) across the 500 replications. The MAD was computed using the formula $\frac{1}{J}\sum_{j=1}^{J}|\hat{\pi}-\pi|$, where $\hat{\pi}$ and π were the estimated and true parameters, and J was the number of items. The RMSE was computed using the formula $\sqrt{\frac{1}{J}\sum_{j=1}^{J}(\hat{\pi}-\pi)^2}$ (While RMSE is a commonly used outcome measure in simulation studies, its use in evaluating the performance of missing data methods may be biased. According to van Buuren [28], the calculation of RMSE ignores the uncertainty nature of the missing responses, and consequently, will favor deterministic methods over stochastic ones. For more details, see van Buuren [28], Section 2.6).

4.3.3. Results

Tables 2 and 3 present the parameter recovery results of the 2PL-IRT and GRM, respectively. Across the specified conditions, both the average MAD and RMSEs are reasonable across the missing data methods. While the main purpose of the simulation is to evaluate the validity for the use of the package, the results revealed patterns regarding the performance of the methods. Explicitly, results of the 2PL-IRT showed that (see Table 2): (1) the LW approach resulted in large biases in item parameter estimations and model non-convergence issues, especially in the presence of a large missing proportion (i.e., 30%); (2) the performance of the LR, EM, RF, PMM, and FIML methods, and the *imputeMissing* function, was robust in recovering both item parameters across conditions, as indicated by the close and relatively small MAD and RMSE values; (3) TW was the optimal method in estimating item difficulties but yielded larger biases for item discriminations than other methods (except for LW and PM), especially when the missing rate was large; (4) the performance of the IN, IM, and PM methods was not superior to the other methods (except for LW) across most conditions.

Similar patterns of the results were found for GRM (see Table 3). The FIML, *imputeMissing* function, PMM, EM, and LR showed robust performance across conditions while the LW yielded the largest amount of bias. The exceptions are associated with the performance of PM and RF: (1) the PM imputation performed as equally optimal as TW in estimating item thresholds but not item discriminations; and (2) the performance of RF was similar to the robust methods (e.g., EM) with slightly larger MAD and RMSE values when the missing rate was 15%. Its performance, however, declined abruptly in the presence of a 30% missing rate, especially for item discriminations, and resulted in the largest MAD and RMSEs across the methods.

Table 2. Item parameter recovery of the 2PL-IRT model across missing response handling methods.

Missing Rate	Missing Treatment	Item Discrimination		Item Difficulty	
		MAD	RMSE	MAD	RMSE
0	\	0.09	0.12	0.10	0.14
15%	LW	0.60	1.14	3.10	8.46
	IN	0.13	0.17	0.28	0.33
	PM	0.25	0.28	0.32	0.38
	IM	0.15	0.19	0.37	0.48
	TW	0.22	0.26	0.11	0.14
	LR	0.11	0.15	0.13	0.17
	EM	0.11	0.14	0.13	0.18
	RF	0.12	0.15	0.14	0.18
	PMM	0.11	0.15	0.13	0.17
	FIML	0.10	0.13	0.12	0.16
	mirt	0.11	0.15	0.13	0.17
30%	LW	\	\	\	\
	IN	0.16	0.21	0.58	0.71
	PM	0.55	0.58	0.54	0.63
	IM	0.25	0.31	0.90	1.20
	TW	0.48	0.53	0.13	0.16
	LR	0.14	0.18	0.16	0.22
	EM	0.15	0.19	0.18	0.25
	RF	0.16	0.21	0.25	0.31
	PMM	0.14	0.18	0.17	0.23
	FIML	0.12	0.15	0.14	0.19
	mirt	0.13	0.17	0.16	0.22

Note. 2PL-IRT = two-parameter item response theory; MAD = mean absolute difference; RMSE = root mean squared error; LW = listwise deletion; IN = treat missing responses as incorrect; PM = person mean imputation; IM = item mean imputation; TW = two-way imputation; LR = logistic regression imputation; EM = expectation–maximization imputation; RF = response function imputation; PMM = predictive mean matching; FIML = full information maximum likelihood; mirt = the *imputeMissing ()* function in the mirt package. Results for LW under the 30% missing rate were not obtained across most of the replications and thus are not presented in the table.

Table 3. Item parameter recovery of the grade response model across missing treatment methods.

Missing Rate	Missing Treatment	Item Discrimination		Item Thresholds							
				b1		*b2*		*b3*		*b4*	
		MAD	RMSE	MAD	RMSE	MAD	RMSE	MAD	RMSE	MAD	RMSE
0	\	0.06	0.08	0.10	0.13	0.07	0.09	0.07	0.09	0.10	0.13
15%	LW	0.30	0.38	1.41	3.67	0.72	1.64	0.63	1.56	1.37	4.23
	IN	0.14	0.17	0.50	0.54	0.33	0.35	0.29	0.32	0.33	0.38
	PM	0.16	0.18	0.14	0.17	0.10	0.12	0.07	0.09	0.09	0.12
	IM	0.19	0.21	0.64	0.69	0.55	0.58	0.33	0.36	0.44	0.49
	TW	0.14	0.16	0.16	0.19	0.09	0.12	0.07	0.09	0.09	0.12
	LR	0.08	0.10	0.16	0.21	0.10	0.13	0.08	0.10	0.12	0.15
	EM	0.09	0.11	0.24	0.30	0.13	0.16	0.07	0.10	0.14	0.19
	RF	0.13	0.15	0.26	0.32	0.16	0.20	0.16	0.20	0.25	0.31
	PMM	0.08	0.11	0.17	0.22	0.11	0.14	0.08	0.10	0.12	0.16
	FIML	0.07	0.09	0.14	0.19	0.09	0.12	0.08	0.10	0.11	0.14
	mirt	0.08	0.10	0.15	0.20	0.10	0.13	0.08	0.10	0.12	0.15
30%	LW	\	\	\	\	\	\	\	\	\	\
	IN	0.22	0.26	1.08	1.13	0.78	0.81	0.68	0.72	0.72	0.77
	PM	0.26	0.29	0.28	0.33	0.20	0.23	0.10	0.13	0.11	0.14
	IM	0.28	0.31	1.28	1.36	1.11	1.17	0.79	0.83	0.93	1.00
	TW	0.25	0.27	0.34	0.38	0.23	0.26	0.11	0.14	0.14	0.17
	LR	0.13	0.16	0.34	0.42	0.20	0.24	0.11	0.14	0.16	0.21
	EM	0.14	0.17	0.50	0.57	0.26	0.30	0.09	0.12	0.24	0.30
	RF	0.33	0.36	1.03	1.12	0.84	0.89	0.83	0.89	1.00	1.09
	PMM	0.14	0.17	0.36	0.44	0.22	0.26	0.12	0.15	0.16	0.22
	FIML	0.11	0.14	0.31	0.37	0.19	0.23	0.11	0.14	0.13	0.18
	mirt	0.12	0.15	0.33	0.41	0.20	0.25	0.11	0.14	0.15	0.20

Note. MAD = mean absolute difference; RMSE = root mean squared error; LW = listwise deletion; IN = treat missing responses as incorrect; PM = person mean imputation; IM = item mean imputation; TW = two-way imputation; LR = logistic regression imputation; EM = expectation–maximization imputation; RF = response function imputation; PMM = predictive mean matching; FIML = full information maximum likelihood; mirt = the *imputeMissing ()* function in the mirt package. Results for LW under the 30% missing rate were not obtained across most of the replications and thus are not presented in the table.

5. Conclusions

The occurrence of missing responses in assessment settings is inevitable and nonignorable. Previous literature has shown that their presence could yield both biased item parameter and ability estimates in psychometric modeling if they are ignored or handled with an improper method. In practice, treating as incorrect and/or ignoring with likelihood-based estimation are the two major treatments of missing responses. A large number of methods are proposed in the context to handle missing data, but their applications remain nominal. Part of the reasons are: (1) there is not sufficient support in the literature for an optimal method; (2) many practitioners and researchers are not familiar with the newly proposed methods; and (3) these methods are usually not employed by psychometric software.

This article discusses the issues of missing responses in assessment data, introduces commonly used methods that are appropriate to handling missing responses in psychometrics, and reviews the literature that examines and compares the performance of these missing data methods. Further, the use of the TestDataImputation package in R is described and illustrated with example data and a simulation study. Corresponding R codes are also provided. It contributes to the context by systematically introducing the missing data handling methods and also serves as a tutorial to guide the implementation of these methods in practice.

As stated previously, the rules and guidelines for an optimal missing response handling method still do not exist in empirical settings. Different methods embody different assumptions about the mechanisms and distributions of the missing responses, the reasons why examinees omit or not reach items, and what missing data would have been if they had not been missed. The decision on the selection of a specific method should be made with caution and based on reasonable explanations.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/psych3040043/s1, Example R code, Table S1: Item parameters used for the simulation study.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of the simulation study are available from the author upon reasonable request.

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

References

1. Braun, H.; von Davier, M. The Use of Test Scores from Large-Scale Assessment Surveys: Psychometric and Statistical Considerations. *Large-Scale Assess. Educ.* **2017**, *5*, 17. [CrossRef]
2. Rutkowski, L.; Gonzalez, E.; Joncas, M.; von Davier, M. International Large-Scale Assessment Data: Issues in Secondary Analysis and Reporting. *Educ. Res.* **2010**, *39*, 142–151. [CrossRef]
3. Martin, M.O.; von Davier, M.; Mullis, I.V.S. (Eds.) *Methods and Procedures: TIMSS 2019 Technical Report*; TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA): Chestnut Hill, MA, USA, 2020 2020; ISBN 978-1-889938-53-0.
4. NCES Treatment of Missing Responses in NAEP. Available online: https://nces.ed.gov/nationsreportcard/tdw/analysis/2000_2001/scaling_missing.aspx (accessed on 13 August 2020).
5. NCES NAEP Analysis and Scaling. Available online: https://nces.ed.gov/nationsreportcard/tdw/analysis/ (accessed on 10 June 2021).
6. OECD. *PISA 2018 Technical Report*; Organization for Economic Co-Operation and Development [OECD]: Paris, France, 2021.
7. 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]
8. Rose, N.; Davier, M.; Xu, X. Modeling Nonignorable Missing Data with Item Response Theory (IRT). *ETS Res. Rep. Ser.* **2010**, *2010*, i-53. [CrossRef]
9. 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]
10. 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]
11. Lord, F.M. Unbiased Estimators of Ability Parameters, of Their Variance, and of Their Parallel-Forms Reliability. *Psychometrika* **1983**, *48*, 233–245. [CrossRef]
12. Mislevy, R.J.; Wu, P.K. Inferring Examinee Ability When Some Item Responses Are Missing. *ETS Res. Rep. Ser.* **1988**, *1988*, i-75. [CrossRef]
13. 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]
14. Finch, H. The Use of Multiple Imputation for Missing Data in Uniform DIF Analysis: Power and Type I Error Rates. *Appl. Meas. Educ.* **2011**, *24*, 281–301. [CrossRef]

15. Ludlow, L.H.; O'Leary, M. Scoring Omitted and Not-Reached Items: Practical Data Analysis Implications. *Educ. Psychol. Meas.* **1999**, *59*, 615–630. [CrossRef]
16. Robitzsch, A.; Rupp, A.A. Impact of Missing Data on the Detection of Differential Item Functioning: The Case of Mantel-Haenszel and Logistic Regression Analysis. *Educ. Psychol. Meas.* **2009**, *69*, 18–34. [CrossRef]
17. Robitzsch, A. About Still Nonignorable Consequences of (Partially) Ignoring Missing Item Responses in Large-Scale Assessment. 2020. Available online: https://osf.io/hmy45 (accessed on 23 August 2021).
18. Zhang, B.; Walker, C.M. Impact of Missing Data on Person—Model Fit and Person Trait Estimation. *Appl. Psychol. Meas.* **2008**, *32*, 466–479. [CrossRef]
19. Little, R.J.; Rubin, D.B. The Analysis of Social Science Data with Missing Values. *Sociol. Methods Res.* **1989**, *18*, 292–326. [CrossRef]
20. 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.
21. Schafer, J.L.; Graham, J.W. Missing Data: Our View of the State of the Art. *Psychol. Methods* **2002**, *7*, 147. [CrossRef] [PubMed]
22. 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] [PubMed]
23. Holman, R.; Glas, C.A. Modelling Non-Ignorable Missing-Data Mechanisms with Item Response Theory Models. *Br. J. Math. Stat. Psychol.* **2005**, *58*, 1–17. [CrossRef] [PubMed]
24. 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).
25. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2021.
26. Rubin, D.B. Inference and Missing Data. *Biometrika* **1976**, *63*, 581–592. [CrossRef]
27. Enders, C.K. *Applied Missing Data Analysis*; Guilford Press: New York, NY, USA, 2010.
28. van Buuren, S. *Flexible Imputation of Missing Data*; CRC Press: Boca Raton, FL, USA, 2018.
29. Peugh, J.L.; Enders, C.K. Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement. *Rev. Educ. Res.* **2004**, *74*, 525–556. [CrossRef]
30. Little, R.J.; Rubin, D.B. *Statistical Analysis with Missing Data*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2002.
31. Frangakis, C.E.; Rubin, D.B. Addressing Complications of Intention-to-Treat Analysis in the Combined Presence of All-or-None Treatment-Noncompliance and Subsequent Missing Outcomes. *Biometrika* **1999**, *86*, 365–379. [CrossRef]
32. Harel, O.; Schafer, J.L. Partial and Latent Ignorability in Missing-Data Problems. *Biometrika* **2009**, *96*, 37–50. [CrossRef]
33. 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.
34. Koretz, D. Omitted and Not-Reached Items in Mathematics in the 1990 National Assessment of Educational Progress. 1993. Available online: http://cresst.org/wp-content/uploads/TECH357.pdf (accessed on 29 August 2021).
35. Köhler, C.; Pohl, S.; Carstensen, C.H. Investigating Mechanisms for Missing Responses in Competence Tests. *Psychol. Test Assess. Modeling* **2015**, *57*, 499–522.
36. Zhang, J. Relationships between Missing Responses and Skill Mastery Profiles of Cognitive Diagnostic Assessment. Ph.D. Dissertation, University of Toronto, Toronto, ON, Canada, 2013, unpublished.
37. Glas, C.A.; Pimentel, J.L. Modeling Nonignorable Missing Data in Speeded Tests. *Educ. Psychol. Meas.* **2008**, *68*, 907–922. [CrossRef]
38. Wilkinson, L. Statistical Methods in Psychology Journals: Guidelines and Explanations. *Am. Psychol.* **1999**, *54*, 594. [CrossRef]
39. Cai, L. *FlexMIRT Version 3.6: Flexible Multilevel Multidimensional Item Analysis and Test Scoring*; Vector Psychometric Group: Chapel Hill, NC, USA, 2021.
40. Muthén, L.K.; Muthén, B.O. *Mplus User's Guide*, 8th ed.; Muthén & Muthén: Los Angeles, CA, USA, 1998.
41. Shi, D.; Lee, T.; Fairchild, A.J.; Maydeu-Olivares, A. Fitting Ordinal Factor Analysis Models With Missing Data: A Comparison Between Pairwise Deletion and Multiple Imputation. *Educ. Psychol. Meas.* **2020**, *80*, 41–66. [CrossRef]
42. Asparouhov, T.; Muthén, B.O. Weighted Least Squares Estimation with Missing Data. 2010. Available online: Statmodel.com (accessed on 17 October 2021).
43. Rizopoulos, D. ltm: Latent Trait Models under IRT (R Package Version 1.1-1). Available online: https://CRAN.R-project.org/package=ltm (accessed on 26 April 2021).
44. 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. [CrossRef]
45. Lord, F.M. *Applications of Item Response Theory to Practical Testing Problems*; Routledge: New York, NY, USA, 1980.
46. Brown, N.J.S.; Svetina, D.; Dai, S. Impact of Methods of Scoring Omitted Responses on Achievement Gaps. In Proceedings of the Annual Meeting of the National Council on Measurement in Education, Philadelphia, PA, USA, 4–6 April 2014.
47. Finch, H. Estimation of Item Response Theory Parameters in the Presence of Missing Data. *J. Educ. Meas.* **2008**, *45*, 225–245. [CrossRef]
48. Robitzsch, A. sirt: Supplementary Item Response Theory Models (R Package Version 3.10-118.). Available online: https://CRAN.R-project.org/package=sirt (accessed on 18 October 2021).

49. 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]
50. Huisman, M. *Item Nonresponse: Occurrence, Causes, and Imputation of Missing Answers to Test Items*; DSWO Press Leiden: Leiden, The Netherlands, 1999.
51. van Ginkel, J.R.; Andries van der Ark, L.; 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]
52. Kalkan, Ö.K.; Kara, Y.; Kelecioglu, H. Evaluating Performance of Missing Data Imputation Methods in IRT Analyses. *Int. J. Assess. Tools Educ.* **2018**, *5*, 403–416. [CrossRef]
53. Enders, C.K. The Impact of Missing Data on Sample Reliability Estimates: Implications for Reliability Reporting Practices. *Educ. Psychol. Meas.* **2004**, *64*, 419–436. [CrossRef]
54. Honaker, J.; King, G.; Blackwell, M. Amelia: A Program for Missing Data. *J. Stat. Softw.* **2021**, *45*, 1–47.
55. Rubin, D.B. The Calculation of Posterior Distributions by Data Augmentation: Comment: 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.
56. Lin, T.H. A Comparison of Multiple Imputation with EM Algorithm and MCMC Method for Quality of Life Missing Data. *Qual. Quant.* **2010**, *44*, 277–287. [CrossRef]
57. IBM SPSS Statistics Impute Missing Data Values (Multiple Imputation). Available online: https://www.ibm.com/docs/en/spss-statistics/SaaS?topic=imputation-impute-missing-data-values-multiple (accessed on 30 August 2021).
58. Yuan, Y.C. Multiple Imputation for Missing Data: Concepts and New Development (Version 9.0). *SAS Inst. Inc. Rockv. MD* **2010**, *49*, 12.
59. van Buuren, S.; Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations (R package version 3.13.0). Available online: https://cran.r-project.org/package=mice (accessed on 19 October 2021).
60. Huque, M.H.; Carlin, J.B.; Simpson, J.A.; Lee, K.J. A Comparison of Multiple Imputation Methods for Missing Data in Longitudinal Studies. *BMC Med Res. Methodol.* **2018**, *18*, 168. [CrossRef] [PubMed]
61. Hu, B.; Li, L.; Greene, T. Joint Multiple Imputation for Longitudinal Outcomes and Clinical Events Which Truncate Longitudinal Follow-Up. *Stat Med.* **2016**, *35*, 2991–3006. [CrossRef] [PubMed]
62. van Buuren, S. Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification. *Stat. Methods Med Res.* **2007**, *16*, 219–242. [CrossRef] [PubMed]
63. 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]
64. Azur, M.J.; Stuart, E.A.; Frangakis, C.; Leaf, P.J. Multiple Imputation by Chained Equations: What Is It and How Does It Work? *Int. J. Methods Psychiatr. Res.* **2011**, *20*, 40–49. [CrossRef]
65. SAS Institute Inc. PROC MI: FCS Statement: SAS/STAT(R) 9.3 User's Guide. Available online: https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_mi_sect008.htm (accessed on 21 October 2021).
66. Wolkowitz, A.A.; Skorupski, W.P. A Method for Imputing Response Options for Missing Data on Multiple-Choice Assessments. *Educ. Psychol. Meas.* **2013**, *73*, 1036–1053. [CrossRef]
67. 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] [PubMed]
68. Sulis, I. A further proposal to perform multiple imputation on a bunch of polytomous items based on latent class analysis. In *Statistical Models for Data Analysis*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 361–369.
69. 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]
70. Vermunt, J.K.; van Ginkel, J.R.; van der Ark, L.A.; Sijtsma, K. Multiple Imputation of Incomplete Categorical Data Using Latent Class Analysis. *Sociol. Methodol.* **2008**, *38*, 369–397. [CrossRef]
71. 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]
72. Albanese, M.T.; Knott, M. TWOMISS: A Computer Program for Fitting a One-or-Two-Factor Logit-Probit Latent Variable Model to Binary Data When Observations May Be Missing: [Handbook]. In *Cadernos de Matemática e Estatística. Série F, Trabalho de Divulgação*; Universidade Federal do Rio Grande do Sul: Porto Alegre, Brazil, 1992; pp. 1–47. Available online: https://www.lume.ufrgs.br/bitstream/handle/10183/204868/000054976.pdf?sequence=1 (accessed on 29 August 2021).
73. 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. Modeling* **2015**, *57*, 523–541.
74. 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* **2000**, *163*, 445–459. [CrossRef]
75. 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* **1999**, *162*, 177–194. [CrossRef]
76. Rose, N.; von Davier, M.; Nagengast, B. Modeling Omitted and Not-Reached Items in IRT Models. *Psychometrika* **2017**, *82*, 795–819. [CrossRef]

77. Pohl, S.; Becker, B. Performance of Missing Data Approaches under Nonignorable Missing Data Conditions. *Methodology* **2020**, *16*, 147–165. [CrossRef]
78. Bacci, S.; Bartolucci, F. A Multidimensional Latent Class IRT Model for Non-Ignorable Missing Responses. *arXiv* **2021**, arXiv:1410.4856. [CrossRef]
79. Bacci, S.; Bartolucci, F. A Multidimensional Finite Mixture Structural Equation Model for Nonignorable Missing Responses to Test Items. *Struct. Equ. Modeling A Multidiscip. J.* **2015**, *22*, 352–365. [CrossRef]
80. 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] [PubMed]
81. Sperrin, M.; Martin, G.P. Multiple Imputation with Missing Indicators as Proxies for Unmeasured Variables: Simulation Study. *BMC Med Res. Methodol.* **2020**, *20*, 185. [CrossRef] [PubMed]
82. 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]
83. Josse, J.; Husson, F. Handling Missing Values in Exploratory Multivariate Data Analysis Methods. *J. De La Société Française De Stat.* **2012**, *153*, 79–99.
84. Wu, M.; Adams, R.; Wilson, M.; Haldane, S. *ACER ConQuest 2.0: General Item Response Modelling Software [Computer Program Manual]*; ACER Press: Camberwell, UK, 2007.
85. Houts, C.R.; Cai, L. *FlexMIRT User's Manual Version 3.6: Flexible Multilevel Multidimensional Item Analysis and Test Scoring, Verson 3.5*; Vector Psychometric Group: Chapel Hill, NC, USA, 2020.
86. Cai, L. *IRTPRO Version 5*; Vector Psychometric Group: Chapel Hill, NC, USA, 2020.
87. Robitzsch, A.; Kiefer, T.; George, A.C.; Uenlue, A. CDM: Cognitive Diagnosis Modeling (R Package Version 7.5–15). Available online: https://cran.r-project.org/package=CDM (accessed on 21 October 2021).
88. Chalmers, R.P. mirt: A Multidimensional Item Response Theory Package for the R Environment. *J. Stat. Softw.* **2012**, *48*, 1–29. [CrossRef]
89. Robitzsch, A.; Kiefer, T.; Wu, M. TAM: Test Analysis Modules (R Package Version 3.7-16). Available online: https://CRAN.R-project.org/package=TAM (accessed on 21 October 2021).
90. Mair, P. CRAN Task View: Psychometric Models and Methods. 2021. Available online: https://ftp.uni-bayreuth.de/math/statlib/R/CRAN/src/contrib/Views/Psychometrics.html (accessed on 18 October 2021).
91. Lim, H.; Wells, C.S. irtplay: Unidimensional Item Response Theory Modeling (R Package Version 1.6.2). Available online: https://cran.r-project.org/package=irtplay (accessed on 19 October 2021).