

Reporting Proficiency Levels for Examinees With Incomplete Data

Sandip Sinharay 

Educational Testing Service, Princeton, NJ, USA

Takers of educational tests often receive proficiency levels instead of or in addition to scaled scores. For example, proficiency levels are reported for the Advanced Placement (AP[®]) and U.S. Medical Licensing examinations. Technical difficulties and other unforeseen events occasionally lead to missing item scores and hence to incomplete data on these tests. The reporting of proficiency levels to the examinees with incomplete data requires estimation of the performance of the examinees on the missing part and essentially involves imputation of missing data. In this article, six approaches from the literature on missing data analysis are brought to bear on the problem of reporting of proficiency levels to the examinees with incomplete data. Data from several large-scale educational tests are used to compare the performances of the six approaches to the approach that is operationally used for reporting proficiency levels for these tests. A multiple imputation approach based on chained equations is shown to lead to the most accurate reporting of proficiency levels for data that were missing at random or completely at random, while the model-based approach of Holman and Glas performed the best for data that are missing not at random. Several recommendations are made on the reporting of proficiency levels to the examinees with incomplete data.

Keywords: IRT models; multiple imputation; regression imputation

1. Introduction

Takers of several large-scale educational tests receive a *pass–fail status* or a *classification* or a *proficiency level* instead of or in addition to a scaled score. For example, those taking any title of the Advanced Placement (AP[®]) examination receive a proficiency classification that is an integer between 1 and 5 indicating how qualified they are to receive college credit and placement; a classification of 5 is equivalent to a college course grade A or A+ (e.g., Patterson & Ewing, 2013). Takers of the Praxis[®] tests and the U.S. Medical Licensing Examination[®] (USMLE) examinations receive a pass or a fail classification (e.g., Educational Testing Service, 2020; USMLE, 2020). For convenience, tests (such as AP,

Praxis, and USMLE) that involve the reporting of classifications will henceforth be referred to as *classification tests*.

Occasionally, portions of educational tests are lost, missing, or unscorable due to unforeseen events that are unrelated to examinee behavior. For example, (a) Gayle (2017) reported that portions of the AP test were lost for dozens of students from a county, (b) poor audio quality or excessive background noise during recording may render some item responses unscorable on speaking items of a test such as the Praxis[®] *Spanish World Language test* (Educational Testing Service, n.d.) that includes Speaking items, and (c) technology-related disruptions may lead to missing item scores, as was observed for a state test described by Byrne (2017). The loss, missingness, or unscorability of a portion of a classification test results in missing item scores and prevents the computation of raw/scaled/pattern scores for the corresponding examinees. Because the classification for an examinee is computed from the examinee's score, missing scores create difficulty in reporting classifications for the corresponding examinees. The test administrators have the option of not reporting classifications to such examinees. Alternately, they have the option of reporting classifications after employing an approach for imputation/projection/estimation of the corresponding classifications. Students whose AP test responses are lost have the option of accepting a projected score based on the rest of their scored examination, having the score canceled, or taking the missing portion of the examination, as implied by Codes 37 and 38 on Page 6 of College Board (n.d.) and as stated by Gayle (2017). Imputed classifications are also reported for the USMLE when some item scores are missing due to reasons such as technical problems (e.g., Jodoin & Rubright, 2020).

In this article, the procedure used to report classifications to examinees whose data are incomplete due to unforeseen events that are unrelated to examinee behavior is referred to as "imputation" and the outcome of this imputation procedure is referred to as an "imputed" classification. Imputation of a classification involves making inferences on the missing portion of the test based on the nonmissing portion and is a special case of imputation of missing data (e.g., Little & Rubin, 2002, p. 20) where an investigator makes inferences from data that include some missing observations. While researchers have examined various problems related to missing data in educational measurement (e.g., De Ayala et al., 2001; Finch, 2008; Holman & Glas, 2005; Sijtsma & van der Ark, 2003; Smits et al., 2002; Xiao & Bulut, 2020), there is a lack of research on imputing classifications in the presence of missing item scores, Feinberg (2021) and Sinharay (2021a) being exceptions. Accordingly, the goal of this article is to explore imputation approaches that allow test administrators to impute and report accurate, and hence fair and valid, classifications to the examinees with missing item scores.

Section 2 includes a review of the literature on the imputation of missing data for educational and psychological tests. Section 3 includes brief descriptions of

how several imputation approaches that have been used to impute missing data in other contexts in educational measurement can be used to impute classification. Section 4 includes a comparison of seven imputation approaches using real data from four classification tests. Section 5 includes discussions and conclusions.

This article does not include any investigation on the estimation of any model parameters (like item response theory or IRT item parameters) or summary statistics (like mean scores or reliability) in the presence of missing item scores. Researchers such as Edwards and Finch (2018), Finch (2008), and Sijtsma and van der Ark (2003) investigated these problems, and the model parameters for the classification tests considered in this article are accurately estimated because of the availability of a large sample of examinees with no missing data. Item responses are often missing due to various types of examinee behavior. The most common of these types of missing responses are omitted and/or not-reached responses. Examples of operational practice to deal with omitted and/or not-reached responses can be found in Allen et al. (2001). Researchers such as De Ayala et al. (2001) and Glas and Pimentel (2008) explored various approaches for handling omitted and not-reached responses—this article does not include an examination of these approaches and does not involve imputation of scores on items with omitted and not-reached responses. Instead, examinee records with such responses, which constitute a very small part of the available data sets, are not included in any computations in this article.

2. Literature Review

Graham (2009), Graham (2012), Schafer (1997), Schafer and Graham (2002), and Sinharay et al. (2001) provided reviews of the literature on missing data analysis in general.

Table 1 shows a list of several research studies that dealt with the problem of missing item scores in the context of educational or psychological measurement. The studies (listed in the second column of the table) are grouped according to their main focus (first column). Some of the studies included in Table 1 have multiple foci (e.g., Huisman & Molenaar, 2001)—so they appear in multiple rows.

Table 1 indicates that a wide variety of problems related to missing item scores have been the focus of existing research. However, there is a lack of research, with the exceptions of Feinberg (2021) and Sinharay (2021a), on the reporting of examinee classifications, which is the problem of interest in the current article.

Feinberg (2021) considered the problem of reporting classifications to examinees who are interrupted during a test and cannot finish the test, so that scores are missing on the items toward the end of the test. He focused on classification tests that only involve “pass” and “fail” classes (i.e., examinees are classified in two classes) and include only dichotomous items and suggested four approaches for estimating the probability of passing on an incomplete classification test. Feinberg (2021) found that none of the approaches is an overwhelming winner

TABLE 1.

The Focus Areas of Some of the Existing Studies on Missing Data in Educational and Psychological Measurement

Focus	Reference
Estimation of item response theory (IRT) item parameters	Edwards & Finch (2018), Finch (2008), Holman & Glas (2005), Glas & Pimentel (2008), and Sulis & Porcu (2017)
Estimation of item proportion correct	Moustaki & Knott (2000)
IRT equating	Bolsinova & Maris (2016) and Shin (2009)
Estimation of reliability	Huisman & Molenaar (2001) and Sijtsma & van der Ark (2003)
Estimation of the IRT ability parameter	De Ayala et al. (2001), Mislevy & Wu (1996), Sulis & Porcu (2017), and Xiao & Bulut (2020)
Estimation of factor loadings	Bernaards & Sijtsma (2000)
Imputation of the total score in psychological tests	van Ginkel et al. (2010) and Huisman & Molenaar (2001)
Imputation of composite scores	Rose et al. (2019)
Imputation of scaled scores	Sinharay (2021b)
Imputation of the missing answer choices on multiple-choice items	Wolkowitz & Skorupski (2013)
Estimation of ability in computerized adaptive multistage testing	Cetin-Berber et al. (2019)
IRT model-fit statistics	Sijtsma & van der Ark (2003)
Mokken's scalability coefficient	Sijtsma & van der Ark (2003)
Modeling omitted and not-reached responses using advanced IRT models	Holman & Glas (2005), Glas & Pimentel (2008), Rose et al. (2017), and Sulis & Porcu (2017)
Estimation of the regression of ability on explanatory variables	Köhler et al. (2017)
Estimation of the parameters of multilevel IRT models	Kadengye et al. (2012)
Parameter estimation in the presence of missing data in background questionnaires	Grund et al. (2020)
Estimation of grade point average	Smits et al. (2002)
Reporting of examinee classification	Feinberg (2021) and Sinharay (2021a)

while an IRT-based approach performs the best overall. Sinharay (2021a) considered the more general problem of estimating the probability of passing on an incomplete classification test that could include polytomous items and could have missing item scores anywhere (and not necessarily at the end) on the test. He suggested two approaches—one each based on IRT and classical test theory—that were found to perform better than the approaches of Feinberg (2021).

The four approaches of Feinberg (2021) and the two approaches of Sinharay (2021a) are briefly described in Online Appendix A.

The current article focuses on a slightly different and more general problem where (a) the interest lies in imputing the classifications themselves and not on estimating the probabilities of classifications, (b) the missing scores could occur on any item and not necessarily on those toward the end of the test, (c) the test may include polytomous items and the item scores are linearly weighted to produce a *weighted sum score* or *total composite score* (TCS) that is then converted to a classification using one or more cut scores, and (d) the examinees may be classified into more than two classes. Therefore, the approaches of Feinberg (2021) and Sinharay (2021a) do not apply to the problem considered this article. Also, a difficulty of applying the approaches of Feinberg (2021) and Sinharay (2021a) to the problem of interest in this article is that Feinberg (2021) and Sinharay (2021a) placed some examinees into an indeterminate class, whereas, for the operational tests considered in this article, the classification resulting from the imputation procedure is not allowed to be “indeterminate.”¹ The imputation approach based on IRT models, which is considered later in this article, is close in spirit to the IRT-based approaches that Feinberg (2021) and Sinharay (2021a) found the best overall.

3. Imputation Approaches

Seven imputation approaches were considered in this article. One of the approaches (linking) is used operationally to impute missing classifications for the tests considered later in this article. The other six approaches have been used in the analysis of missing data in educational measurement but never have been applied to impute missing classifications. The first two imputation approaches are not based on any rigorous prediction models while the last five approaches are. It is important for an imputation approach to be based on a rigorous prediction model in case the imputed classifications are questioned by the users of these high-stakes tests. The imputation approaches are explained below using the hypothetical example of a classification test that includes nine items and a hypothetical Examinee 1 for whom scores on Items 1 and 2 are missing and scores on Items 3 through 9 are available. Let us consider that the TCS on the test is computed as the weighted sum of the item scores before being converted to a classification using a set of cut scores. Let us further assume that the possible scores on Item 1 are 0, 1, 2, 3, and 4 and those on Item 2 are 0, 1, 2, and 3 and that the generalized partial credit model (GPCM; Muraki, 1992) provides an adequate fit to the data from the test.

3.1. Person-Mean Imputation (PMI)

The PMI or *proration* approach to impute missing scores (e.g., Huisman, 1999) involves the imputation of each missing item score of an examinee by the

mean score on the nonmissing item scores for that examinee. If the maximum possible score varies over the items, all item scores are converted to a proportional score (by dividing the item scores by the maximum possible score on the respective items) before applying the PMI approach to impute a proportional score; then the imputed proportional score is multiplied by the maximum possible score on the item to obtain the imputed item score. To apply this approach to Examinee 1, one has to first compute average of the proportional scores over Items 3 through 9; imputed scores on Items 1 and 2 for the examinee can be obtained by multiplying this average by 4 and 3, respectively. Then, the *imputed TCS* for Examinee 1 can be computed as the weighted sum of the actual/observed scores on Items 3 through 9 (or, the observed *partial composite score* or PCS on Items 3–9) plus the *imputed PCS* on Items 1 and 2, which is the weighted sum of the imputed scores on Items 1 and 2. Then, the abovementioned cut scores can be used to convert the imputed TCS to an imputed classification for Examinee 1. The PMI approach is not based on any statistical theory or prediction model but is simple and does not require any specialized software. Because this approach implicitly assumes that all items are of equal difficulty, the approach may lead to inaccurate imputation when the items with missing scores differ in difficulty from the other items.

3.2. Linking

In the application of the linking approach to impute a score for Examinee 1, score linking (e.g., Kolen & Brennan, 2014, p. 487) is performed of the observed PCS based on the last seven items to the observed TCS using the data from the subsample of examinees whose scores are available on all nine items on the test. Then, the imputed TCS of the examinee can be obtained as the value of the TCS that is equivalent to (or linked to) the examinee's PCS. Finally, the cut scores for the test can be used to convert the imputed TCS to an imputed classification for Examinee 1. The single-group equipercentile equating (e.g., Kolen & Brennan, 2014, p. 14) approach was used to perform score linking in this article. In this approach, one finds the linked/equivalent TCS corresponding to a PCS of S as the value of the TCS that has the same percentile rank as S . Thus, if G and F , respectively, denote the cumulative distribution function corresponding to the distribution of the TCS and the PCS, then the linked TCS corresponding to a PCS of S is obtained as $G^{-1}(F(S))$ after defining the inverse function $G^{-1}(\cdot)$ properly. The linking approach is operationally used to impute classifications for the tests considered later in this article. Given that the linking approach is not based on any statistical theory or prediction model (e.g., Braun & Holland, 1982, p. 14), one goal of this article is to examine whether other imputation approaches, especially those based on statistical modeling of the data, lead to more accurate imputation compared to the linking approach. Given that the problem of estimating missing scores or classification is more a prediction problem and linking is

not the optimum approach for prediction (e.g., Fayers & Hays, 2014), the linking approach is expected to perform worse than the more advanced imputation approaches.

3.3. Imputation Based on Linear Regression

In the application of the imputation approach based on regression, a linear regression that predicts $Y = w_1X_1 + w_2X_2$ from X_3, X_4, \dots, X_9 is fitted using the subsample of examinees who had scores available on all nine items on the test, where w_i is the weight placed on the score on item i in the computation of the TCSs of the examinees and X_i is the score on item i , $i = 1, 2, \dots, 9$. Then, the scores of Examinee 1 on Items 3–9 are entered in the fitted regression equation to compute the imputed PCS on Items 1 and 2 for the examinee; this imputed PCS is added to the (observed) PCS on Items 3 through 9 to obtain an imputed TCS for Examinee 1. Finally, the cut scores can be used to convert the imputed TCS to an imputed classification for Examinee 1. The advantage of this approach is the conceptual simplicity and the ubiquitous nature of linear regression. The approach may perform well when the assumptions underlying the model (e.g., that the regression is linear, the variances of the scores are homogeneous, the interitem correlation matrix is not collinear) are met, but the approach may not perform well if the assumptions are violated.

3.4. Imputation Based on Cumulative Logistic Regression

Because the proficiency level or classification of an examinee is an ordinal variable, the *cumulative logistic regression model* or the *proportional odds model* (e.g., Agresti, 2013, p. 301) is a natural approach for imputing the classification. Let the variable C denote the classification of an examinee. In the application of this approach to impute a classification for Examinee 1 on the abovementioned test, a cumulative logistic regression model for predicting C from X_3, X_4, \dots, X_9 is employed. According to the model, the probability of C smaller than or equal to c is given by

$$P(C \leq c) = \frac{\exp(\alpha_c + \beta_3X_3 + \beta_4X_4 + \dots + \beta_9X_9)}{1 + \exp(\alpha_c + \beta_3X_3 + \beta_4X_4 + \dots + \beta_9X_9)}, \quad (1)$$

where α_c and the β_i 's are unknown coefficients of the model and are estimated from the subsample of examinees who had scores available on all nine items on the test (so that the actual values of C of these examinees are known as well). Then, the scores of Examinee 1 on Items 3 through 9 are entered in the fitted Equation 1 to compute the estimated probabilities of the different classifications for the examinee. The imputed classification for the examinee can then be obtained as the classification with the largest estimated probability. This approach is slightly different from the other approaches considered in this article in that the missing classification is directly imputed without imputing the item

scores or the TCS. The approach may perform well when the assumptions underlying the model (e.g., that the relationships between the logit of the probabilities of various classifications and the item scores are linear, the interitem correlation matrix is not collinear) are met.

3.5. Imputation Based on IRT

The missing score on an item for an examinee can be imputed by its posterior expectation given the available item scores under an IRT model after the model parameters have been estimated using an examinee sample (e.g., Korobko et al., 2008). In this article, the three-parameter logistic model (3PLM) was used for the dichotomous items and the GPCM was used for the polytomous items—this combination of IRT models is used in several large-scale assessments including the National Assessment of Educational Progress (Allen et al., 2001, pp. 229–230). The steps for the IRT-based imputation approach for Examinee 1 are the following:

1. Estimate the parameters of the IRT model from the data. Ignore the items with missing scores in this step.
2. Impute the PCS on Items 1 and 2 for the examinee as

$$\int_{\theta} E(Y|\theta) p(\theta|X_3, X_4, \dots, X_9) d\theta, \quad (2)$$

where $Y = w_1X_1 + w_2X_2$ and $p(\theta|X_3, X_4, \dots, X_9)$ is the posterior distribution of the ability of Examinee 1 (e.g., Baker & Kim, 2004, p. 159) given the scores on Items 3 through 9 and $E(Y|\theta)$ is the expected value of the PCS on Items 1 and 2 conditional on θ . The quantity $E(Y|\theta)$ can be computed as

$$\begin{aligned} E(Y|\theta) &= w_1E(X_1|\theta) + w_2E(X_2|\theta), \\ &= w_1 \sum_{k=1}^4 kP(X_1 = k|\theta) + w_2 \sum_{k=1}^3 kP(X_2 = k|\theta), \end{aligned} \quad (3)$$

where $P(X_1 = k|\theta)$ and $P(X_2 = k|\theta)$ are computed under the GPCM as

$$P(X_1 = k|\theta) = \frac{\exp[k\alpha\theta - \sum_{h=1}^k \beta_h]}{1 + \sum_{h=1}^m \exp[h\alpha\theta - \sum_{l=1}^h \beta_l]}, \quad (4)$$

where α and β_h 's, respectively, denote the slope and difficulty parameters for the item. For convenience, the notation in this article does not reflect the fact that the quantities such as $p(\theta|X_3, X_4, \dots, X_9)$, $E(Y|\theta)$, and $P(X_1 = k|\theta)$ are estimates (as they depend on the item parameter estimates).²

3. Compute the imputed TCS of Examinee 1 as the imputed PCS on Items 1 and 2 plus the actual PCS on Items 3 through 9.
4. Use the abovementioned cut scores to convert the imputed TCS to an imputed classification for Examinee 1.

This approach is similar in spirit to the (IRT-based) Lord–Wingersky approach of Feinberg (2021) and the modified Lord–Wingersky approach of Sinharay (2021a). The R package *mirt* (Chalmers, 2012) was used to fit the IRT models in this article. The integral in Equation 2 was approximated using numerical integration.

In application of this approach to the operational data sets in this article, the few omitted and not-reached responses that are observed for the data sets are not included in any computations. The application of the approach involves the assumption that the IRT model fits the data. So, the approach may not perform well when there is misfit of the IRT model to the data.

3.6. Multiple Imputation (MI) Using Data Augmentation and Chained Equations

Researchers in several fields including education and psychology (e.g., Finch, 2008; Smits et al., 2002; Sulis & Porcu, 2017) have found the MI approach (e.g., Little & Rubin, 2002, p. 85) to lead to the most accurate estimation of various quantities of interest (such as item parameter and reliability) in the presence of missing data. To apply MI, one assumes a probability model for the data and computes a predictive (or conditional) distribution of the missing data given the observed data and draws multiple values from this predictive distribution.

A recent MI approach that is gaining popularity is MI using chained equations (MICE), also known as fully conditional specification (FCS; Raghunathan et al., 2001). The MICE approach specifies the imputation model on a variable-by-variable basis by a set of conditional densities, one for each incomplete variable. Starting from an initial imputation, the MICE approach draws imputations by iterating over the univariate conditional densities. Variables are imputed one at a time, as opposed to all being simultaneously imputed as in some other MI approaches. A major advantage of the MICE approach over other MI approaches is that the conditional distributions of the variables can be specified to be models such as the ordered logit model that are appropriate for item scores that are the variables of interest in this article. The MICE approach has been found to lead to accurate imputation in comparison studies by researchers such as Horton and Lipsitz (2001). However, application of the MICE approach to educational measurement is rare, with the exception of Edwards and Finch (2018) and Xiao and Bulut (2020). While the R package *mice* (e.g., van Buuren & Groothuis-Oudshoorn, 2011) can be used to implement the FCS approach, the stand-alone BLIMP software (Enders et al., 2018) was used for the data examples in this article.³ The BLIMP software uses an ordered probit model to impute incomplete ordinal variables and utilizes a fully Bayesian estimation and imputation (e.g., Enders et al., 2020). In this

article, five sets of draws/imputations of missing item scores were used for the MICE approach; each set of draws was used to compute an imputed TCS; the final imputed TCS was obtained as the simple average of the five imputed TCSs. This strategy is in agreement with the recommendation on combining results from MIs by Rubin (1987, p. 76), who suggested estimating a quantity of interest by the simple average of the quantity computed from the MIs. Finally, the abovementioned cut scores were used to convert the final imputed TCS to an imputed classification for the examinee. This approach is based on the assumption that the ordered probit model assumed by the BLIMP software fits the data.

3.7. A Method Based on Modeling of Nonignorable or Missing Not at Random (MNAR) Data

Sinharay (2021b) provided an example where item responses missing apparently due to technical problems could be MNAR. So, there was the need to include an approach that can model MNAR data. Holman and Glas (2005) suggested such an approach in which one assumes that an IRT model fits the item scores X_i s. One also defines, for item i , a missing-score indicator variable d_i that is 1 or 0, respectively, depending on whether the score on item i is missing or not. One also assumes a probability model for the d_i s. While Holman and Glas (2005) suggested several models for the d_i s, the model assumed in this article is given by⁴

$$P(d_i = 1) = \frac{\exp(\xi - \gamma_i)}{1 + \exp(\xi - \gamma_i)}, \quad (5)$$

where γ_i is a location parameter and ξ is a latent missingness propensity parameter for the examinee. Thus, the larger ξ is, the more is the tendency for the response to be missing. As in Holman and Glas (2005), the joint population distribution $g(\theta, \xi|m, \Sigma)$ for (θ, ξ) , where θ is the examinee ability parameter, was assumed to be a bivariate normal distribution. The marginal likelihood for an examinee is then given by

$$\int_0 \int_{\xi} \prod_i [f(x_i|\theta, \alpha_i, b_i)h(d_i|\xi, \gamma_i)]g(\theta, \xi|m, \Sigma)d\xi d\theta, \quad (6)$$

where $f(x_i|\theta, \alpha_i, b_i)$ and $h(d_i|\xi, \gamma_i)$, respectively, are the probability distributions of x_i and d_i and are computed using Equations 4 and 5. To ensure model identifiability, Holman and Glas (2005) set μ equal to the zero vector. The matrix Σ denotes the unknown variance matrix of the population distribution of (θ, ξ) . One then estimates the parameters α_i s, β_i s, γ_i s, and Σ by maximizing the product of the aforementioned likelihood over all examinees in the sample. While Holman and Glas (2005) did not estimate examinee abilities or proficiency classifications, it is possible to adopt their approach to impute proficiency classifications. After

computing the item-parameter estimates, one can compute the posterior distribution of θ given the d_i s and x_i s as

$$p(\theta|x_1, x_2, \dots, d_1, d_2, \dots) \propto \int_{\xi} \prod_i [f(x_i|\theta, \alpha_i, b_i)h(d_i|\xi, \gamma_i)]g(\theta, \xi|\mu, \Sigma)d\xi,$$

where the α_i s, β_i s, γ_i s, and Σ are replaced by their estimates, and μ is set to the zero vector. Then, one can impute the PCS on Items 1 and 2 for Examinee 1 as

$$\int_{\theta} E(Y|\theta)p(\theta|x_1, x_2, \dots, d_1, d_2, \dots)d\theta,$$

for $E(Y|\theta)$ given in Equation 3, compute the imputed TCS of Examinee 1 as the imputed PCS on Items 1 and 2 plus the observed PCS on Items 3–9 and use the cut scores to convert the imputed TCS to an imputed classification for the examinee. The application of the approach involves the assumption that the bivariate IRT model provided in Equation 6 fits the data.

4. Methods: Comparison of Imputation Approaches for Four Classification Tests

The seven imputation approaches were compared using one data set each from four high-stakes operational tests—the data sets included only examinees with scores available on all items or examinees with complete records. These four data sets are referred to as the “complete data sets.”⁵ In the comparison study, different parts of the complete data sets were assumed missing in various ways and the missing parts were imputed by seven imputation approaches.

4.1. The Four Tests and the Data Sets

Three of the four data sets were from three titles of a large-scale classification test. These three titles are henceforth referred to as Tests A1, A2, and A3 and are intended to measure the mastery of the examinees on an arts subject, a science subject, and a language subject, respectively. The fourth data set originated from another large-scale classification test—this test is henceforth referred to as Test B. For each test, a weighted sum of the item scores is computed to yield a TCS for each examinee. Several cut scores are used to convert the TCSs to classifications that are reported to the examinees.

Occasionally, portions of these classification tests are lost, missing, or unscorable due to various reasons, which leads to the problem of missing item scores, or, incomplete test, for the corresponding examinees. When scores on a small portion of the test are missing for an examinee on any of these tests, an imputed classification, which is based on the available item scores, is made available to the examinee. No imputation is performed for the examinees for whom the missing portion of the test contributes roughly more than 50% to the TCS—these examinees are allowed a free retest.

TABLE 2.
Some Information About the Four Tests

Quantity	Test A1	Test A2	Test A3	Test B
Number of MC items	70	40	60	60
Number of CR items	6	6	8	6
Mean interitem correlations	.46, .09, .19	.60, .13, .28	.41, .09, .15	.42, .09, .15
Contribution of MC items to TCS (%)	50	50	50	55
Rough sample size	21,000	101,000	27,000	1,000
Average percentage score on MC items	56	64	54	75
Average percentage score on CR items	57	32	63	62
Maximum possible TCS	140	108	120	96
Mean of TCS	78	46	69	68
Mean as a percentage of the maximum	56	43	57	71
<i>SD</i> of TCS	25	21	18	16
Reliability of TCS	.91	.94	.90	.89
Number of proficiency classes	5	5	5	3
Percentage of misfitting items	6.6	7.9	6.7	4.9

Note. The three mean interitem correlations, respectively, represent the mean interitem correlation among the constructed response (CR) items, among the multiple-choice (MC) items, and among combinations of an MC and a CR item. *SD* = standard deviation; TCS = total composite score.

Table 2 includes some information about the four tests including the number of items, mean interitem correlation, rough sample size, average percent scores,⁶ reliability, and the number of proficiency levels or classes for the available data sets. The table also shows the percentages of misfitting items that are the percentages of the $S - \chi^2$ statistic (Kang & Chen, 2008; Orlando & Thissen, 2000) that are significant at 5% significance level from the fitting of a combination of the 3PLM (to the dichotomous items) and GPCM (to the polytomous items) to the data sets. The table shows that the sample sizes for Tests A1 through A3 are between 21,000 and 101,000 while that for Test B is about 1,000. Each test includes somewhere between 45 and 70 MC items and between six and eight CR items. The means of the TCSs (and the means as percentages of the maximum provided in the next row) indicate that Test B is the easiest and Test A2 is the most difficult. The reliability of the TCS for the whole test, computed using the Feldt-Raju procedure (Qualls, 1995), was found to be .89 or above for all the test forms considered here. The reliability for Tests A2 and B are the largest and the smallest, respectively, among the four tests. The extent of item misfit seems smallest (and close to nominal level) for Test B and somewhat elevated for Tests

A1 through A3, but this result could be an outcome of much larger sample sizes for the latter tests (the extent of IRT model fit has been noted to be sensitive to sample size by experts such as Hambleton & Han, 2005). The interitem correlations are larger on average for Test A2. It was also found that the classifications are uniformly distributed among the classes for Test A2, the lowest and highest classifications are more frequent for Test B, and the middlemost classification is the most frequent for Test A3. More details (like the names, exact sample sizes, cut scores, and frequencies in the different classes) on Tests A1 through A3 and B cannot be provided for confidentiality purposes.

4.2. Study Design and Computation

Descriptions of the factors that were varied in the comparison study are provided below, followed by a description of the steps of the comparison study.

4.2.1. Missing score patterns *considered*. Historically, for each of Tests A1 through A3 and B, numerous patterns of missing item scores have been observed due to various factors that are not related to examinee behavior. To make the comparison study realistic while keeping the size of the study manageable, the following five patterns of missing item are included in the comparisons: (a) 100% MC item scores missing, (b) 100% CR item scores missing, (c) 50% MC item scores missing, (d) 50% CR item scores missing, and (e) 25% CR item scores missing. These are the most common patterns of missing item scores across these four tests.

For Tests A1 through A3 and B, the number of CR items are not necessarily multiples of 4, but the percentage of CR item scores missing over all the replications of the comparison study is equal to 25% for the fifth missing score pattern.⁷

4.2.2. Missing data mechanisms considered

For a data set that includes some missing values, the probability that a value is missing is related to the underlying values of the variables in the data set according to one of the following three *missing data mechanisms* (e.g., Little & Rubin, 2002, p. 11)—(a) missing completely at random (MCAR), (b) missing at random (MAR), and (c) MNAR. The approaches used to analyze missing data should ideally depend on the nature of the dependencies implied by these missing data mechanisms. For data from classification tests such as Tests A1 through A3 or B, missing data for an examinee are MCAR when the missingness or the probability of the values missing is unrelated to any data including the item scores (either observed or unobserved) and examinee covariates. Given that this article focuses on scores that are missing due to unforeseen events (such as technology-related disruptions) that are unrelated to examinee behavior, most such scores are expected to be MCAR—so the MCAR mechanism was considered. The missing item scores for these tests are MAR if the probability of a missing score for an

examinee on an item depends on the examinee's scores on the items on which scores are available; this type of missingness may occur if one who has performed poorly on the other items of the test makes random noises on a speaking item to produce an poor/unscorable audio recording.⁸ The MNAR missing data mechanism may arise for classification tests if the probability of a missing score for an examinee on an item depends on the (hypothetical) score the examinee would have received on the item if it were not missing; this type of missingness may occur if one who is likely to perform poorly on a speaking item makes random noises to produce an unscorable audio recording. Sinharay (2021b) provided examples where item responses missing apparently due to technical problems could actually be MAR and also MNAR. Therefore, MAR and MNAR mechanisms were also considered in the comparison study.

In the comparison study, to simulate under the MCAR mechanism, the examinees and the items with missing scores were chosen at random. To simulate missing scores under the MAR mechanism for, for example, the "50% missing CR" case, the PCS was calculated for each examinee by computing a weighted sum of all the MC items. The examinees were then divided into four groups based on their PCSs using the three quartiles of the sample distribution of the PCS. The members of each group were assigned the probabilities .14, .03, .02, and .01 of missing responses—thus, smaller scores had larger probabilities of missing responses. The mean of these probabilities across the groups is equal to .05—thus about 5% examinees had 50% missing CR item scores. For each examinee, a uniform random number u between 0 and 1 was generated and compared with the probability of missing responses (p) assigned to the examinee. The scores on 50% randomly chosen CR items were marked missing for the examinees for whom u was smaller than p . The set of 50% CR items with missing scores varied over the examinees.

The model of Holman and Glas (2005), described in Equation 5, was used to simulate MNAR scores. To simulate MNAR scores, ξ s for the examinees were simulated from a univariate normal distribution with mean $\hat{\rho}\hat{\theta}$ and variance $1 - \rho^2$, where the $\hat{\theta}$ s are the estimated examinee abilities in the full sample for the complete data set.⁹ For all patterns of missing scores, the item scores were assumed to be missing for the 5% examinees with the largest simulated values of ξ . For the "100% CR item scores missing" and "100% MC item scores missing" cases, the item scores on all CR or MC items, respectively, were assumed missing for these 5% examinees. For the other cases, values of γ_i were chosen, by a trial and error method, to ensure that the desired proportion of examinees had missing scores¹⁰ and then values of d_i were simulated from a Bernoulli distribution with success probability given by Equation 5. As recommended in Pohl et al. (2014), the correlation ρ between ξ and θ determines the amount of nonignorability or the extent to which the data deviate from the MAR assumption—so this correlation was set to two values in two different simulations. In one set of

MNAR simulations, ρ was set equal to $-.18$ as in Pohl et al. (2014) to represent a moderate extent of nonignorability. In another set of MNAR simulations, ρ was set equal to $-.8$ as in Rose et al. (2010) to represent a stronger extent of nonignorability.¹¹ Note that it is possible to simulate missing scores under the MAR mechanism from the model of Holman and Glas (2005), using an observed scores in place of ξ , but this procedure was not used.¹²

4.2.3. Steps in the comparison and computation. The comparison of the imputation approaches was based on 1,000 replications of the following steps for each missing data pattern for each missingness mechanism for each of the four tests:

1. Draw the examinee-item combinations whose scores are to be treated as missing in the comparison:
 - If the missingness mechanism is MCAR, then draw 5% examinees randomly from the full sample. If the missing score pattern is “100% MC item scores missing” or “100% CR item scores missing,” then mark the scores on all MC or CR items of the sampled examinees as missing. For any of the other three missing data patterns, for each sampled examinee, randomly draw the set of appropriate items (such as 50% MC items) and mark their scores as missing.
 - If the missingness mechanism is MAR or MNAR, then follow the steps described earlier (in Section 4.2.2) to simulate missing item scores.

For the missing data patterns other than the patterns “100% MC item scores missing” and “100% CR item scores missing,” the set of items with missing scores differs over the examinees in each replication.
2. Estimate the “imputation models,” which are the psychometric/statistical models underlying the imputation approaches, based on the 95% examinees that did not have any missing item scores—data for these examinees constitute the *model-building data*. For example, for the IRT-based approach, this step involves the fitting of an IRT model to the model-building data.
3. Impute the classifications for the 5% examinees drawn in the first step (data for these examinees are referred to constitute the *test data set*) using the imputation model estimated in the second step and their item scores that are not marked missing in the first step above. For example, for the IRT-based approach, this step involves the application of Equation 2 to impute the PCS on the items with missing scores followed by the computation of the imputed TCS on the total test and, finally, the imputed classification for the 5% examinees with missing scores.

The above steps provided 1,000 sets of imputed classifications for 5% examinees for each imputation approach for each combination of a missing data pattern (of a total of five), a missing data mechanism (of three), and a test (of four). Because the actual or observed classifications of all examinees were available, it is possible to compare the imputed classifications with the corresponding actual classifications to evaluate the accuracy of the imputation approaches. The following measures, which are appropriate agreement measures for ordinal data

(e.g., Williamson et al., 2012), were used in the comparison of the different imputation approaches:

- *the percent exact agreement* between the actual and imputed classifications or the percentage of examinees for whom the imputed classifications were identical to the actual classifications;
- *the classification boundary percentage agreement (CBPA)* between the actual and imputed classifications; these are the percentage of times when both the actual and imputed classifications indicated that the examinees had a classification of c or higher. For Test B that involves three possible classifications (1, 2, and 3), there are two CBPAs, one each for $c = 2$ and $c = 3$. For Tests A1 through A3, there are four CBPAs each.
- *Cohen's κ* : Experts such as Williamson et al. (2012) recommended the use of a measure like Cohen's κ (Cohen, 1960) instead of or in addition to the percentage agreement measure. The Cohen's κ , or, κ henceforth, is defined as $100 \frac{p_o - p_e}{1 - p_e}$, where p_o is the proportion of actual agreement and p_e is the hypothetical probability of a chance agreement. The multiplication by 100 is meant to express κ as a percentage in this article. Thus, κ is smaller than the percentage agreement and is widely considered as a robust measure of agreement. The use of quadratically weighted κ (QWK; e.g., Williamson et al., 2012) led to similar conclusions as the use of κ —so the QWK is not considered henceforth.

Accurate imputation of the missing classifications would result in each of percent exact agreement, CBPAs, and κ to be close to 100. In addition, because the accuracy in imputing the TCSs may be of interest, the bias and the root mean squared difference (RMSD) in imputing the TCSs for the various imputation approaches were also computed, where the bias was computed as the mean of the differences between the imputed and actual TCSs, and RMSD was computed as the square root of the average of the squared differences between the imputed and actual TCSs. The TCS is not imputed in the logistic regression approach—so bias and RMSD were not computed for this approach.

Online Appendix B includes R code for imputation of classification using the PMI, linking, linear regression, logistic regression, and MICE approaches for one replication for the case when 100% MC items are missing for several examinees (the steps for simulating missingness are not shown in the code; instead, it is assumed that the missingness has been simulated and the data set was partitioned into a model-building data set that does not include any missing item scores and a test data set that includes missing item scores).

4.3. Results Under the MCAR Missingness Mechanism

Figures 1 and 2 show the RMSDs and percent exact agreements for all the imputation approaches for the five missing score patterns for the four tests for the MCAR mechanism. In these figures, the missing score patterns are shown along the horizontal axis and the accuracy measures are shown along the vertical axis.

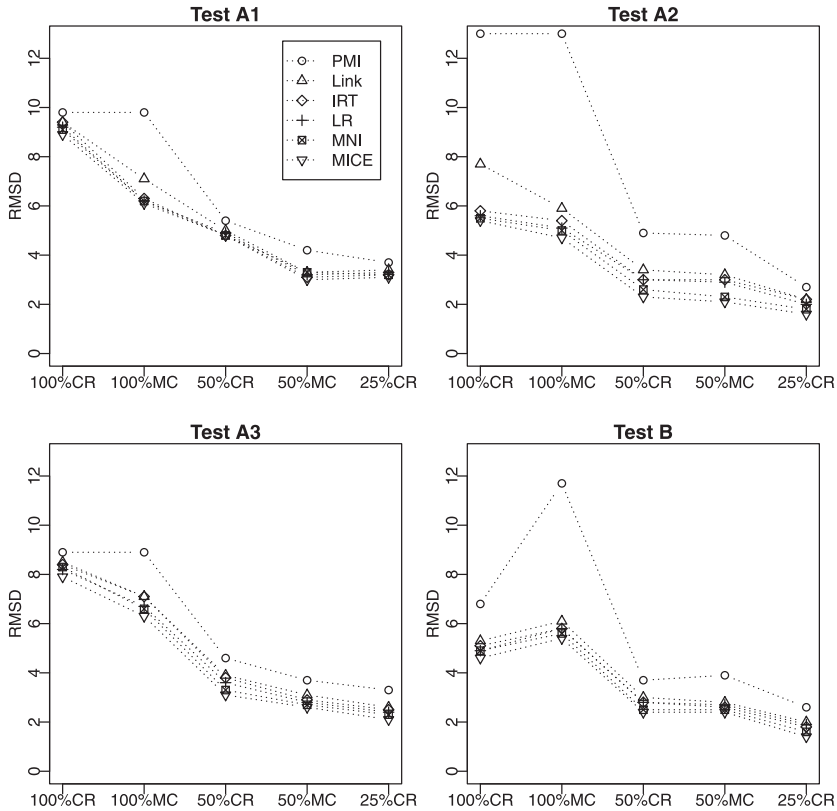


FIGURE 1. Root mean squared differences of the imputation approaches for missing completely at random data. Link = linking; LR = linear regression; MNI = modeling of nonignorable data.

In each figure, the range of the vertical axis is the same over the four panels. The modeling of nonignorable data (Holman & Glas, 2005), linear regression, cumulative logistic regression, and linking approaches are denoted as MNI, LR, CLR, and Link, respectively. The CBPA's and κ show the same patterns as the exact agreement—Table C1 and Figure C1 in Online Appendix C show their values. The values of bias for all imputation approaches and all missing score patterns were 0.0 up to one decimal place except for being equal to 0.1 for the PMI approach for “100% MC item scores missing” and “100% CR item scores missing” patterns—so the values of bias are not presented in this article.

To put the number of missing item scores into context, the reliability of the PCS on the nonmissing part, as a percentage of the reliability of the TCS, where the reliability is computed using the Feldt-Raju procedure (Qualls, 1995), is provided in Table 3.

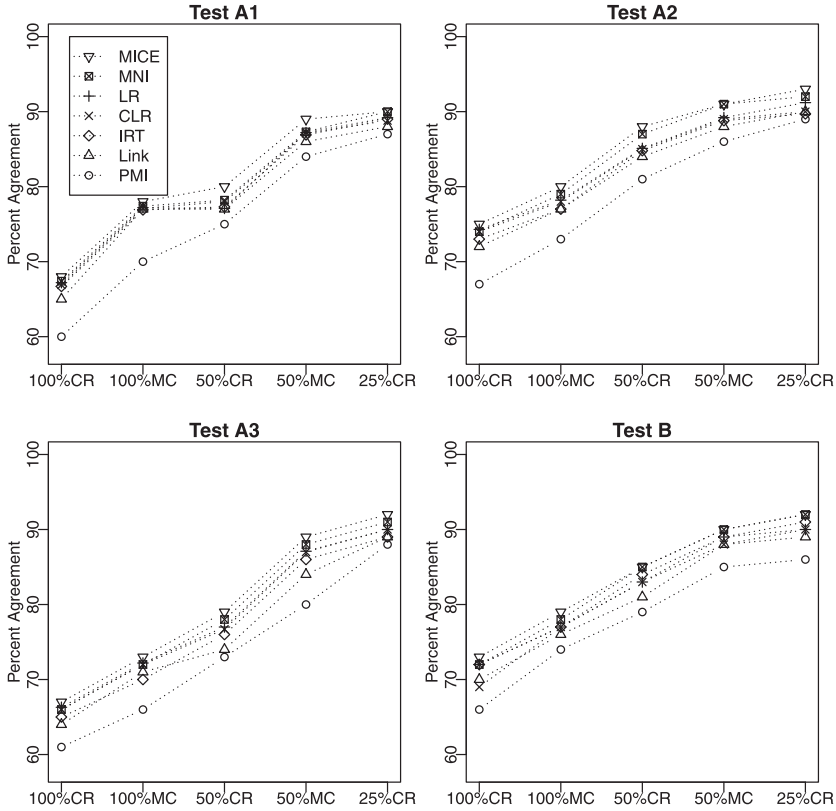


FIGURE 2. Percentage agreement of the imputation approaches for missing completely at random data. MNI = modeling of nonignorable data; LR = linear regression; CLR = cumulative logistic regression; Link = linking.

Figures 1 and 2 show that when data are MCAR,

- The extent of the agreement between the actual and imputed TCS and between the actual and imputed classifications increases on average as the number of missing item scores decreases. For a given data set and a given imputation approach, each of percent exact agreement, CBPA, and κ increases and RMSD decreases as the percentage of missing CR item scores decreases from 100 to 50 to 25.
- The values of percentage agreement for Test A3 are smaller overall than those for the other tests. This is presumably due to the smaller interitem correlations (see Table 2) and smaller percentage of examinees in the two extreme proficiency levels for Test A3 compared to the other tests. The values of percentage agreement are larger in general for Test A2, presumably due to the larger interitem correlations and large score reliability for the test.

TABLE 3.
The Reliabilities as a Percentage of Reliability of the Whole Test When Scores Are Missing Completely at Random

Test	Missing Item Scores				
	100% MC	50% MC	100% CR	50% CR	25% CR
A1	92	96	95	98	99
A2	96	98	93	98	99
A3	92	96	96	98	99
B	95	97	97	98	99

Note. MC = multiple choice.

- For each test, the performances of the imputation approaches do not differ much, with larger differences being observed for more missing item scores. For example, for Test A1, the values of percentage agreement for the approaches vary over a narrow range of 87 to 90 for 25% missing CR item scores while the values vary between a wider range of 60 to 68 for 100% missing CR item scores (see Figure 2).
- The MICE approach seems to lead to the most accurate imputation overall, closely followed by the MNI and regression approaches. The other approaches did not perform much worse except for the PMI approach that was considerably worse than the other approaches. A primary contributor to the accurate imputation for all approaches, presumably, is that the large multiple correlation coefficient of the TCS on the nonmissing item scores is larger than .93 on average for all the conditions.

4.4. Results Under the MAR Missingness Mechanism

The performance of any imputation approach under the MAR mechanism was very similar to that under the MCAR mechanism. The bias, RMDS, and proportion agreement for any combination of imputation approach, missing score pattern, and test was the same up to two decimal places over the MCAR and MAR missingness mechanisms. The similarity of the results over the MCAR and MAR mechanisms is expected from discussion in Finch (2008), Graham (2009, p. 553), and Schafer and Graham (2002) who asserted that the imputation approaches that operate the assumption that data are MAR perform quite well under the MAR mechanism.

4.5. Results Under the MNAR Missingness Mechanism

The mean of the TCS for the model-building data (that comprised 95% of the sample examinees) and the test data (5% of the sample examinees) under the MNAR mechanism for Test A1 were 82.6 and 74.3, respectively, for $\rho = -.18$,

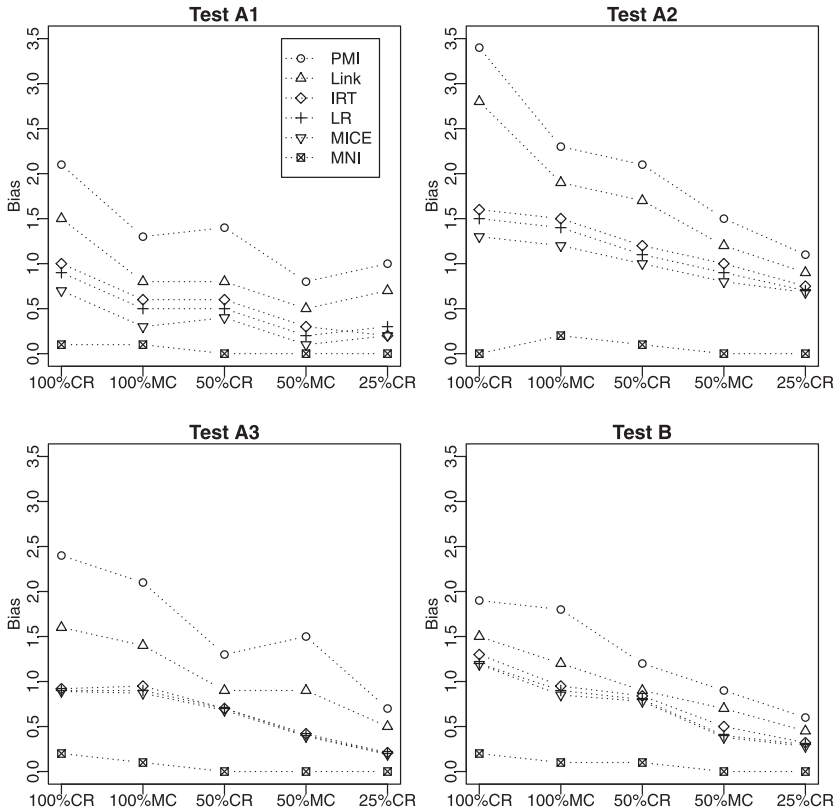


FIGURE 3. Bias of the imputation approaches for missing not at random data with $\rho = -.18$. Link = linking; LR = linear regression; MNI = modeling of nonignorable data.

and 97.1 and 59.9, respectively, for $\rho = -.80$. Thus, the examinees in the model-building data were of higher ability than those in the test data under the MNAR missingness mechanism and more so for the MNAR condition with $\rho = -.80$.

Figures 3 through 5 show the bias, RMSD, and percent exact agreements for all the imputation approaches for the five missing score patterns for the four tests for the MNAR mechanism when the correlation (ρ) between the examinee ability and missingness propensity parameters is equal to $-.18$.

Figures 6 through 8 show the bias, RMSD, and percent exact agreements for all the imputation approaches for the five missing score patterns for the four tests for the MNAR mechanism when ρ is set equal to $-.8$.

The figures show that the MNI approach is the most accurate, followed by the MICE approach, in imputing missing classification for MNAR data.

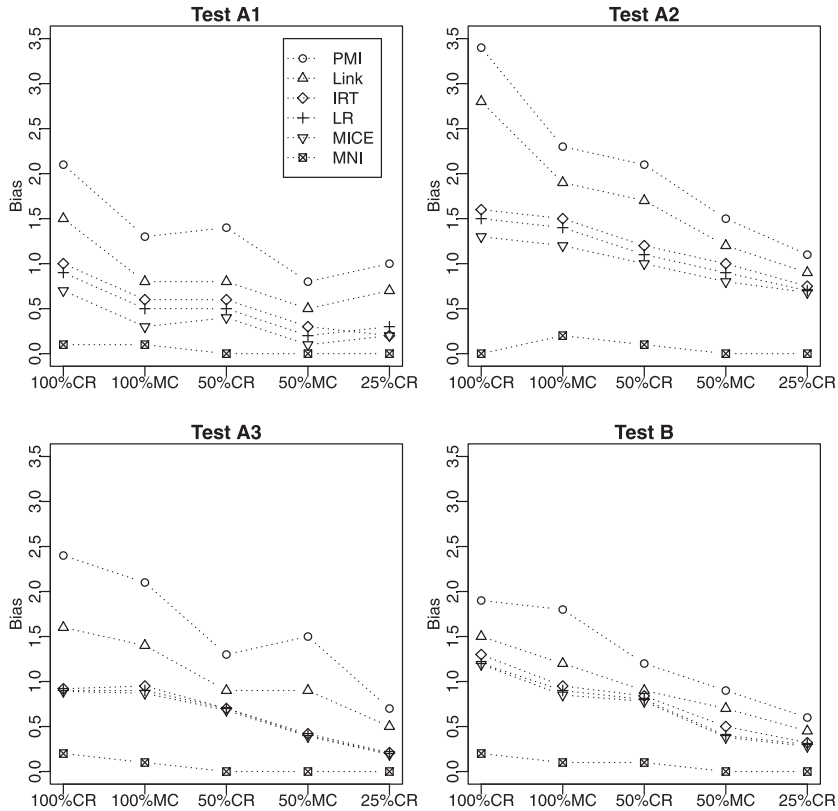


FIGURE 4. Root mean squared differences of the imputation approaches for missing not at random data with $\rho = -.18$. Link = linking; LR = linear regression; MNI = modeling of nonignorable data.

A comparison of Figures 1 and 4 and of Figures 2 and 5 show that for any given missing data pattern, the RMSD or percent exact agreement for any imputation approach other than the MNI approach under a moderate extent of nonignorability ($\rho = -.18$) is close to that under the MCAR mechanism, a result that may seem counterintuitive because experts such as Finch (2008) noted that MNAR data can create great difficulties in accurately imputing missing scores. However, the closeness of the RMSDs and exact agreement is most likely an outcome of the fact that when the scores on up to 50% items are MNAR for an examinee and the extent of nonignorability is small or moderate, the nonmissing item scores provide a substantial amount of information on the missing scores because all the item scores are influenced by the examinee's ability parameter. In other words, even when up to 50% of the item scores are MNAR, the available item scores of

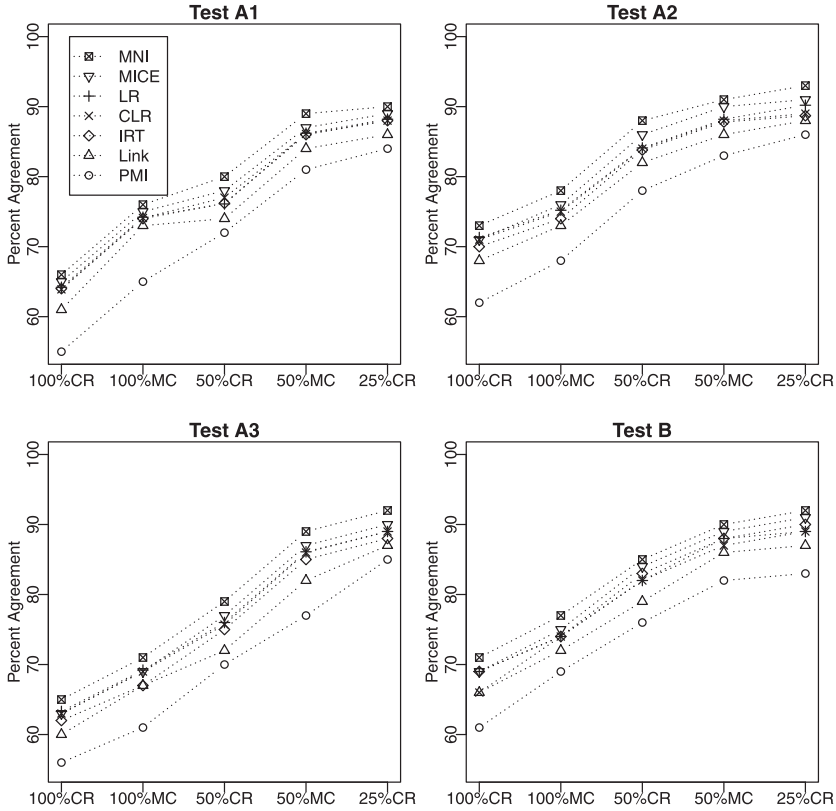


FIGURE 5. Percentage agreement of the imputation approaches for missing not at random data with $\rho = -.18$. MNI = modeling of nonignorable data; LR = linear regression; CLR = cumulative logistic regression; Link = linking.

an examinee provide a substantial amount of information about the examinee ability that largely determines the missing item scores, which, in turn, leads to accurate imputation of the missing scores.¹³ Researchers such as Graham (2009) emphasized the importance of including covariates that predict the missing values in imputation models and asserted that such imputation will yield reasonable estimates even under the MNAR mechanism in the presence of strong covariates; the available item scores of the examinees act as the strong covariates that Graham (2009) referred to. The (mostly) accurate prediction of several approaches under the MNAR mechanism is in agreement similar results in van Ginkel et al. (2010), Xiao and Bulut (2020),¹⁴ and Rose et al. (2010).¹⁵

However, a comparison of Figures 1 and 7 and Figures 2 and 8 show that for any given missing data pattern, the RMSD or percent exact agreement for any approach except for the MNI approach is considerably worse under the MNAR

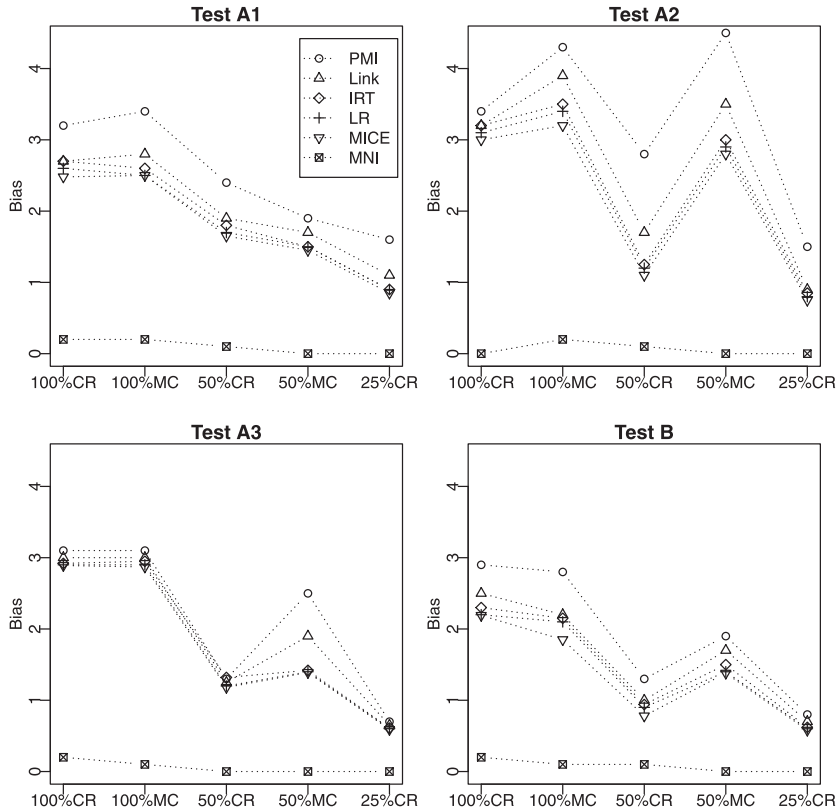


FIGURE 6. Bias of the imputation approaches for missing not at random data with $\rho = -.8$. Link = linking; LR = linear regression; MNI = modeling of nonignorable data.

mechanism than under the MCAR mechanism for the stronger nonignorability condition ($\rho = -.8$). The differences between the results under the MNAR and MCAR mechanisms are larger for the patterns with more missing item scores including the 100% MC and 100% CR missing cases. Figure 6 indicates that this deterioration of performance under strong nonignorability is associated with a large bias in the TCS for the imputation approaches that are based on the MAR assumption.¹⁶

The relative ranking of the imputation approaches with respect to RMSD is similar to that with respect to exact agreement, but the differences in RMSDs between the approaches are often larger than the differences in exact agreement. For 100% missing CR or MC item scores and for Tests A2 or B, the RMSD for the PMI approach is occasionally more than twice that of the MICE approach (Figures 1, 4, and 7), but Figures 2, 5, and 8 show that the percent exact agreement of the PMI approach for the two tests is 90% or above of that of the MICE

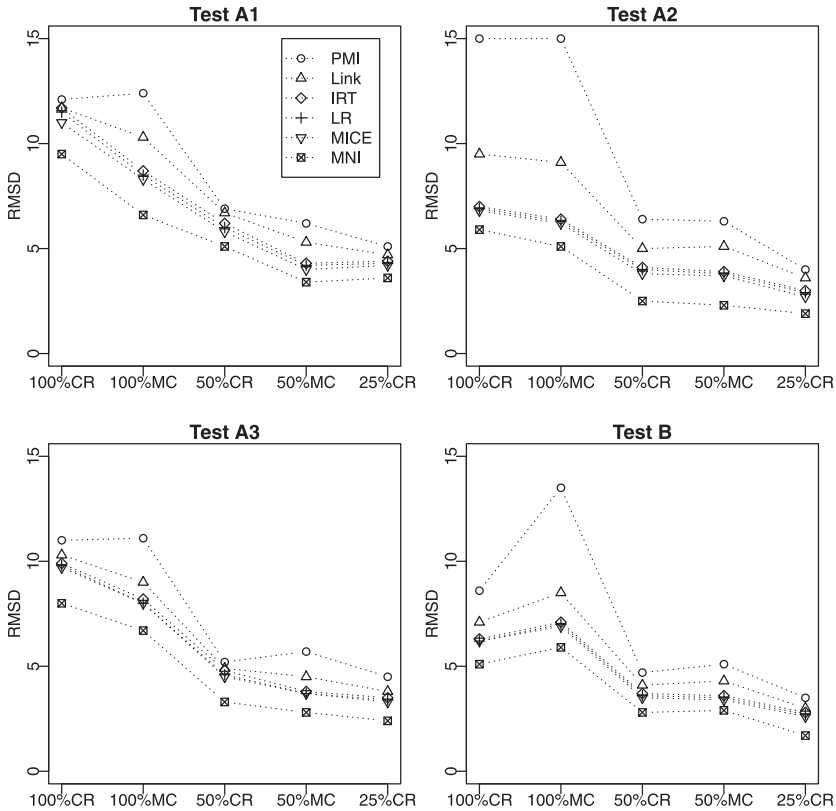


FIGURE 7. Root mean squared differences of the imputation approaches for missing not at random data with $\rho = -.8$. Link = linking; LR = linear regression; MNI = modeling of nonignorable data.

approach. This phenomenon is presumably due to the loss of information resulting from the conversion of the TCSs to a few classifications before computing the percent exact agreement.

4.6. Discussion on the Results of the Comparison of the Imputation Approaches

The results from the above comparisons have the practical implication that if a practitioner has reasons to believe that the missing data are MNAR and is willing to use an approach that is complex and computation-intensive, then the MNI approach (Holman & Glas, 2005) should be chosen as the method of choice. If the practitioner thinks that the missing data are more likely to be MAR, then they should choose the MICE approach if computational burden is not an issue and the linear regression approach if the practitioner is interested in a simple approach

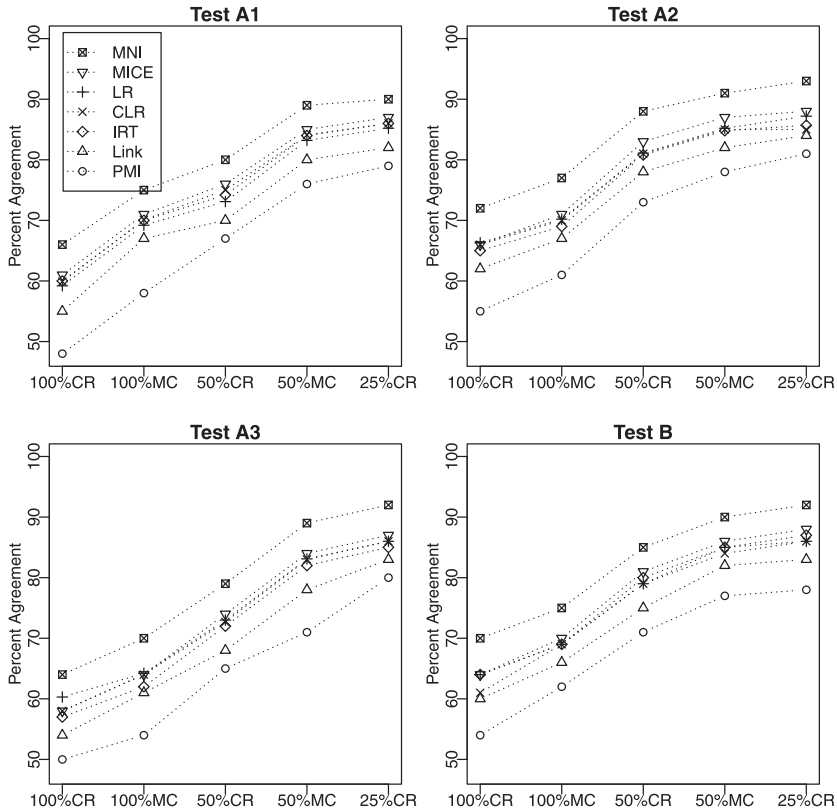


FIGURE 8. *Percentage agreement of the imputation approaches for missing not at random data with $\rho = -.8$. MNI = modeling of nonignorable data; LR = linear regression; CLR = cumulative logistic regression; Link = linking.*

(Sijtsma & van der Ark, 2003, emphasized the importance of using simple approaches in operational practice). A couple of approaches that are more computation-intensive than linear regression—the IRT approach and the logistic regression approach—provided no substantial benefits over the regression approach. It was found from further analysis that most incorrect classifications occurred when the individuals were in the middlemost classes. Also, more than 98% incorrect classifications were off by only one class.

5. Conclusions and Recommendations

Six imputation approaches were brought to bear on the problem of imputing classifications of examinees with incomplete data. These approaches and the operational approach currently used for four high-stakes classification tests were

compared with respect to their accuracy in imputing classifications using data from the corresponding tests. When the data were MNAR, an approach based on modeling the nonignorable missingness (Holman & Glas, 2005) led to the most accurate imputation of classification. An approach based on MI—the MICE approach (Raghunathan et al., 2001)—led to the most accurate imputation of the classification for data that were MCAR or MAR. All approaches except the simple PMI approach performed better than the operational approach. Given that the operational approach (based on linking) is not based on a (statistical) prediction model and the other approaches are, this article demonstrates how statistical methods can be used to improve upon operational practice in an important problem in educational testing.

The differences in the accuracy of classification between the imputation approaches were small for MCAR and MAR missingness and a moderate extent of nonignorability, a result that is in agreement with the finding of various imputation approaches leading to similar estimates in Finch (2008), Huisman and Molenaar (2001), Sinharay (2021b), and Xiao and Bulut (2020) and is an outcome of the large correlations between the item scores on the tests. The difference between the imputation approaches was larger when the extent of nonignorability was strong, that is, when the correlation between examinee ability and missingness propensity was set equal to $-.8$.

The accurate imputation of the classification for the multiple linear regression approach (that performed only slight worse than the best-performing MICE and MNI approaches) for MCAR and MAR missingness and a moderate extent of nonignorability should be good news to practitioners and operational testing programs given the simplicity and ubiquitous nature of linear regression. Researchers such as Sarle (1998) noted that the literature on missing data deals almost exclusively with estimation¹⁷ and the use of linear regression to impute missing data typically leads to poor estimation. Sarle (1998) also noted, however, that when prediction of values is required in the presence of missing data (as in the context of this article), linear regression may lead to excellent predictions. Therefore, the satisfactory performance of the multiple linear regression approach in this article may not be surprising.

One important and somewhat surprising finding in this article is the accurate imputation from several of the imputation approaches including the linear regression and MICE approaches for up to a moderate extent of nonignorability even though these approaches did not explicitly model the missingness. This finding is essentially an outcome of the high interitem correlations, agrees with similar findings in van Ginkel et al. (2010) and Xiao and Bulut (2020), and has the practical implication that the consequences of MNAR missingness in the context of imputing classifications may not be as disastrous as in other contexts (such as in Enders, 2011, who found MAR-based approaches to be highly biased for MNAR data in latent growth curve analysis), at least when faced with a small to moderate extent of missing item scores. Box and Draper (1987, p. 54)

commented that all models are wrong, but some are useful. The results of this article indicate that MAR-based approaches may be wrong but are useful when applied to impute classifications in the presence of up to 50% MNAR data when the extent of nonignorability is small to moderate.

One practical implication of the results from the comparison study is that practitioners may use the MICE or the regression-based approach if computational complexity is an issue and (a) there is a small to moderate extent of missing item scores or (b) the practitioner strongly believes that the extent of nonignorability is not large. Note that it is possible to get some idea of the extent of nonignorability for item-response data using approaches used by, for example, Holman and Glas (2005) and Rose et al. (2010).¹⁸ However, if computational complexity is not an issue, then practitioners should use the MNI approach that would protect them from the worst scenario of strong nonignorability in the data.

The practitioners who would like to adopt an approach to impute classifications in the face of missing item scores for their own data sets could perform a simulation study as performed in this article to compare various approaches. Ideally, they would start with a representative subset of examinees with no missing item scores and then artificially generate missing data according to the mechanism that is anticipated to have occurred for the data, where the missingness pattern is what they actually observed for their data (i.e., if they want to impute classifications of some examinees with scores missing on Items 1 through 10 on a 50-item test, then they should start with the examinees with no missing item scores, simulate missing scores on Items 1 through 10 under different assumptions about the missingness mechanism, and impute their classifications using various approaches). It would then be straightforward to compare the various approaches because the actual classifications are available for the subset of examinees that they started with.

One important consideration regarding imputation of classifications is the uncertainty inherent in the imputation of the imputed classifications, its impact on the decisions based on the classifications, and the ways to report the uncertainty to the score users. One way to report the uncertainty would involve the reporting of the probabilities of the various classifications, which can be readily computed from the cumulative logistic regression approach described earlier. For example, for 100% missing MC items and for Test B, the probabilities of classifications of 1, 2, and 3, respectively, are .94, .04, and .02 for an examinee while they are .59, .24, and .17 for another examinee. While both of them would receive a reported classification of 1, the uncertainty associated with the classification is much smaller for the first examinee compared to the second. However, a problem with classification probabilities is that they may not be easily understood by the score users.

Although the findings of this article may have important practical implications, this article has several limitations, and, consequently, it is possible to perform future research in several related areas. First, one could compare the

imputation approaches using more data sets, both simulated and real, preferably from other types of classification tests including tests that (a) have smaller reliability and/or interitem correlation compared to the tests considered in this article, (b) include items with low psychometric quality (e.g., data from field trials), and (c) include testlets, with missingness depending on testlet membership. Second, it is possible to employ, in future comparison studies, more advanced approaches such as other model-based approaches that apply when data are MNAR (e.g., Enders, 2011; Glas & Pimentel, 2008; Rose et al., 2017) and data mining methods (e.g., Hastie et al., 2009). However, these approaches, especially those designed for the MNAR data, are case-specific and may not be easy or practical to use for large-scale tests due to their complexity. Third, the operational rules for handling missing item scores (such as rules for treating omitted and not-reached responses and threshold for imputation) for the tests were not questioned in this article, but future research could examine whether such rules are too strict or too liberal. Fourth, the results from this study apply only to tests for which the composite score is a linear combination of the item scores and the weights in the linear combination are known in advance—future research could evaluate the comparative performance of the imputation approaches for tests in which the composite score is not a linear combination of the item scores or the weights in the linear combination are not known in advance. Fifth, the MNI approach performed well under the specific MNAR mechanism (the one based on the model of Holman & Glas, 2005) that was considered here and may not perform as well under other MNAR mechanisms. Finally, it is possible to examine whether the relative performance of the imputation approaches is similar over relevant demographic subgroups.


Acknowledgments

The author wishes to express sincere appreciation and gratitude to Steven Culpepper, the editor, and the four anonymous reviewers for their helpful comments. The author would also like to thank Gautam Puhan, Hongwen Guo, and Sooyeon Kim for their helpful comments on an earlier version.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: The author prepared the work as employee of Educational Testing Service. Any opinions expressed in this publication are those of the author and not necessarily of Educational Testing Service.

ORCID iD

Sandip Sinharay  <https://orcid.org/0000-0003-4491-8510>

Notes

1. Note that this decision is made by the test administrators and, in principle, it is possible to allow an indeterminate class for these tests.
2. In practice, one has to plug in the estimated item response theory (IRT) model parameters computed in Step 1 in these expressions.
3. While the BLIMP software leads to very similar results as the mice package for continuous variables, their results may differ when some variables are categorical. That is because the full joint distribution of a set of latent variables underlying the categorical variables is modeled in BLIMP, unlike in the mice package where only the conditional distributions are modeled. The results from BLIMP were found slightly more accurate than those from the mice package for the data sets analyzed in this article.
4. This model, which essentially is the Rasch model, was referred to as G_2 by Holman and Glas (2005) and is a special case, with $\alpha_{di} = 1$, of Equation 5 of Rose et al. (2010). In limited simulations, other models provided by Holman and Glas (2005) performed similar to model G_2 for our data and are not considered henceforth.
5. The small percentage of examinees who had any omitted and not-reached answers were not included in these complete data sets.
6. The average percentage score for a constructed response (CR) item is the average score on the item divided by the maximum possible score on the CR items.
7. For Tests A1, A2, and B that include six CR items, the simulation procedure for the “25% CR item scores missing” case alternates between one missing CR item in one replication and two missing CR items in the next replication—so the percentage of missing CR items over all replications is 25.
8. The testing company often does not have irrefutable evidence that the poor audio quality was due to the examinee behavior and the only option is to report a score to the examinee.
9. This normal distribution is the conditional prior distribution of ξ given θ , computed at $\theta = \hat{\theta}$ under the assumption that ξ and θ follow a joint bivariate normal prior distribution with means 0, variances 1, and correlation ρ .
10. For example, in the case of 50% missing MC scores, each MC item has missing responses for about 50% examinees while, in the case of 25% missing MC scores, each MC item has missing responses for about 25% examinees. To match these values, γ_i had to be set equal to 0.0 and 1.315, respectively.
11. Rose et al. (2010) defined $d_i = 1$ when the score on item i is available and 0 otherwise whereas we define $d_i = 1$ when the score on item i is missing—so their assumed correlation of .8 is equivalent to a correlation of $-.8$ in this article.
12. Limited simulations confirmed that the results for the MAR missingness mechanism did not depend on the procedure used to simulate MAR scores.

13. van Ginkel et al. (2010) provided a similar explanation of a similar finding for missing not at random (MNAR) data.
14. See their Table 2 in which a nonzero bias is observed only when the percentage of MNAR item scores is 70.
15. They concluded that missing at random (MAR)-based imputation led to similar results as MNAR-based imputation for up to a moderate extent of missing data.
16. Note that the MAR-based approaches had bias problems, although to a lesser extent, under moderate nonignorability.
17. In the context of this article, that means, for example, the estimation of the IRT model parameters.
18. In one approach, the larger the absolute value of the estimated correlation in the model of Holman and Glas (2005), the larger is the extent of nonignorability.
19. κ was computed using the function `cohen.kappa` in the R package `psych` (Revelle, 2020).

References

- Agresti, A. (2013). *Categorical data analysis* (3rd ed.). John Wiley Publisher.
- Allen, N. A., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 technical report* (NCES No. 2001-452). U.S. Department of Education, Institute of Education Sciences, Office for Educational Research and Improvement.
- Baker, F. B., & Kim, H. S. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). Marcel Dekker.
- Bernaards, C. A., & Sijtsma, K. (2000). Influence of imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. *Multivariate Behavioral Research*, 35, 321–364.
- Bolsinova, M., & Maris, G. (2016). Can IRT solve the missing data problem in test equating? *Frontiers in Psychology*, 6. <https://doi.org/doi=10.3389/fpsyg.2015.01956>
- Box, G. E. P., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. Wiley.
- Braun, H. I., & Holland, P. W. (1982). Observed score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9–49). Academic Press.
- Byrne, M. R. (2017). *Decisions, concerns, and questions pertaining to two 2017 statewide assessment events involving Algebra I and English II EOCs of the Missouri assessment program*. Report submitted to Governor Eric Grietens. <https://www.moagainstcommoncore.com/2017EOCAssessmentIssues-10-23-17.pdf>
- Cetin-Berber, D. D., Sari, H. I., & Huggins-Manley, A. C. (2019). Imputation methods to deal with missing responses in computerized adaptive multistage testing. *Educational and Psychological Measurement*, 79, 495–511.
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1–29.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.

- College Board. (n.d.). *AP student datafile for schools and districts 2020 layout format*. Retrieved September 20, 2021, from <https://apcentral.collegeboard.org/pdf/ap-data-file-layout-2020.pdf>
- De Ayala, R. J., Plake, B. S., & Impara, J. C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement*, 38, 213–234.
- Educational Testing Service. (2020). *The Praxis test information bulletin 2020-21*. ETS.
- Educational Testing Service. (n.d.). *The Praxis study companion for Spanish: World language*. Retrieved September 20, 2021, from <https://www.ets.org/s/praxis/pdf/5195.pdf>
- Edwards, J. M., & Finch, W. H. (2018). Recursive partitioning methods for data imputation in the context of item response theory: A Monte Carlo simulation. *Psicológica Journal*, 39, 88–117.
- Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods*, 16, 1–16.
- Enders, C. K., Du, H., & Keller, B. T. (2020). A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and non-linear terms. *Psychological Methods*, 25, 88–112.
- Enders, C. K., Keller, B. T., & Levy, R. (2018). A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychological Methods*, 23, 298–317.
- Fayers, P. M., & Hays, R. D. (2014). Should linking replace regression when mapping from profile-based measures to preference-based measures? *Value in Health*, 17, 261–265.
- Feinberg, R. (2021). Estimating classification decisions for incomplete tests. *Educational Measurement: Issues and Practice*, 40(2), 96–105.
- Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, 45, 225–245.
- Gayle, A. (2017). *Dozens of high school students' AP exams missing multiple choice sections, parents angry*. Retrieved May 25, 2021, from <https://wjla.com/news/local/dozens-of-high-school-students-ap-exams-missing-multiple-choice-sections-parents-angry>
- Glas, C. A. W., & Pimentel, J. L. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement*, 68, 907–922.
- Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60, 549–576.
- Graham, J. W. (2012). *Missing data*. Springer.
- Grund, S., Lüdtke, O., & Robitzsch, A. (2020). On the treatment of missing data in background questionnaires in educational large-scale assessments: An evaluation of different procedures. *Journal of Educational and Behavioral Statistics*. Advance online publication. <https://doi.org/10.3102/1076998620959058>
- Hambleton, R. K., & Han, N. (2005). Assessing the fit of IRT models to educational and psychological test data: A five step plan and several graphical displays. In W. R. Lenderking & D. Revicki (Eds.), *Advances in health outcomes research methods, measurement, statistical analysis, and clinical applications* (pp. 57–78). Degnon Associates.

- Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.
- Holman, R., & Glas, C. A. W. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, 58, 1–17.
- Horton, N. J., & Lipsitz, S. R. (2001). Multiple imputation in practice. *The American Statistician*, 55, 244–254.
- Huisman, M. (1999). *Item nonresponse: Occurrence, causes, and imputation to missing answers to test items*. DSWO Press.
- Huisman, M., & Molenaar, I. W. (2001). Imputation of missing scale data with item response models. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 221–244). Springer.
- Jodoin, M. G., & Rubright, J. D. (2020). When examinees cannot test: The pandemic's assault on certification and licensure. *Educational Measurement: Issues and Practice*, 39(3), 31–33.
- Kadengye, D. T., Cools, W., Ceulemans, E., & Van den Noortgate, W. (2012). Simple imputation methods versus direct likelihood analysis for missing item scores in multi-level educational data. *Behavior Research Methods*, 44, 516–531.
- Kang, T., & Chen, T. T. (2008). Performance of the generalized $S - \chi^2$ item-fit index for polytomous IRT models. *Journal of Educational Measurement*, 45, 391–406.
- Köhler, C., Pohl, S., & Carstensen, C. H. (2017). Dealing with item nonresponse in large-scale cognitive assessments: The impact of missing data methods on estimated explanatory relationships. *Journal of Educational Measurement*, 54, 397–419.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking* (3rd ed.). Springer.
- Korobko, O. B., Glas, C. A. W., Bosker, R. J., & Luyten, J. W. (2008). Comparing the difficulty of examination subjects with item response theory. *Journal of Educational Measurement*, 45, 139–157.
- Little, R. A., & Rubin, D. (2002). *Statistical analyses with missing data* (2nd ed.). John Wiley & Sons.
- Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score “equatings.” *Applied Psychological Measurement*, 8, 453–461.
- Mislevy, R. J., & Wu, P.-K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (ETS Research Report Series No. RR-96-30-ONR). Educational Testing Service.
- Moustaki, I., & Knott, M. (2000). Weighting for item non-response in attitude scales by using latent variable models with covariates. *Journal of the Royal Statistical Society: Series A*, 163, 445–459.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159–176.
- Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24, 50–64.
- Patterson, B. F., & Ewing, M. (2013). *Validating the use of AP exam scores for college course placement* (College Board Research Report No. 2013-2). College Board.
- Pohl, S., Graefe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests. *Educational and Psychological Measurement*, 74, 423–452.

- Qualls, A. L. (1995). Estimating the reliability of a test containing multiple item formats. *Applied Measurement in Education*, 8, 111–120.
- Raghunathan, T., Lepkowski, J. M., van Hoewyk, J., & Solenberger, P. W. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27, 85–96.
- Revelle, W. (2020). *psych* (R package version 2.0.9). Evanston, Illinois: Northwestern University.
- Rose, N., von Davier, M., & Nagengast, B. (2017). Modeling omitted and not-reached items in IRT models. *Psychometrika*, 82, 795–819.
- Rose, N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT)* (ETS Research Report No. RR-10-11). Educational Testing Service.
- Rose, N., Wagner, W., Mayer, A., & Nagengast, B. (2019). Model-based manifest and latent composite scores in structural equation models. *Collabra: Psychology*, 5(1), 9.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, Inc.
- Sarle, W. S. (1998). Prediction with missing inputs. In P. P. Wang (Ed.), *JCIS 98 proceedings* (pp. 399–402). Research Triangle Park.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147–177.
- Shin, S. (2009). How to treat omitted responses in Rasch model-based equating. *Practical Assessment, Research, and Evaluation*, 14, 1–8.
- Sijtsma, K., & van der Ark, L. A. (2003). Investigation and treatment of missing item scores in test and questionnaire data. *Multivariate Behavioral Research*, 38, 505–528.
- Sinharay, S. (2021a). Estimating probabilities of passing for examinees with incomplete data in mastery tests. *Educational and Psychological Measurement*. Advance online publication. <https://doi.org/doi=10.1177/00131644211023797>
- Sinharay, S. (2021b). Score reporting for examinees with incomplete data on large-scale assessments. *Educational Measurement: Issues and Practice*, 40(1), 79–91.
- Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods*, 6, 317–329.
- Smits, N., Mellenbergh, G. J., & Vorst, H. C. M. (2002). Alternative missing data techniques to grade point average: Imputing unavailable grades. *Journal of Educational Measurement*, 39, 187–206.
- Sulis, I., & Porcu, M. (2017). Handling missing data in item response theory. Assessing the accuracy of a multiple imputation procedure based on latent class analysis. *Journal of Classification*, 34, 327–359.
- U.S. Medical Licensing Examination. (2020). *2020 bulletin of information*. U.S. Medical Licensing Examination. <https://www.usmle.org/pdfs/bulletin/bulletin2020.pdf>
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45, 1–67.
- van Ginkel, J. R., Sijtsma, K., van der Ark, L. A., & Vermunt, J. K. (2010). Incidence of missing item scores in personality measurement, and simple item-score imputation. *Methodology*, 6, 17–30.

- Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1), 2–13.
- Wolkowitz, A. A., & Skorupski, W. P. (2013). A method for imputing response options for missing data on multiple-choice assessments. *Educational and Psychological Measurement*, 73, 1036–1053.
- Xiao, J., & Bulut, O. (2020). Evaluating the performances of missing data handling methods in ability estimation from sparse data. *Educational and Psychological Measurement*, 80, 932–954.

Author

SANDIP SINHARAY is a Distinguished Presidential Appointee at Educational Testing Service, MS-12T, Rosedale Road, Princeton, NJ 08541, USA; email: ssinharay@ets.org. His research interests include item response theory, assessment of model fit, reporting of subscores, statistical methods for detecting test fraud, and Bayesian methods.

Manuscript received July 31, 2020

Revision received September 8, 2021

Accepted September 13, 2021