Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Spring 2021, Vol. 40, No. 1, pp. 79–91

Score Reporting for Examinees with Incomplete Data on Large-Scale Educational Assessments

Sandip Sinharay, *Educational Testing Service*

Abstract: *Technical difficulties occasionally lead to missing item scores and hence to incomplete data on computerized tests. It is not straightforward to report scores to the examinees whose data are incomplete due to technical difficulties. Such reporting essentially involves imputation of missing scores. In this paper, a simulation study based on data from three educational tests is used to compare the performances of six approaches for imputation of missing scores. One of the approaches, based on data mining, is the first application of its kind to the problem of imputation of missing data. The approach based on data mining and a multiple imputation approach based on chained equations led to the most accurate imputation of missing scores, and hence to most accurate score reporting. A simple approach based on linear regression performed the next best overall. Several recommendations are made regarding the reporting of scores to examinees with incomplete data.*

Keywords: data mining, IRT models, multiple imputation, regression imputation

Computer-based or online tests provide several benefits over paper-and-pencil tests, but are occasionally marred by technical difficulties that result in missing item scores for some examinees. In this paper, technical difficulties refer to *technology-related disruptions* of the examinees, or to *technical problems* in recording, storing, and/or scoring the item responses. Technology-related disruptions are events that disrupt the examinees' testing experiences and are primarily caused by malfunctions of the computers, online systems, or other technological devices through which the tests are delivered. Some examples of such disruptions are involuntary logouts, computer slowdowns, and inaccurate content display. For example, a computer slowdown that occurred when several high school students were in the middle of completing their 2017 end-of-course exams in the state of Missouri led to missing item responses (and hence missing item scores) for 176 students (Byrne, 2017). Technical problems in recording, storing, and/or scoring the item responses may affect online tests that involve electronic recording and online scoring of examinee responses. Examples of such problems include poor audio quality or excessive background noise during recording. Ramanarayanan, Lange, Evanini, Molloy, and Suendermann-Oeft (2017) provided an example where technical problems in recording, storing, and/or scoring the item responses caused some item responses to be unscorable, leading to missing item scores.

Item scores that are missing due to technical difficulties lead to the problem of missing/incomplete data and pose a problem in scores-reporting for the corresponding examinees. The test administrators have the option of not reporting a score to such examinees. Alternately, they have the option of reporting a score after employing an approach for imputation/projection/estimation of the score on the missing part of the test; for example, an imputation approach based on score-linking (e.g., Kolen & Brennan, 2014, p. 487) was employed to report scores to the aforementioned 176 examinees in the state of Missouri. Unfortunately, the administrators of tests marred by technical difficulties have, at their disposal, a scarcity of guidance because of a lack of research on how to report a score to the examinees for whom technical difficulties led to missing data. The goal of this paper is to fill this gap and explore several methodological approaches that would allow test administrators to report accurate, and hence fair and valid, scores to the examinees for whom technical difficulties led to missing item-response data. The approaches examined in this paper essentially are those for imputation of missing data (e.g., Little & Rubin, 2002, p. 20).

This paper is only concerned with the reporting of examinee scores and not with the estimation of population-level/model parameters such as item response theory or IRT item parameters, reliability, and mean scores in the presence of missing item scores. The latter issue has been addressed by researchers such as Finch (2008) and Sijtsma and van der Ark (2003). In addition, technical difficulties typically affect a small fraction of examinees (e.g., Byrne, 2017) and model parameters can be accurately estimated from the large fraction of unaffected examinees for whom complete data are typically available. For almost all educational tests that are marred by technical difficulties, there is usually a well-established policy on the maximum number of missing item scores that will still permit the test administrators to report a score to an examinee. For example, for each of the speaking tests considered later, the test administrators do not impute scores and instruct an examinee to retake the test if the number of missing item scores for the examinee is larger than a predetermined number. No attempt is made in this paper to challenge those policies.

© 2020 by the National Council on Measurement in Education

17453992, 2021, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/emip.12396 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Omitted and/or not-reached responses, which also lead to missing item scores, are common in tests with multiple-choice (MC) items even when technical difficulties do not occur. Researchers such as De Ayala, Plake, and Impara (2001), Glas and Pimentel (2008), and Rose, von Davier, and Nagengast (2017) explored various approaches of handling omitted and not-reached responses. The goal of this paper is not to examine or recommend any approach for the imputation of scores for the omitted and/or not-reached responses, which is mainly because most operational tests have well-established policies (which typically do not involve imputation) about omitted and not-reached responses in the absence of technical difficulties. For example, in the National Assessment of Educational Progress, not-reached items are treated as not presented and the omitted responses are assigned a score equal to the reciprocal of the number of answer choices if the item is MC and the lowest possible score if the item is a constructed-response (CR) item (e.g., Allen, Donoghue, & Schoeps, 2001). The operational practices regarding the omitted and not-reached responses for the tests considered in this paper will be assumed to be appropriate.

Data: Two Speaking Tests and a State Test

Data were available from two large-scale online and nonadaptive speaking tests that are administered in several countries. The tests are administered on computers in various testing centers that are strictly proctored. The tests will be referred to as Tests A and B, respectively. The available data from Tests A and B included 11 and 6 items, respectively. The sample sizes were 3,920 and 10,227 for Tests A and B, respectively. The test items measure various aspects of the candidates' ability to speak English. For each item, the examinees receive specific directions including the time allowed for preparing and speaking the responses before they have to speak their responses. The responses are captured by a microphone, digitized, and later assigned an integer score between 0 and m, by trained human raters using a holistic scoring rubric, where m is a positive integer that varies over items. The sum of the item scores is computed to yield a total/raw score for each examinee. Then a transformation is used to convert the total score to a scale score that is reported to the examinees. Occasionally, technical problems in recording, storing, and/or scoring the item responses make it impossible to assign scores to some examinees on some items, which leads to the problem of missing scores for the corresponding examinees. Technical problems constitute the only source of missing item scores for these two tests. When only a few item scores are missing for an examinee, a score is reported to the examinee after imputing the missing item scores based on the available item scores. No score is reported to the examinees for whom more than a few item scores are missing—these examinees are allowed a free retest.

Data from 12,312 examinees were also available from a computerized nonadaptive mathematics test that belongs to a United States state testing program. The test included 28 MC (scored 0/1) and 8 CR (CR; scored 0–2 or 0–4) items. The maximum possible total raw score across the CR items is about one-third of the maximum possible raw score on the test. A raw total score is computed for all the examinees, and then, the IRT true score equating procedure (e.g., Kolen & Brennan, 2014, p. 193) is used to convert the raw total score to a scale score. Currently, omitted and not-reached responses are the only known sources of missing scores for the test—these responses are assigned item scores of 0. However, given the several recent instances of technical difficulty during state tests (e.g., Byrne, 2017; Sinharay et al., 2014) and the potential future technology-related disruptions due to poor Internet access on at-home tests (e.g., Michel, 2020), the test administrators should have a strategy for reporting scores to the examinees when some item scores are missing due to technical difficulties.

Background

Literature Review and Motivation of the Research

Extensive reviews of the literature on missing data analysis were performed by Graham (2009), Graham (2012), Schafer (1997), Schafer and Graham (2002), Sinharay, Stern, and Russell (2001), and Vriens and Sinharay (2006). In addition, researchers have considered a wide variety of problems regarding missing data in educational or psychological measurement (e.g., Cetin-Berber, Sari, & Huggins-Manley, 2019; De Ayala, Plake, & Impara, 2001; Finch, 2008; Holman & Glas, 2005; Huisman & Molenaar, 2001; Smits, Mellenbergh, & Vorst, 2002; Sijtsma & van der Ark, 2003; Xiao & Bulut, 2020). A brief summary of most of these studies is included in the first subsection of the Supporting Information. The main findings of these studies varied—one finding that is important in the context of this paper is that several of the researchers including Finch (2008), Smits et al. (2002), and Sulis and Porcu (2017) found the multiple imputation (MI) approach (e.g., Rubin, 1987, p. 2) to perform the best or close to the best in estimating the quantity of interest. The summary of the studies reveals a lack of research focusing on the reporting of scores to examinees with missing item scores on large-scale and high-stakes educational tests. Huisman and Molenaar (2001) examined the imputation of the total/raw score, but they dealt with psychology tests, did not use data from any educational tests, and did not use more recent imputation approaches such as MI (Rubin, 1987, p. 2)

Missing Data Mechanisms and Technical Difficulties

The missingness or the probability of having a missing value 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) missing not at random (MNAR). Alternatively, the missing data mechanisms are described as the assumptions that investigators make about the process that led to the missingness in their data, or as explanations of why the missing data are missing. The properties of the methods to analyze missing data depend strongly on the nature of the dependencies predicated by these missing data mechanisms.

If missingness or the probability of having a missing value does not depend on the values of the data, including missing and observed data, then the data (and the missing data mechanism) are referred to as MCAR. If missingness depends on the component of the data that is observed, then the data are referred to as MAR. If missingness depends on the component of the data that is missing, then the data are referred to as MNAR. There exist approaches suggested by, for example, Little (1988) and Sijtsma and van der Ark (2003), for

17453992, 2021, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/emip.12396 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

testing whether data are MCAR, but it is not possible to examine whether data are MAR or MNAR (e.g., Rubin, 1987, p. 155) though some forms of MAR can be assessed by the test of Potthoff, Tudor, Pieper, and Hasselblad (2006). Unfortunately, the approaches for assessing whether the data are MCAR or MAR have low power so that they are likely to not be helpful for most educational applications.

For educational tests, missing data for an examinee is MCAR when the missingness is unrelated to any data including the item scores and examinee covariates. The missing data are MAR if, for example, the probability of a missing score for an examinee on an item depends on the examinee's total score on the remaining items or on the examinee's background characteristics. The MNAR missing data mechanism may arise for educational tests if the probability of a missing score for an examinee on an item is directly related to the score itself; this type of missingness may occur if those who are likely to perform poorly on an item on a speaking test make random noises to cause the recording to be unscorable. Note that the term MNAR is slightly misleading. For example, if a random sample of examinees has missing scores on the two most difficult items on the test, the resulting data are MCAR and are not MNAR.

Item scores that are missing due to technical difficulties may appear to be MCAR because of the perception that such difficulties are completely random occurrences. However, there is some evidence that such scores may actually be MAR or MNAR. For example, Sinharay et al. (2014) reported that more technology-related disruptions occurred in the first few days of a state testing window and more low-performing students typically test on the first few days; as a consequence, the average score on the previous year's test of the students who experienced technology-related disruptions in the current year was 11 scale-score points smaller than those who did not experience technology-related disruptions. Thus, for the test, technology-related disruptions were related to an examinee covariate (previous year's scores) and the item scores missing due to the disruptions were MAR. In addition, some research has revealed that for one of the speaking tests discussed earlier, the missing item responses are very likely to be MNAR (J. R. Lockwood, personal communication, December 30, 2019).

Imputation Approaches

Six imputation approaches were suggested or considered in this paper. The first approach was chosen because it is similar to the approach that is operationally used for Tests A and B to report imputed scores and is a very simple approach. The next four approaches were chosen because they arguably are the most popular imputation approaches in educational and psychological measurement. The last approach is based on data mining (e.g., Hastie, Tibshirani, & Friedman, 2009) and has not been applied to impute missing scores, but was chosen because data-mining approaches have been found useful in other prediction problems and have the potential to lead to accurate imputation. The approaches are explained below using the hypothetical example of a test that includes 10 items and a hypothetical Examinee 1 whose scores on Items 9 and 10 are missing due to technical difficulties and scores on Items 1–8 are available. Let us further assume that the possible scores on Item 9 are 0, 1, 2, 3 and those on Item 10 are 0, 1, 2, 3, 4 and that the generalized partial credit model (GPCM; Muraki, 1992) is operationally used to report scores on the test. The description of each imputation approach focuses on the computation of the imputed (raw) total score for an examinee. In operational practice, after the computation of the imputed total score, a transformation has to be applied to the imputed total score to obtain an imputed scale score.

Person-Mean Imputation

The person-mean imputation (PMI) approach (e.g., Huisman, 1999), also known as *proration* of the test scores, involves the imputation of the missing score on an item of an examinee by the mean score on the other items 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 the proportional scores on all items for the examinee and then compute the average of the proportional scores over Items 1–8; imputed scores on Items 9 and 10 for the examinee can be obtained by multiplying this average by 3 and 4, respectively. Then the imputed total score on the test for Examinee 1 can be computed as the sum of the actual scores on Items 1–8 plus the sum of the imputed scores on Items 9 and 10. The PMI approach is similar (but not identical) to the approach that is operationally used for Tests A and B to report scores to examinees with incomplete data. The PMI approach is likely to lead to biased imputed scores irrespective of the missing data mechanism, especially when the items vary in difficulty, because the difficulty of the items is ignored in the approaches (e.g., van Ginkel, Sijtsma, van der Ark, & Vermunt, 2010).

Linking

In the application of the linking approach to impute a score for Examinee 1 for the above-mentioned test, score linking (e.g., Kolen & Brennan, 2014, p. 487) is performed of the total score on the first eight items to the total score on all the items using the data from the subsample of examinees whose scores are available on all 10 items on the test. Then, the imputed total score on the whole test for the examinee is obtained as the equated value of the examinee's total score on Items 1–8. The linking approach was used to report missing scores for 176 examinees for the Missouri state test on algebra in 2017 (Byrne, 2017). The single-group equipercentile equating (e.g., Kolen & Brennan, 2014, p. 14, 36) approach was used to perform the score linking in this paper.

Regression Imputation

In the application of the regression imputation approach to impute a score for Examinee 1 for the above-mentioned test, a linear regression of Y on $X_1, X_2, \ldots, X_8$ is fitted using the subsample of examinees who had scores available on all 10 items on the test, where

$$\mathrm{Y} = \text{Score on item 9} + \text{Score on item 10},$$

$$X_1 = \text{Score on Item 1}, X_2 = \text{Score on Item 2}, \ldots, X_8 = \text{Score on Item 8}.$$

Then, the examinee's scores on Items 1–8 are entered in the fitted regression equation to compute the imputed sum score

© 2020 by the National Council on Measurement in Education

17453992, 2021, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/emip.12396 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

on items 9 and 10 for the examinee; this imputed sum score is added to the actual total score on Items 1–8 to obtain an imputed total score for Examinee 1. The advantage of this approach is the conceptually simplicity and the ubiquitous nature of linear regression, but the approach may not perform well if the underlying relationship between the item scores is nonlinear. Regression imputation was found to lead to accurate prediction of missing grades by Smits et al. (2002). The use of logistic regression instead of linear regression in the regression imputation approach was found to lead to slightly worse imputations in the simulations reported later and hence is not considered henceforth.

Imputation Based on Item Response Theory

Researchers such as Korobko, Glas, Bosker, and Luyten (2008) suggested that the missing score on an item for an examinee can be imputed by its posterior expectation under an IRT model after the IRT model parameters have been estimated using an examinee sample. In this paper, the three-parameter logistic model 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). To implement this approach, one first estimates the parameters of the IRT model using the subsample of examinees who had scores available on all 10 items of the test. One then imputes the sum of the scores on Items 9 and 10 for Examinee 1 as

$$\int_{\theta} E(Y|\theta)p(\theta|X_1, X_2, \ldots, X_8)d\theta, \tag{1}$$

where $p(\theta|X_1, X_2, \ldots, X_8)$ is the posterior distribution of the ability of Examinee 1 (e.g., Baker & Kim, 2004, p. 159) given the scores on Items 1–8 and $E(Y|\theta)$ is the expected value of the sum of the scores on Items 9 and 10 conditional on θ. The quantity $E(Y|\theta)$ can be computed as

$$\begin{aligned} E(Y|\theta) &= E(X_9|\theta) + E(X_{10}|\theta) \\ &= \sum_{k=1}^{3} kP(X_9 = k|\theta) + \sum_{k=1}^{4} kP(X_{10} = k|\theta), \end{aligned}$$

where $P(X_9 = k|\theta)$ and $P(X_{10} = k|\theta)$ are given by, for example,

$$P(X_9 = k|\theta) = \frac{\exp[k\alpha\theta - \sum_{h=1}^{k}\beta_h]}{1 + \sum_{h=1}^{m}\exp[h\alpha\theta - \sum_{l=1}^{h}\beta_l]},$$

where α and β_h, respectively, denote the slope and difficulty parameters for the item. For convenience, the notation in this paper does not reflect the fact that the quantities such as $p(\theta|X_1, X_2, \ldots, X_8)$, $E(Y|\theta)$, and $P(X_9 = k|\theta)$ are estimates (as they depend on the item parameter estimates). One finally computes the imputed total score of Examinee 1 as the imputed sum of the scores on Items 9 and 10 plus the actual total score on Items 1–8. The R package mirt (Chalmers, 2012) was used to fit the IRT models in this paper. The integral in Equation (1) was approximated using numerical integration. Huisman and Molenaar (2001) found an approach based on the Rasch model to lead to the most accurate imputation of the total score for psychological test data sets. The IRT-based approach in this paper is based on the more general three-parameter logistic model and the GPCM is expected to lead to accurate imputation. Note that it is possible to use other IRT-based imputation approaches including one where $\hat{\theta}$, the estimated examinee ability based on the scores on Items 1–8, is computed followed by the addition of $E(Y|\hat{\theta})$ to the total score on Items 1–8. In limited simulations, these other approaches performed very similarly to the approach used in this paper and hence are not considered henceforth.

Multiple Imputation Using Data Augmentation and Chained Equations

Several researchers (e.g., Finch, 2008; Smits et al., 2002; Sulis & Porcu, 2017) have found that the MI (e.g., Rubin, 1987, p. 2) approach leads to the most accurate estimation of various quantities of interest (such as item parameter and reliability) in the presence of missing data including education and psychology. To apply MI to impute missing data, one first creates several (typically, somewhere between 5 and 20) copies of the incomplete data set, filling in each missing value with a different set of plausible replacement values, typically using a probability model for the data. The complete data sets are then analyzed, and the resulting parameter estimates and standard errors are pooled into a single set of results. For example, to apply MI to Examinee 1, one would fill in the (missing) values of X_9 and X_{10} multiple times and treat those filled-in values as the multiple imputed values of the scores on Items 9 and 10. The MI approach used by, for example, Finch (2008) is based on data augmentation (DA) in which the joint distribution of the variables is assumed to be multivariate normal—the assumption may not be appropriate for dichotomous and polytomous item scores. An alternative MI approach was suggested by Raghunathan, Lepkowski, van Hoewyk, and Solenberger (2001) and is referred to as MI using chained equations (MICEs). The approach is also known as the fully conditional specification (FCS) approach. 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. A major advantage of the MICE approach is that the conditional distributions of the variables (item scores in the context of this paper) do not have to be assumed to be normal and can be specified to be models such as the ordered logit model that are more appropriate than normal models for (discrete) item scores. The MICE approach has been found to lead to accurate imputation in comparison studies by researchers such as Horton and Lipsitz (2001) and van Buuren, Brand, Groothuis-Oudshoorn, and Rubin (2006). The R package *mice* (e.g., van Buuren & Groothuis-Oudshoorn, 2011) was used to implement the FCS approach for the data examples in this paper. In the application of the MICE approach in this paper, the conditional distributions are assumed to be the logistic regression model for binary item scores and the ordered logit model for polytomous item scores. In this paper, five sets of draws/imputations of missing item scores were used for the MICE approach; each draw was used to compute an imputed total score; the final imputed total score was obtained as the simple average of the five imputed total scores. This strategy is in agreement with the recommendation on combining results from MIs by (Rubin, 1987, p. 76), who

17453992, 2021, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/emip.12296 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

suggested estimating a quantity of interest by the simple average of the same quantity computed from the MIs.

Data Mining Methods

The problem of imputation of missing item scores can be considered to be one of prediction of some item scores given the scores on some other items (e.g., Huisman & Molenaar, 2001). Several computation-intensive methods for prediction are available in the field of *data mining* (e.g., Hastie et al., 2009; Strobl, 2013). Data mining methods for prediction are becoming increasingly popular in various scientific fields including social sciences, genetics, epidemiology, medicine, education, and psychology (see, for example, Hastie et al., 2009; Sinharay, 2016; Strobl, Malley, & Tutz, 2009). These methods typically are not based on probability models and are algorithmic in nature, but they often provide better prediction than traditional prediction methods such as linear and logistic regression (e.g., Fernandez-Delgado, Cernadas, Barro, & Amorim, 2014) and have the potential to perform satisfactorily in predicting missing item scores. Data mining methods for prediction have been applied in educational measurement to predict essay scores by Chen, Fife, Bejar, and Rupp (2016) and Sinharay, Zhang, and Deane (2019) and to predict essay scores, item parameters, and high-school dropout status by Sinharay (2016). However, applications of data mining methods to impute missing item scores are rare, and hence, this paper is one of the first to apply data mining methods to impute missing scores. Among the data mining methods for prediction, the gradient boosting machine (GBM; e.g., Friedman, 2001) approach was considered because of its superior performance in, for example, Sinharay (2016). The GBM can be considered to be a nonparametric method for prediction, whereas the regression imputation, imputation based on IRT models, and MICE approaches can be considered as parametric methods for prediction.

Tukey (1977) suggested a smoothing procedure referred to as *twicing*. In the first step of twicing, a regression model is used to predict a response variable from several predictors and the residuals are computed from the regression; in the second step, a regression model is used to predict these residuals from the same set of predictors. The final prediction is obtained as the sum of the predictions from the two steps; that is, the final prediction is a combined prediction from two regressions. Tukey (1977) mentioned the possibility of iterating the process. The GBM approach is a refined version of twicing and combines predictions from several regression trees (e.g., Breiman, Friedman, Olshen, & Stone, 1984), which are a type of nonparametric regression models, fitted to subsets of the same data set. A description of regression trees is provided in the second subsection of the Supporting Information. In an application of the GBM, the regression trees are fitted sequentially and each regression tree is constructed using information from previously grown regression trees, just like in *twicing*—the construction process of a regression tree places an increased emphasis on observations that were poorly modeled by the previously constructed tree. Any step of GBM involves the fitting of a nonparametric regression tree that is rather small, with just a couple of predictors (James, Witten, Hastie, and Tibshirani, 2013, p. 322), to the data. Then, one computes residuals from the small regression tree and fits a new regression tree, again small, to these residuals. The process of computing residuals and fitting small trees continues for a large number of times. By fitting small trees to the residuals, one slowly improves the prediction in areas where prediction till then has not been satisfactory. To obtain a predicted value for a new observation from the GBM approach, each of the aforementioned regression trees is used to obtain a predicted value for the observation. The final predicted value is calculated as the average of these predicted values multiplied by a *shrinkage parameter* that is typically a positive number considerably smaller than 1—the multiplication by the shrinkage parameter effectively shrinks the contribution of each regression tree to the final predicted value and avoids overfitting the observations (Elith, Leathwick, & Hastie, 2008). The GBM approach also includes a random/stochastic component as the approach employs only a random subset of the data to construct each tree. Research has shown that random subsets that are about half as large as the data set leads to the most accurate prediction (Elith et al., 2008).

The GBM approach can be implemented using several R packages. Among them, the *gbm* (Ridgeway, 2017) package has been found satisfactory in several disciplines by, for example, James et al. (2013) and Sinharay (2016)—so that package was used in this paper. The tuning parameters—the number of iterations, the shrinkage parameter, and the *depth* of the regression trees—for the GBM approach were determined from a preliminary analysis aimed at choosing the values of these parameters that lead to superior predictions. Such a preliminary analysis also ensured that the approach did not overfit the data. The values 5,000, .001, and 5, respectively, of these parameters were chosen because those led to near-optimum performance of the GBM approach for all the three data sets. The *gbm* package utilizes subsampling from the whole sample to fit the prediction model at any step—so, imputation using the package is similar to the application of a single imputation approach. Because the MI approach has several advantages over the single imputation (e.g., Rubin, 1987, p. 16), the *gbm* package was run five times for each data set, the total score was computed for each of these runs, and the average of the total scores from the five runs was used as the final imputed total score from the GBM approach.

Data mining approaches such as the GBM have been criticized for their black-box nature and for not explaining the relationships among the variables, but are widely acknowledged as useful in prediction of an outcome variable from several explanatory variables (e.g., Breiman, 2001; Shmueli, 2010). The GBM is used as a prediction tool in this paper—so the approach would adequately serve the purpose of this paper.

Missing Data Mechanisms and the Imputation Approaches

Among the imputation approaches discussed earlier, the last four, that is, all but the PMI and linking, are expected to lead to very accurate imputed scores if the missing item scores are MAR or MCAR, given the findings and/or recommendations of researchers such as Finch (2008), Schafer and Graham (2002), and Smits et al. (2002). The missing data literature shows that if the data are MNAR, an imputation approach has to correctly model the missing data mechanism to lead to accurate imputation. However, approaches specific to MNAR are rarely convenient due to at least three reasons. First, these approaches are case-specific. Second, these approaches are likely to be incorrect because it is rarely feasible to model the missing data mechanism with any degree of confidence (Little & Rubin, 2002, p. 22), which leads to the resulting imputations being inaccurate and often worse

© 2020 by the National Council on Measurement in Education

17453992, 2021, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/emip.12396 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

than those under MAR or even MCAR assumptions. Third, these approaches may be computationally intensive. Instead, several experts such as Sinharay et al. (2001) recommended that even if the MAR assumption is incorrect, MI with a good set of covariates may produce estimates with little bias and may be more convenient than an imputation approach that is designed for MNAR data. Little and Rubin (2002, p. 19) noted that in some empirical settings, the MAR assumption has been found to yield more accurate predictions of the missing values than methods based on the more natural MNAR mechanism. van Ginkel et al. (2010) also found their imputation approach, which is expected to work only for MAR item-response data, to lead to accurate estimates of summaries such as mean score and coefficient α under MNAR. It will be interesting to find out how the imputation approaches perform when item scores are MNAR in the simulations described in the next section.

Methods: A Comparison of Six Imputation Approaches

In this section, the above-mentioned six imputation approaches were compared using data that were simulated based on real data. Some item scores were artificially assigned to be missing in the simulated data to make the latter look like those containing missing item scores due to technical difficulties. The need for imputation approaches that are not too complex has been emphasized by researchers such as Sijtsma and van der Ark (2003) who argued that many practitioners who have the responsibility to impute scores do not have a statistician who can help them to implement complex approaches; thus, the complexity of the approaches will also be considered in the comparison and in determining the best imputation approach.

In the first step of the analysis, subsets of data from the original data sets from the three above-mentioned tests were chosen so as to only include the examinees whose scores were available on all items (or, examinees with complete records). The original data from the state test included no missing scores—so the whole of the state test data set was chosen in this step. The three resulting data sets include no scores missing due to technical difficulties and would be referred to as the "complete data sets." To compare the imputation approaches, different parts of the complete data sets were assumed missing and the missing parts were imputed by the different approaches.

Study Design and Computation

Several simulation conditions were considered, where a simulation condition is characterized by one level of each of three factors—the missing data pattern, the missing data mechanism, and a test—with each factor being fully crossed with the others. Three missing data mechanisms were considered and the three above-mentioned tests were considered. The number of missing data patterns was three for Tests A and B and five for the state test. Thus, there were 27 simulation conditions for each of Tests A and B and 45 conditions for the state test.

Missing data patterns considered. Historically, for both Tests A and B, numerous patterns of missing item scores have resulted due to technical difficulties. For example, past data for Test B reveal that the number of items on which an examinee experienced technical difficulties (leading to missing scores) can be any value between 0 and the number of items on the test. However, for Tests A and B, the comparison study includes only the patterns with up to three item scores missing due to technical difficulty. For the state test, five missing data patterns were considered—each pattern involved some missing scores on MC items and some missing scores on CR items. Larger numbers of missing item scores would most often lead to retests (and no score reporting) according to the policy for these tests.

Missing data mechanisms considered. Missing data were simulated under each of MCAR, MAR, and MNAR mechanisms. To simulate MCAR data, the items with missing scores were chosen completely at random from the set of all items.

To generate MAR scores for one item (referred to as *the target item*), borrowing on the methodologies outlined in Finch (2008) for creating MAR item-response data, the raw score was calculated for each examinee on all but the target item. The examinees were then divided into four fractiles based on this raw score. Members of each fractile were assigned a probability of a missing response, with lower scores having a higher probability of a missing response. The average of these probabilities across the fractiles was equal to .30 (thus, 30% examinees were assumed to have some missing item scores). For each examinee, a uniform random number between 0 and 1 was generated and compared with the probability of a missing response.

To generate MNAR scores for one item, borrowing on the methodologies outlined in Finch (2008), the examinees were divided into as many groups as the number of possible scores on the item. Each examinee was then assigned a probability of a missing response, with individuals having smaller scores being assigned a higher probability of that item score being missing. The mean of these probabilities of a missing item score was equal to .30. As a simple example, consider a data set with 10 examinees in which two examinees each receive scores of 0, 1, 2, 3, and 4 on an item. In this case, the examinees with scores of 0, 1, 2, 3, and 4 would be assigned probabilities of missing item scores of .5, .4, .3, .2, and .1, respectively, so that the average probability of missing on the item was .30. For each examinee, a uniform random number between 0 and 1 was generated and the item score was set as missing if the resulting value was lower than the assigned probability of a missing value. Thus, the examinees who are chosen as those with missing scores under the MNAR mechanism have smaller scores on average compared to the remaining examinees on the items with missing scores.

Steps in the comparison and computation. The comparison of the imputation approaches was based on 1,000 iterations of the following steps for each simulation condition:

1. Randomly choose one or more items whose scores will be assumed missing for 30% examinees. The number of items with missing scores is determined by the missing data pattern that is being considered. For example, if the missing data pattern is "two missing item scores," then a simple random sample of two items is drawn from the set of all items and the scores of 30% examinees are assumed to be missing on the two items. The set of two items with missing scores is fixed in each iteration, but is allowed to vary over the 1,000 iterations.

17453992, 2021, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/emip.12396 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2. From the complete data set, draw a subsample of 30% examinees[1] who would be treated as those with missing item scores due to technical difficulty. This subsample is typically referred to as the *validation* (or *cross-validation*) data. The remaining 70% examinees in the complete data set comprise the "model-building data," or data on which the imputation model is estimated. Depending on the missingness mechanism for the simulation case, draw the sample of 30% examinees with missing scores on an item (the item must have been drawn in the previous step as one with missing scores) according to the MCAR, MAR, or MNAR mechanism. The set of 30% examinees with missing scores is allowed to vary over the 1,000 iterations.
3. Estimate the psychometric or statistical models (henceforth referred to as the "imputation model") inherent in the imputation approaches based on 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.
4. Impute the missing scores for the validation data drawn in the second step above using (a) the imputation model estimated in the third step above and (b) the item scores that are not assumed missing in the first step above. For example, for the IRT-based approach, this step involves the application of an equation like Equation (1) to impute the sum score on the items with missing scores followed by the computation of the imputed total score on the test for the examinees in the validation data.
5. For the examinees in the validation data, the imputed total scores were converted to imputed scale scores using the appropriate conversion table for the test.

The above steps provided 1,000 sets of imputed scale scores for the examinees in the validation data for each imputation approach for each simulation condition.[2] Because the actual (observed) total and hence scale scores of these examinees are available, it is possible to compare the imputed scale scores with the actual scale scores to evaluate the accuracy of the imputation approaches. Four measures were used in the comparisons: bias, root mean squared difference (RMSD), percent of zero-scale score differences, and graphical plots. For each imputation approach and each simulation condition, bias was computed as the average of the differences between the imputed and actual scale scores, the RMSD was computed as the square root of the average of the squared differences between the imputed and actual scale scores, and percent of zero-scale score differences was computed as the percent of the differences between the imputed and actual scores that are equal to zero. For convenience of interpretation, the bias and RMSD were expressed in units of the standard deviation (*SD*) of the scale score, where the *SD* was computed from the complete data set. For convenience, these measures would be referred to as standardized bias and RMSD, respectively. Denoting T_{ji} and $\hat{T}_{ji}$ as the actual and imputed scale scores of examinee j (in the validation data) in iteration i, the standardized bias, and RMSD for a simulation condition can be computed as $\frac{1}{S}\frac{\sum_i\sum_j(T_{ji}-\hat{T}_{ji})}{1{,}000n_v}$ and $\frac{1}{S}\sqrt{\frac{\sum_i\sum_j(T_{ji}-\hat{T}_{ji})^2}{1{,}000n_v}}$, respectively, where n_v is the number of examinees in the validation data set and S is the *SD* of the scale score computed from the whole sample. The percent of zero-scale score differences is given by $100\frac{\sum_i\sum_j I_{T_{ji}=\hat{T}_{ji}}}{1{,}000n_v}$, where $I_{T_{ji}=\hat{T}_{ji}}$ is an indicator function denoting whether T_{ji} is equal to $\hat{T}_{ji}$. Accurate imputation corresponds to a standardized bias very close to zero, a small standardized RMSD, and a large percent of zero-scale score differences. The following types of graphical plots were examined: histograms of the actual and imputed scale scores from the imputation approaches, bivariate scatterplots of the actual and imputed scale scores from each imputation approach, bivariate scatterplots of the imputed scale scores from pairs of imputation approaches, and bivariate scatterplots of the differences and absolute differences between the actual and imputed scale scores from pairs of imputation approaches.

Results under the MCAR Missingness Mechanism

Figure 1 shows the absolute differences between the actual and imputed scale scores as a multiple of the *SD* for the GBM approach (along the *X*-axis) versus that for the PMI approach (along the *Y*-axis) for the case of 50% missing MC scores and 0% missing CR scores. Figure 1 was created after *jittering*, that is, adding a small random noise to the values of the absolute differences. The jittering, along with the fact that the scale scores and hence the absolute difference can only be integers, lead to what appear to be boxes in the bottom left corner of the figure. A diagonal solid line is provided for convenience—points falling close to this line are supposed to indicate little difference between the two imputation approaches. The horizontal and vertical dashed lines in Figure 1 represent the absolute difference of 1.0 times the *SD* are provided for convenience. The ranges of the *X* and *Y* axes are the same. Figure 1 shows that the differences for the PMI approach agree in general with those of the GBM approach, but also shows that the absolute difference for the PMI is larger than that for the GBM approach for a substantial number of examinees. For example, the number of absolute differences larger than the *SD* is nine for the GBM approach, but more than 40 for the PMI approach.

Under the MCAR missingness mechanism, the standardized bias did not exceed .02 in absolute value for any of the imputation approaches in any simulation condition with the exception of a value of .06 for PMI for three missing item scores for Test A—so all the imputation approaches essentially led to unbiased imputation in all the cases. The standardized bias was zero up to two decimal places for several approaches in several cases.

Rows 3–8 of Table 1 show the standardized RMSDs, for one, two, and three missing item scores for Tests A and B for the MCAR mechanism. To put the number of missing item scores into context, the maximum possible raw score on the nonmissing part, as a percentage of the maximum possible raw score on the total test, is provided in the first row of the table ("Raw Score %"); and the reliability of the score on the nonmissing part, as a percentage of the reliability of the total test, is provided in the second row of the table ("Reliability %"), where the coefficient α was used as a measure

[1]Changing this percentage to other values such as 10% or 20% did not affect the comparative performance of the imputation approaches.

[2]While 1,000 iterations are performed in this comparison, a practitioner would use only one iteration for the chosen imputation approach in practice and report imputed scores based on the results from that one iteration.

© 2020 by the National Council on Measurement in Education

17453992, 2021, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/emip.12396 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

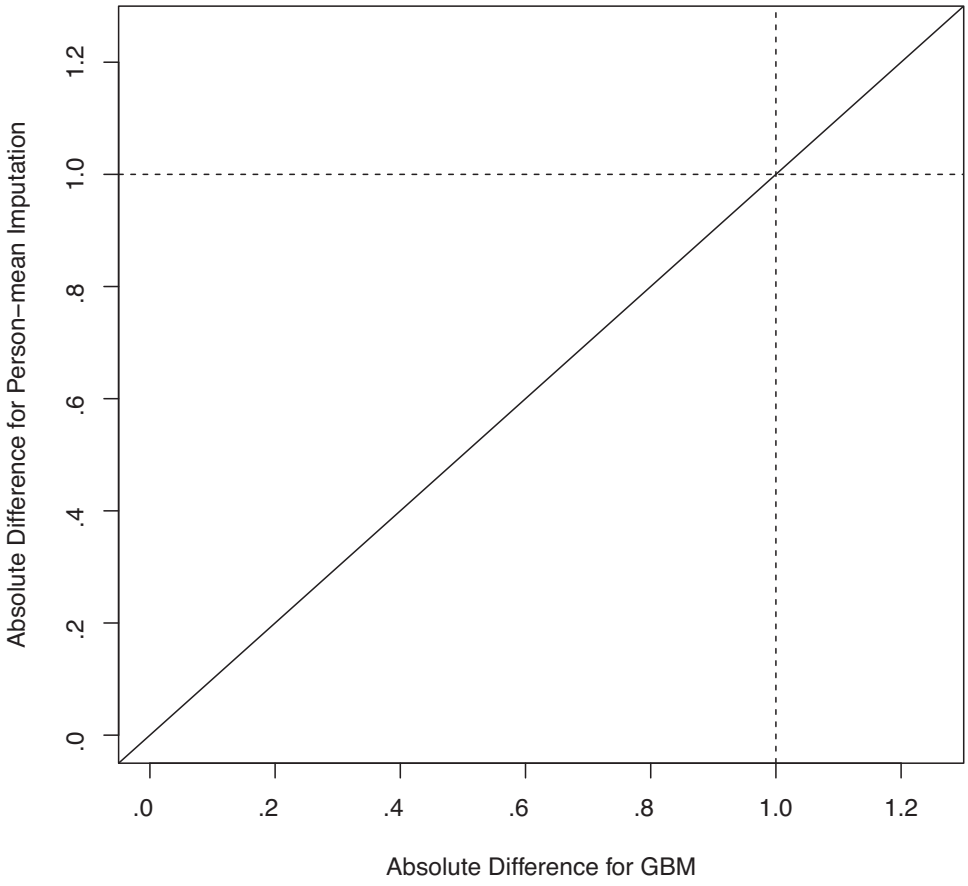

FIGURE 1. A scatterplot of the absolute differences between the actual and imputed scale scores (in *SD* units) for the GBM approach versus that for the PMI approach for the state test.

Table 1. The Standardized RMSDs and Percent of Zero-Scale Score Differences for the Different Imputation Approaches for the Two Speaking Tests When Scores Are MCAR

		Test A			Test B		
		1 miss.	2 miss.	3 miss.	1 miss.	2 miss.	3 miss.
Raw Score %		94	87	80	85	71	58
Reliability %		99	98	95	97	93	88
Standardized RMSD	GBM	.14	.20	.23	.19	.28	.36
	MICE	.15	.20	.25	.18	.29	.37
	Regn	.15	.21	.25	.19	.30	.38
	IRT	.15	.21	.26	.19	.31	.43
	Link	.17	.22	.26	.20	.32	.33
	PMI	.17	.24	.28	.20	.33	.46
Percent of zero scale score difference	GBM	88	78	67	66	47	35
	MICE	87	77	68	66	48	35
	Regn	87	76	66	65	47	34
	IRT	87	75	66	65	46	35
	Link	84	72	63	61	43	32
	PMI	83	70	55	63	44	31

Note. "miss." = missing; "Regn" = regression; "Link" = linking.

of reliability. The first two rows of numbers in Table 1 imply, for example, that the maximum possible raw score on Test A with one missing item score is about 94% of that on Test A as a whole, and the reliability of the score on Test A with one missing item score is about 99% of the reliability of the score on Test A as a whole. The *SD*s of the scale scores for the full data sets were about 25 and 4, respectively, for Tests A and B. The value of .15 for regression imputation for one missing item for Test A in Table 1 means that, on average, the absolute difference between the imputed and actual scale scores for the simulation case was $25 \times .15 \approx 3.75$. The last six rows of Table 1 show the percent of zero-scale score differences for one, two, and three missing item scores for Tests A and B for the MCAR mechanism.

Table 2 shows the maximum possible raw score on the non-missing part as a percentage of the maximum possible raw

17453992, 2021, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/emip.12396 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 2. The Standardized RMSDs and Percent of Zero Scale Score Differences for the Different Imputation Approaches for the State Test When Scores Are MCAR

		Percent of Items with Missing Scores				
		(0,30)	(0,50)	(20,0)	(20,30)	(50,0)
Raw Score %		87	80	87	74	70
Reliability %		98	96	98	96	96
Standardized RMSD	GBM	.11	.17	.13	.15	.22
	MICE	.12	.17	.13	.16	.23
	Regn	.12	.18	.13	.16	.23
	IRT	.12	.19	.13	.17	.24
	Link	.13	.20	.14	.17	.25
	PMI	.14	.21	.15	.19	.27
Percent of zero-scale score difference	GBM	91	80	89	82	72
	MICE	90	79	90	81	72
	Regn	89	79	87	82	71
	IRT	89	78	86	79	73
	Link	85	72	86	75	66
	PMI	84	72	85	74	65

score on the total test (first row of numbers); the reliability of the score on the nonmissing part as a percentage of the reliability of the total test (second row); the standardized RMSDs (rows 3–8); and the percent of zero-scale score differences (rows 9–14), for the State test for the MCAR mechanism. The five columns of numbers in the table correspond to the five combinations of percent of MC item scores missing and percent of CR item scores missing—these two percentages are shown in column (1), with a comma separating them. The missing data patterns are sorted according to the increasing order of the percent of missing MC item scores. The *SD* of the scale score for the complete data set from the state test was about 35.

The standardized RMSDs in Tables 1 and 2 can be considered as effect sizes roughly representing the average absolute errors in imputing the scale scores. Thus, standard guidelines on effect sizes (e.g., Cohen, 1988, p. 40) can be used to interpret the standardized RMSDs in Tables 1 and 2. According to Cohen's guidelines, effect sizes larger than .2, .5, and .8 can be considered small, medium, and large, respectively. Therefore, for example, for Test A, the effect size corresponding to the absolute error in imputing scale scores using all the approaches can be considered "negligible" for one missing item and "small" for two or three missing items.

Tables 1 and 2 show that when data are MCAR, the absolute difference between the actual and imputed scale scores increases on average as the number of missing item scores increases. For example, in Table 1, the standardized RMSDs for Test A for all the approaches are smaller than or equal to .17 for one missing item score, but are .23 or larger for three missing item scores. The tables also show that for any given test, the performances of the approaches do not differ much for any missing data pattern. This result could be an outcome of the fact that when scores are available on a large proportion of the test, all approaches perform very similarly to each other. The GBM and the MICE approaches seem to lead to the most accurate imputation overall, closely followed by the regression approach. Though the standardized RMSDs for the GBM and MICE approaches are only slightly smaller than that of the regression approach for any pattern, the difference is statistically significant (a difference of .01 or larger in standardized RMSD in Tables 1 and 2 is statistically significant given the large number of replications). A primary contributor to accurate imputation for the regression approach is the large multiple correlation coefficient of the total score on the nonmissing item scores—this multiple correlation coefficient is larger than .96 on average in all the regressions performed for Test A and larger than .86 on average in all the regressions performed for Test B. The PMI approach seems to be worse than all the other approaches, but does not perform much worse than the other approaches. This result implies that the operational imputation approach (which is similar to the PMI approach) for Tests A and B performs quite respectably compared with the other imputation approaches considered in this paper.

For the state test, when there are no technical difficulties, the missing item scores are treated as 0. Had the same approach been used even under technical difficulties, the RMSDs would have been much larger. For example, for the state test data available to us, treating scores missing due to technical difficulties as 0 would have led to the standardized RMSD of .48 (versus .12 for regression) for the case represented by the first row of Table 2 and of 1.13 (versus .23 for regression) for the case represented by the last row of Table 2, and a major contributor to this large standardized RMSD is a large bias or underestimation of the examinee score. The approach of treating missing item scores as 0 is unfair to the examinees because the approach penalizes the examinees for no fault of their own. Therefore, the standardized RMSDs for the approach and those in Table 2 show that all the imputation approaches considered in Table 2 perform much better than an unfair approach and hence can be argued to lead to more fair scores.

Results under the MAR Missingness Mechanism

The performance of any imputation approach under the MAR mechanism was very similar to that under the MCAR missingness mechanism. The standardized bias and RMSD for any imputation approach and any combination of missing data pattern and test were the same up to two decimal places over the MCAR and MAR missingness mechanisms. The rounded percent of zero-scale score differences is the same for the MAR and MCAR missingness mechanisms for any approach for any simulation case. The similarity of the results over the MCAR and MAR mechanisms is expected,

© 2020 by the National Council on Measurement in Education

17453992, 2021, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/emip.12396 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

for example, from the discussion and recommendations of Finch (2008) Graham (2009, p. 553), and Schafer and Graham (2002) who asserted that the common imputation approaches like the MI approach yield accurate imputations under the MAR mechanism.

Results under the MNAR Missingness Mechanism

The performance of any imputation approach under the MNAR mechanism was only slightly worse than that under the MAR or MCAR missingness mechanism. The difference in standardized RMSDs under the MNAR and MCAR mechanisms for any imputation approach and any combination of missing data pattern and test was .00 up to two decimal places for one or two missing item scores for Tests A and B and for the missing data patterns represented in the first three columns of Table 2 for the state test; the difference was .01 for three missing item scores for Tests A and B and for the missing data patterns represented in the last two rows of Table 2 for the state test. For example, for three missing item scores for Test A, the standardized RMSD for the regression approach was .25 and .26, respectively, under the MCAR and MNAR mechanisms. The slightly larger value for the MNAR mechanism for this case was the outcome of a small positive standardized bias of about .05 in all the imputation approaches. Thus, the imputed scores were slightly larger than the actual scores on average. A positive bias for this case is expected given that, to introduce MNAR data, smaller item scores were assigned larger probabilities to be missing. The rounded percent of zero-scale score differences is the same or 1 less for the MNAR missingness mechanism compared to the MCAR missingness mechanism for any imputation approach for any combination of missing data pattern and data set. Such closeness of the results under the MNAR and the MCAR/MAR mechanisms may seem counterintuitive because, for example, Finch (2008) noted that MNAR data can create great difficulties in accurately imputing missing values. However, the closeness of the results is most likely an outcome of the fact that, for example, when the score on Item 1 is MNAR for an examinee, the scores on the other items provide a substantial amount of information on the missing score because all the item scores are influenced by the examinee's ability parameter. In other words, even if the score on Item 1 is MNAR, the scores on Items 2, 3, . . ., of the examinee provide a substantial amount of information about the examinee ability that largely determines the score on Item 1, which, in turn, leads to accurate imputation of the score on Item 1. Researchers such as Graham (2009) and Sinharay et al. (2001) emphasized the importance of including covariates that predict the missing values in the imputation model and asserted that imputation will yield reasonable estimates even under the MNAR mechanism in the presence of strong covariates; the scores on Items 2, 3, . . ., of the examinee (which predict the score on Item 1 of the examinee quite well) act as strong covariates in this context. The accurate prediction under the MNAR mechanism is in agreement with the finding of (van Ginkel et al., 2010, p. 27) that their imputation approaches produced unbiased estimates of data summaries such as the reliability coefficient under the MNAR mechanism. van Ginkel et al. (2010) attributed their finding to the fact that even if an extreme MNAR missingness mechanism results in several missing item scores for an examinee, the observed item scores for the individual contain enough information to impute the missing scores reasonably well.

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 intends to choose the imputation approach that is expected to lead to the most accurate imputation and is willing to use an approach that is complex and computation-intensive, then the GBM approach or the MICE approach should be chosen as the method of choice. However, if a practitioner is interested in a simple approach (as asserted by Sijtsma & van der Ark, 2003), then the regression imputation approach may be chosen. Two approaches that are more computation-intensive than regression imputation—the IRT approach and the MIDA approach—provided no benefits over the regression approach. Note that if the RMSDs in Tables 1 and 2 or the losses in reliability (reported earlier) seem too large to the practitioner (for example, if the test scores are used to make high-stakes decisions, then an RMSD of .15 or a loss of reliability of 5% or more may be perceived as too large), then the practitioner may conclude that imputation should not be used and a test-taker with any missing item score should retake the test.

Conclusions and Recommendations

Six approaches for imputing missing scores were suggested or considered and their performances were compared using data from three educational tests. Of the imputation approaches, an approach based on data mining—the GBM approach—was the first of its kind to be applied to impute missing scores. The differences between the imputation approaches were small, which is in agreement with findings of small differences between various imputation approaches by, for example, Finch (2008), Huisman and Molenaar (2001), and Smits et al. (2002). Imputations based on GBM and the MICE approach were found to lead to the most accurate imputation of missing scores, with the regression imputation approach being a close second. All the imputation approaches performed rather well. The accurate prediction for the regression imputation approach should be good news to practitioners and operational testing programs given the simplicity and ubiquitous nature of linear regression.

This paper considers tests for which a scale score computed from a unweighted sum of the item scores is reported. All the approaches suggested in this paper, with the exception of the IRT-based approach, apply in a straightforward manner to tests that report a scale score computed from a weighted sum of the item scores. For example, consider a 10-item test whose first five items have possible scores of 0, 1, and 2 and the last five items have possible scores of 0 and 1 and Items 1–5 and 6–10 receive weights of 2 and 3, respectively, in a weighted sum score that is later converted to a scale score; one can apply the aforementioned imputation approaches to such a test after making the assumption that the possible scores on Items 1–5 are 0, 2, and 4, and those on Items 6–10 are 0 and 3 and a scale score computed from a unweighted sum of the item scores is reported for the test. This paper deals with imputation of scores for tests that are marred by technical difficulties. However, the imputation approaches can also be applied to report scores on tests for which item scores were missing due to some reasons other than technical difficulties, especially if the reason is not related to any examinee behavior. Examples of such reasons are

17453992, 2021, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/emip.12396 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

answer papers on a part of the test getting lost in transit and a part of a test getting canceled for some examinees due to a natural disaster. In addition, if some examinees experience technical difficulties during a test, their answers to all items are available, and their performances after the technical difficulties are found to have been adversely affected (from an analysis suggested by, for example, Sinharay & Jensen, 2019; Sinharay, Wan, Choi, & Kim, 2015), one can consider the possibility of treating the item scores after the technical difficulties as missing and imputing the scores of the corresponding examinees using the approaches considered in this paper.

Although the PMI or proration approach has been criticized for leading to biased estimates (e.g., van Ginkel et al., 2010), the approach, which is similar to the approach currently used by Tests A and B, did not perform much worse than the other approaches in this study. While one reason of this phenomenon is the consideration of only up to a few item scores as missing in the simulations, another reason is that the items chosen as missing were selected randomly in the simulations. If the missing item scores had corresponded to the hardest or easiest items, then the imputed scores would have been positively or negatively biased, respectively (this phenomenon was verified in some additional simulations whose results are not reported and can be obtained from the author).

All of the imputation approaches (considered in this paper) allow the investigator to obtain a "completed" or "filled-in" data set (e.g., Little & Rubin, 2002, p. 85) that includes the actual/observed item scores and imputed missing item scores of all examinees. For approaches such as the MI approaches, missing item scores can be obtained as the outputs of standard software packages. For some other approaches such as the IRT-based and the regression-based approaches, it is possible to obtain completed data sets after some simple calculations (like rounding of the imputed/estimated scores).

One important finding in this paper is the accurate imputation of most of the imputation approaches when the item scores were MNAR even though these approaches did not explicitly model the missingness. This finding is essentially an outcome of the high interitem correlations, agrees with a similar finding in van Ginkel et al. (2010), and implies that testing programs may not need to worry too much about MNAR data while imputing the scores, at least when technical difficulties lead to a small to moderate extent of missing item scores.

Any approach of score imputation involves uncertainty that is typically ignored while reporting an imputed score to an examinee. For example, when a score imputed by the regression-based approach is reported to an examinee, the uncertainty quantified in the variance of the score predicted by the regression (e.g., Draper & Smith, 1998, pp. 81–82) is ignored. The uncertainty may be reported to the examinees. Somewhat related is the issue of variance of the reported scores—it is possible to perform further research on the computation of variance of imputed scale scores and on examining how much the variance of an imputed scale score is inflated compared to a scale score that is not imputed. In addition, the score users should have an easy way to know which scores were imputed. The imputed examinee scores, which do not reflect the uncertainty associated with the imputation, should not be used to compute any population-level estimates (such as item-parameter estimates or equating functions).

One question that is closely related to the aforementioned issue of uncertainty is "What is the reliability of the imputed scale scores?" or "What values of reliability should be reported for tests on which the scores of some examinees are imputed? One value of reliability for those without any missing item scores and other values for each unique pattern of missing item scores?" One may be tempted to assume that the reliability of the imputed scale scores is identical to the reliability of the scale scores computed using the "completed" data sets. For each of the three aforementioned tests, the reliability of the scaled scores computed using a "completed" data set was identical, up to two decimal places, to that of the original data set for the regression, GBM, and MICE approaches (this result is in agreement with, for example, table 6 of Sijtsma & van der Ark, 2003). However, the reliability of the completed data sets, while providing some useful information, probably pertain more to the examinees with no missing data and less to the typically small percentage of examinees with missing data.[3] On the other hand, all the imputation approaches involve imputation of the raw score based on the nonmissing data so that one may be tempted to assume that the reliability of the imputed scale score is equal to the reliability of the raw score on the nonmissing part; however, this assumption is not appropriate either because the test administrators are reporting an imputed scale score and not the raw score on the nonmissing data (this is true in spite of the fact that the reliability of the raw score and scaled score for the original data set was within .01 of each other for all the data sets in this paper). The reliability of imputed scores is a potential topic for future research.

In some cases, the true ability may not be reflected by the available item scores of an examinee who experienced technical difficulties. The phenomenon may occur if, for example, an examinee experiences a technical difficulty after the fifth item of the test, becomes nervous, completes the test, and then responses to items 30-40 become missing for the examinee. Therefore, one should impute scores for an examinee only if one is confident that the observed scores are not adversely impacted by the technical difficulties. Also, if there are some missing responses for an examinee who experienced technical difficulties on a test, it is often difficult or impossible to figure out which of those missing responses are missing due to technical difficulties and which, for example, the examinee was going to omit even without technical difficulties. While a conservative approach comprises the treatment of all the missing responses as missing due to technical difficulties, it would be ideal to find a way (possibly by better record-keeping on the actions of the examinees during the test) to find out exactly which missing responses are missing due to technical difficulties.

This paper is concerned with the imputation of the examinee scores themselves rather than estimation of the corresponding true scores. That is because an estimated true score (or an equated version of it) is not reported for the examinees with complete data—so reporting of an estimated true score for those with missing data may lead to unfair

[3]Consider a large item-response data set that includes no missing data. If one marks as missing all the item scores of one examinee in the data set and imputes the item scores by completely random numbers, the reliability of the completed data set will be identical up to two decimal places to that of the original data set. This closeness does not mean that the imputed score is reliable for that examinee.

© 2020 by the National Council on Measurement in Education

17453992, 2021, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/emip.12396 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

scores (for example, an estimated true score will be pulled toward the mean and may place the high-scoring examinees at a disadvantage). In addition, though terms such as "imputation of missing score" and "imputed scores" are used in this paper, the imputation of the scores using some of the imputation approaches (like the regression or IRT-based approaches) does not involve any random draws and is conceptually similar to the prediction of an individual observation in linear regression (e.g., Draper & Smith, 1998, p. 81).

Although the findings of this paper may have important practical implications, this paper has several limitations, and consequently, additional research is needed in several related areas. First, this paper considered only one, two, or three missing item scores for each examinee, the items with missing scores were randomly chosen and the same one, two, or three items were assumed to have missing scores for 30% examinees in each data set. Though the comparative performance of the imputation approaches was the same in all simulation cases considered in this paper, it is possible that the comparative performance would have been different for some cases (including the case of different examinees having missing scores on different sets of items or most difficult items having missing scores) not considered in this paper. Second, one could compare the imputation approaches using more data sets, both simulated and real, preferably from other types of tests. Specifically, the imputation approaches, which are designed primarily for MAR data, may lead to inaccurate imputation for some types of MNAR data—research on finding these types of data/situations may be useful. Third, it is possible to employ, in future comparison studies, more advanced approaches such as the model-based methods that apply when data are MNAR (e.g., Enders, 2011; Glas & Pimentel, 2008; Holman & Glas, 2005; Rose et al., 2017; Sulis & Porcu, 2017) and more data mining methods (e.g., Strobl, 2013). However, as noted by Sijtsma and van der Ark (2003), these approaches, especially those designed for the MNAR data, may not be easy or practical to use for large-scale tests due to their complexity. Fourth, the operational rules for handling missing item scores for the tests were not questioned in this paper, but future research could examine whether such rules are too strict or too liberal. Finally, it is possible to examine if the relative performance of the imputation approaches are similar over relevant demographic subgroups.

Acknowledgments

The author wishes to express sincere appreciation and gratitude to Deborah Harris, the editor, and the three anonymous reviewers for their helpful comments. The author would also like to thank Marna Golub-Smith, Hongwen Guo, and J. R. Lockwood for their helpful comments on an earlier version. Any opinions expressed in this publication are those of the author and not necessarily of ETS. The opinions expressed are solely those of the authors.

References

Allen, N. A., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 technical report* (NCES No. 2001-452). Washington DC: United States Department of Education, Institute of Education Sciences, Department of Education, Office for Educational Research and Improvement.

Baker, F. B., & Kim, H. S. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Marcel Dekker.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*, 199–231.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

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. Retrieved from 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.

Chen, J., Fife, J. H., Bejar, I. I., & Rupp, A. A. (2016). *Building e-rater® scoring models using machine learning methods* (ETS Research Report No. RR-16-04). Princeton, NJ: ETS.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

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.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York, NY: John Wiley.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*, 802–813.

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods*, *16*, 1–16.

Fernandez-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, *15*, 3133–3181.

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, *45*, 225–245.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *28*, 1189–1232.

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*. New York, NY: Springer.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: 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*. Leiden, The Netherlands: 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). New York, NY: Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. New York, NY: Springer.

Kohler, 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.). New York, NY: Springer.

 © 2020 by the National Council on Measurement in Education

17453992, 2021, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/emip.12396 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of American Statistical Association*, *83*, 1198–1202.
Little, R. A., & Rubin, D. (2002). *Statistical analyses with missing data* (2nd ed.). New York, NY: John Wiley & Sons.
Michel, R. S. (2020). Remotely proctored K-12 high stakes standardized testing during COVID-19: Will it last? *Educational Measurement: Issues and Practice*, *39*(3), 28–30.
Milborrow, S. (2016). *rpart.plot: Plot 'rpart' models: An enhanced version of 'plot.rpart'*. (R package version 2.1.0)
Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.
Potthoff, R. F., Tudor, G. E., Pieper, K. S., & Hasselblad, V. (2006). Can one assess whether missing data are missing at random in medical studies? *Statistical Methods in Medical Research*, *15*, 213–234.
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.
Ramanarayanan, V., Lange, P. L., Evanini, K., Molloy, H. R., & Suendermann-Oeft, D. (2017). Human and automated scoring of fluency, pronunciation and intonation during human-machine spoken dialog interactions. In *Interspeech 2017*. ISCA.
Ridgeway, G. (2017). *gbm: Generalized boosted regression modeling*. (R package version 2.1.3)
Rose, N., von Davier, M., & Nagengast, B. (2017). Modeling omitted and not-reached items in IRT models. *Psychometrika*, *82*, 795–819.
Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: John Wiley & Sons, Inc.
Schafer, J. (1997). *Analysis of incomplete multivariate data*. New York, NY: 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.
Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*, 289–310.
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. (2016). An NCME instructional module on data mining methods for classification and regression. *Educational Measurement: Issues & Practice*, *35*, 38–54.
Sinharay, S., & Jensen, J. L. (2019). Higher-order asymptotics and its application to testing the equality of the examinee ability over two sets of items. *Psychometrika*, *84*, 484–510.
Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods*, *6*, 317–329.
Sinharay, S., Wan, P., Choi, S. W., & Kim, D. (2015). Assessing individual-level impact of interruptions during online testing. *Journal of Educational Measurement*, *52*, 80–105.
Sinharay, S., Wan, P., Whitaker, M., Kim, D., Zhang, L., & Choi, S. W. (2014). Determining the overall impact of interruptions during online testing. *Journal of Educational Measurement*, *51*, 419-440.
Sinharay, S., Zhang, M., & Deane, P. (2019). Prediction of essay scores from writing process and product features using data mining methods. *Applied Measurement in Education*, *32*, 116–137.
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.
Strobl, C. (2013). Data mining. In T. Little (Ed.), *The Oxford handbook of quantitative methods in psychology* (Vol. 2, pp. 678–700). New York, NY: Oxford University Press.
Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*, 323–348.
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.
Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*, 1049–1064.
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.
Vriens, M., & Sinharay, S. (2006). Dealing with missing data in surveys and databases. In R. Grover & M. Vriens (Eds.), *The handbook of marketing research* (pp. 178–191). New York, NY: Sage.
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.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:
https://onlinelibrary.wiley.com/doi/10.1111/emip.12396

© 2020 by the National Council on Measurement in Education

17453992, 2021, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/emip.12396 by Arizona State University Acq & Analysis, Lib Continuations, Wiley Online Library on [20/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License