ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(3); 209-220

The Efficacy of the IRTree Framework for Detecting Missing Data Mechanisms in Educational Assessments

Yeşim Beril SOĞUKSU*

Abstract

The effectiveness of methods for handling missing data in educational assessments depends on understanding the underlying missing mechanisms. This study investigates the performance of the IRTree framework in detecting missing data mechanisms using a Monte Carlo simulation. Omitted responses were simulated at varying proportions according to three mechanisms: MCAR, MAR, and MNAR, across tests with different lengths and sample sizes. The IRTree was employed to model the omitted responses and detect the mechanisms based on the correlations between the propensity to omit and proficiency. Results indicate that the IRTree accurately identifies all three missing data mechanisms, with no relationship between propensity to omit and proficiency under MCAR, and negative correlations for MAR, reaching up to -0.3, and for MNAR, as high as -0.8. Furthermore, the detection of MAR and MNAR mechanisms became more pronounced with higher proportions of omitted responses, longer tests, and larger sample sizes. IRTree framework not only enables educators and researchers to accurately understand the nature of missing data but also guides them in using appropriate methods for handling it.

Keywords: IRTree, missing data, missing data mechanism, simulation, R language

Introduction

The issue of missing data, which emerges in measurements conducted across various fields such as educational sciences, psychology, healthcare, and social sciences, presents a significant challenge for researchers. Particularly in the fields of education and psychology, where critical decisions about individuals are made, it is essential to understand the nature of missing data and apply appropriate handling methods to ensure that estimates are unbiased and accurate. Missing responses can occur for various reasons across a wide range of testing situations, from classroom to large-scale educational assessments. For instance, in classroom achievement tests, students may omit questions even if they have enough time to answer them due to reasons such as not knowing the answer, difficulty in choosing between options, fatigue, motivation decline or stress. In speed tests, where students must answer as many questions as possible within a given time, it is common for students to leave some items unanswered, particularly towards the end of the test (De Ayala et al., 2001; Graham, 2012; Little & Rubin, 1987). Additionally, in large-scale educational assessments, missing data may result from administering a subset of items to participants to reduce their response burden. This type of planned missing data, known as nonadministered data, typically does not threaten the validity of the assessment (Rose et al., 2015). However, missing responses due to omitted and not-reached items remain one of the most common and problematic issues researchers face during data analysis.

When missing data is not addressed properly, it can lead to biased parameter estimates, Type I and Type II errors, and reduced statistical power. For example, in high-stakes tests where critical decisions are made about students, if ability scores are estimated higher (positive bias) or lower (negative bias) than their actual scores, this bias can result in incorrect decisions about their academic placement or progression. The importance of obtaining unbiased estimates becomes particularly critical in tests where such decisions are based on ability scores. Additionally, errors in hypothesis testing, such as incorrectly

* Dr., Ministry of National Education, Vali Hilmi Tolun Middle School, Kahramanmaraş-Türkiye, e-mail: berilsoguksu@gmail.com, ORCID ID: 0009-0004-0870-4974

To cite this article:
Soğuksu, Y.-B. (2024). The efficacy of the IRTree framework for detecting missing data mechanisms in educational assessments. *Journal of Measurement and Evaluation in Education and Psychology, 15*(3), 209-220. https://doi.org/10.21031/epod.1514741

Received: 11.07.2024
Accepted: 16.10.2024

rejecting a true null hypothesis (Type I error) or failing to reject the null hypothesis (Type II error), further complicate research findings and diminish the validity of studies (Little et al., 2016; Newman, 2014; Roth, 1994). The importance of unbiased and precise estimates becomes even more evident when considering that countries shape their educational policies based on the results of international assessments like Program for International Student Assessment - PISA (Damiani, 2016; Martens et al., 2016).

The extent to which missing data can affect the validity of measurements depends on their proportions, patterns, and mechanisms (Tabachnick & Fidell, 2007). The proportion of missing responses can significantly impact the generalizability of the study and the statistical inferences drawn from the data. Missing data patterns indicate which responses are observed and which are missing in the dataset but do not provide information about the reasons behind the missing data. Similarly, missing data mechanisms describe the statistical relationships between missing and observed data without explaining how or why the data are missing (Enders, 2010; McKnight et al., 2007).

Rubin (1976) proposed three different mechanisms for missing data: Missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). In the MCAR mechanism, the presence of missing data is not associated with any variables in the dataset, including the variable itself (Allison, 2002). For instance, in an attitude scale, if students randomly forget to answer some items regardless of their attitudes, this exemplifies the MCAR mechanism. In contrast, the MAR mechanism occurs when the missingness of a variable is related to other observed variables but not to the variable itself (McKnight et al., 2007). For example, in a survey on online learning platforms, students with limited access to technology may have more missing responses. When controlling for other variables, these missing responses are associated with the students' access to technology, not their attitudes towards online learning. The MNAR mechanism occurs when the probability of missing data is directly related to the value of the variable itself (Enders, 2010). An example of this is in competence tests, where lower-performing students tend to leave more questions unanswered, indicating that the missing responses are directly related to their ability levels.

MAR indicates the presence of ignorability, meaning there is no need to model missing data for accurate parameter estimation because the missingness is irrelevant to the parameters being estimated. Many traditional and modern missing methods assume that the missing data are ignorable. When this assumption is not met, it results in biased parameter estimations. Conversely, MNAR data are not ignorable and must be modeled accordingly. Even many modern methods commonly used in the literature can produce biased parameter estimates under the MNAR mechanism (Cheema, 2014; Enders, 2010; Graham, 2012; Rose et al., 2015).

In this context, detecting the missing data mechanisms in a dataset is crucial for researchers to determine the most appropriate methods for handling the missing data. Missing data mechanisms guide the selection of methods that will yield the best performance (Peugh & Enders, 2004). Pigott (2010) emphasized that different methods for handling missing data come with varying assumptions, and if these assumptions are not met, the results can be biased and misleading. For instance, traditional methods often used by researchers assume that missing data meet the MCAR assumption. If this assumption is violated, parameter estimates may be biased, and Type I and Type II errors may occur. In contrast, modern and more robust approaches such as Maximum Likelihood (ML) and Multiple Imputation (MI) operate under the assumption that missing data are ignorable, thereby allowing for unbiased parameter estimates (Graham, 2012; Little et al., 2016; Peugh & Enders, 2004). Therefore, it is essential for researchers to identify the missing data mechanisms in their datasets to apply the most suitable handling techniques.

Several methods can be used to determine whether the missing data mechanism is MCAR. For example, in a scenario where students are surveyed about their math attitudes at the beginning of the semester and their final math exam scores are collected at the end, if students who did not report their attitudes are expected to be no different, on average, from those who did, a t-test can be performed. By comparing the average final math scores between students with missing and complete survey responses, researchers can test for differences. If the missing data mechanism is MCAR, the average math scores should be the same within the sampling error. However, using this method for multiple variables with missing data

can increase the risk of Type I error. Little's MCAR test (Little, 1988) is more effective in such cases, as it helps avoid Type I errors (Enders, 2010).

Little's MCAR test compares the observed variable means for each missing data pattern with the expected population means and calculates a total weighted squared deviation. If the dataset meets the MCAR assumption, any subsample with a given missing data pattern should produce the same means for each variable as those calculated for the entire dataset using a robust parameter estimation method. When there are many patterns of missing data and each pattern tends to produce different means for each variable, the data are unlikely to be MCAR. In this case, the data deviate from a completely random process, and the chi-square test will be significant (McKnight et al., 2007).

Little (1988) contends that MCAR is the only missing data mechanism that can be empirically tested. Similarly, Enders (2010) argues that while methods exist to detect MCAR, it is not feasible to reliably distinguish between MAR and MNAR. Consequently, if the dataset does not meet the MCAR assumption, it is assumed to be either MAR or MNAR. This uncertainty poses a problem because using methods suitable for ignorable missing data on nonignorable missing data can lead to biased parameter estimates. Allison (2002) emphasized that if the MNAR mechanism is incorrectly assumed to be MCAR or MAR, the missing data process will not be modeled accurately, resulting in inaccurate parameter estimates. Similarly, if the MAR mechanism is mistakenly assumed to be MCAR, the estimated parameters will not be generalizable to the population. Huismann (2000) stated that the success of methods used to address missing data depends on accurately identifying the mechanism causing the missing data. In this context, the Item Response Tree (IRTree) framework is recommended as a model-based approach for dealing with missing data, as it provides insights into the missing data mechanisms present in the dataset (De Boeck & Partchev, 2012; Debeer et al., 2017; Jeon & De Boeck, 2016).

IRTree and missing data

The Item Response Tree (IRTree) framework, which integrates item response theory (IRT) and cognitive psychology theories, has garnered significant attention from researchers in recent years (Böckenholt, 2012; Jin et al., 2022). This framework explains response processes through tree-based model structures. There are various studies on the IRTree framework that focus on response styles (Alagöz & Meiser, 2023; Alarcon et al., 2023; Böckenholt, 2017; Dibek, 2019; Plieninger, 2021; Quirk & Kern, 2023; Spratto et al., 2021). Additionally, some studies use the IRTree framework to model answer change behavior (Jeon et al., 2017) and address missing data (Debeer et al., 2017; Huang, 2020; Jeon & De Boeck, 2016). By formulating response probabilities based on tree-based structures, the IRTree method provides a comprehensive framework for understanding response styles while also serving as a robust tool for handling missing data, thereby enhancing the accuracy and validity of parameter estimates in educational and psychological assessments. The ability of the IRTree framework to model non-ignorable missing data, which distinguishes it from both traditional and modern missing data methods, is one of its notable strengths. Unlike commonly used approaches such as Listwise Deletion, which result in a reduction of the dataset, IRTree preserves the integrity of the data. Furthermore, its applicability to both dichotomous and polytomous data, as well as its capability to provide insights into the underlying mechanisms of missing data, positions IRTree as a robust tool for handling missing data in research (Debeer et al., 2017).

The probability of selecting a particular response category in IRTree models depends on the path an individual takes through the tree model. The tree model shown in Figure 1, adapted from Debeer et al. (2017), can be used to model omitted items in a dichotomously scored test. The branching points of this tree model are called ‘nodes,’ and each node represents a different feature based on the underlying assumption of the model.

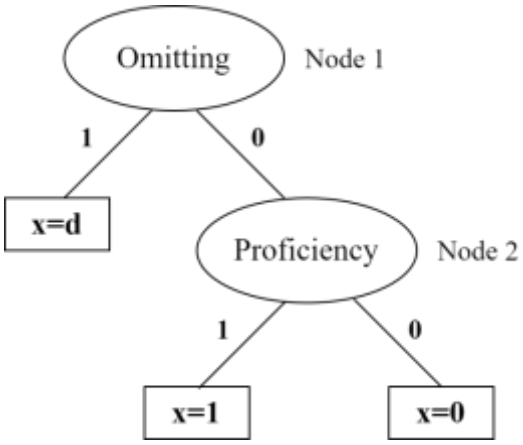

Figure 1. IRTree model for omitted items in dichotomously scored tests

In the tree model shown in Figure 1, when a test taker encounters an item, they first choose one of the behaviors, answer or omit, and then attempts to answer the item correctly. This behavioral process for the test taker is hypothetical. Different tree models can be constructed based on different assumptions. However, it is important that the tree model to be created can logically explain the test-taker's behavioral processes and is appropriate for the test situations. In Figure 1, the tree model has 3 different response categories. These are: Omitting (x=d), correct answer (x=1), incorrect answer (x=0). In addition, there are 2 nodes in this tree model: omitting process and proficiency. The first node, the omitting process, is connected to the proficiency node with two possible response outcomes (two branches from the node), and these nodes can be modeled with different IRT models. Since Birnbaum's (1968) Two Parameter Logistic IRT (2PL) model is used in this study, latent traits for the nodes are $\theta_j^{(1)}$and $\theta_j^{(2)}$ for the person j, slope parameters for item i (i= 1..... k) are $\alpha_i^{(1)}$and $\alpha_i^{(2)}$, and item difficulty parameters for item i are $\beta_i^{(1)}$ and $\beta_i^{(2)}$. This multidimensional model generates the probability of the observed response categories (X_{ij}=0, 1 or d) based on the combination of the probabilities of the sub-processes (Debeer et al., 2017; Jeon & De Boeck, 2016; Park & Wu, 2019).

Since the tree model 1 is a binary model, the right and left branches are coded as 0 and 1, and the observed response categories can be mapped with this coding for the branches. Table 1 shows the mapping matrix for the tree model in Figure 1. If the test taker omits the item in the first stage, the response output for node 1 is 1 and the response output for node 2 is NA; if the test taker answers the item correctly, the response output for node 1 is 0 and is 1 for node 2; if the test taker answers the item incorrectly, the response output for node 1 is 0 and is 0 for node 2. In this way, the data matrix consisting of the test takers' item responses is transformed into a large data matrix containing the responses for nodes 1 and 2. In formulating the probabilities of the response outputs, the response output for each node is modeled with IRT models and the product of these probabilities is taken.

Table 1.
The mapping matrix for omitted items

Original Responses	Node 1	Node 2
Omitted (x=d)	1	NA
Correct (x=1)	0	1
Incorrect (x=0)	0	0

When modeling omitted items in a test, different tree models can be constructed based on different assumptions. At this point, Akaike's Information Criterion (AIC) or Bayesian Information Criterion (BIC) can be used to determine which of the different tree models with the same response categories best explains the data structure, thus determining which tree model better fits the data. However, when the tree models have different response categories, it is difficult to compare them directly because the two tree models model different situations. Likelihood ratio tests (LRT) can be used to compare nested tree models (De Boeck & Partchev, 2012; Jeon & De Boeck, 2016).

As a result, after the nodes of tree model in Figure 1 are modeled with the IRT model, the propensity to omit is treated as a latent variable that is different from the measured latent variable of test takers, and the relationship between these two latent variables can be determined (Glas & Pimentel, 2008; Holman & Glas, 2005). At this point, Debeer et al. (2017) state that based on this relationship, it is possible to detect the existing missing data mechanism in a dataset. For example, before choosing the missing data method to be used in a test, determining the extent to which the propensity to omit is related to proficiency using IRTree can guide researchers in terms of using the appropriate handling method. For example, in a competence test if there is a strong negative relationship between these two latent variables (i.e., students with higher abilities omit less items), using traditional missing data methods to deal with missing data will lead to biased results. In this case, the researcher must choose methods that can be used to deal with nonignorable missing data.

When identifying the missing data mechanism in the dataset using the IRTree framework, the correlation between the latent variables for missing propensity and proficiency is examined. While the absence or low correlation between these two latent variables indicates the assumption of MCAR or MAR, a high relationship indicates the presence of nonignorable missing data. In the literature, there are various simulation studies in which different missing data mechanisms are created by manipulating the relationship between these two latent variables. For example, Debeer et al. (2017) generated missing data for MAR based on the absence of a relationship between these two latent variables, and for MNAR based on the relationship being -0.5. Similarly, Huang (2020) generated missing data for MNAR by determining the correlation between these latent variables as -0.5. Holman and Glas (2005) stated that as the correlation between two latent variables increases, the assumption of ignorability weakens. In their study, they showed that when the correlations exceeds 0.4, the nonignorability becomes evident. Köhler et al. (2017) varied the correlation between latent variables as 0.0, 0.2, 0.4, and 0.6, and treated the correlation of 0.0 as MCAR. Glas and Pimentel (2008) varied the correlations between latent variables as 0.0, 0.2, 0.4, 0.6 and 0.8, and stated that the violation of ignorability increases as the correlations move away from 0.0. Glas et al. (2015) interpreted a correlation of 0.0 between latent variables as ignorability, 0.4 as a slight violation of ignorability, and 0.8 as a serious violation of ignorability.

Consequently, the studies in the literature are simulative, and missing data for the MCAR, MAR, and MNAR mechanisms were generated by manipulating the correlations between the missing propensity and proficiency. At this point, a gap exists in the literature, as no study has employed the IRTree framework to identify missing data mechanisms under varying test conditions. Therefore, this study aims to demonstrate the efficacy of IRTree in detecting missing data mechanisms in the presence of omitted items in dichotomously scored tests by examining the correlations between the propensity to omit and proficiency. Specifically, the study explores how the IRTree framework can distinguish between MCAR, MAR, and MNAR mechanisms under varying testing conditions, such as different test lengths, sample sizes, and proportions of omitted responses. After modeling the missing datasets with IRTree, the following research questions were addressed:

- What are the mean correlations between the propensity to omit and proficiency for the missing datasets under MCAR, MAR and MNAR mechanisms?

- Do these mean correlations accurately reflect the underlying missing data mechanisms?

To provide an overview of this paper's structure, the subsequent sections are organized as follows: The Methods section outlines the methodology, including the Monte Carlo simulation used for data generation and the modeling of missing data with IRTree. The Results section provides the results of

the analyses, focusing on how well the IRTree framework was able to detect the missing data mechanisms under different test conditions. Finally, the Discussion section discusses the implications of these findings in the context of existing research and highlights the contributions of this study.

Methods

Data Generation

The study was conducted as a Monte Carlo simulation, using the mirt package (Chalmers, 2012) in R to generate the datasets. Data generation was performed according to Birnbaum's (1968) Two-Parameter Logistic (2PL) model. The formula for the 2PL model is as follows:

$$P(\theta_j) = \frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}} \quad (1)$$

In the Formula 1, θ_j refers to the latent trait of the person j, a_i refers to the item discrimination parameter, b_i refers to the item difficulty parameter for item i. While selecting the distributions of item and ability parameters, the previous studies were considered (Baker, 2001; DeMars, 2010; Feinberg & Rubright, 2016; Hambleton et al., 1991; Harwell et al., 1996). In this context, the discrimination parameters follow a uniform distribution a ~ U(0, 2). Item difficulty parameters and ability parameters were drawn from the normal distribution θ ~ N(0, 1). To avoid convergence issues and ensure accurate parameter estimation, test lengths were set at 20 and 40 items (Alarcon et al., 2023; DeMars, 2010). As in previous studies (Alarcon et al., 2023; Leventhal, 2019), a sample size of 500 was included in this study to assess the robustness of the IRTree framework with small sample sizes, which is commonly encountered in educational assessments. In addition, sample sizes of 1000 and 2000 were chosen based on previous studies on modeling missing data with IRTree (Debeer et al., 2017; Huang, 2020). Since it is necessary to have at least 5% omitted responses to effectively model the omitting process with IRTree (Debeer et al., 2017), the lowest missing data proportion in the study was set at 5%. Proportions of 10%, 30%, and 50% were also chosen to create test scenarios with varying levels of missing data and assess the robustness of the IRTree. The number of replications was set at 100. The study was conducted using the R programming language, version 4.3.2.

Types of missing data

After data generation, missing data were created for three different missing data mechanisms. For MCAR, %5, 10%, 30%, and 50% of the responses were randomly deleted from the datasets. For MAR, following the approach of Collins, Schafer, and Kam (2001), a covariate variable was generated that correlated with the total test scores, with the correlation set at 0.3. The covariate variable was divided into three group based on its quartiles (.00 - .33, .33 - .66, .66 - 1.00), and different probabilities of missing data were assigned to each group. At this point, test takers in the lower quartile group have a higher probability of missing data, whereas test takers in the upper quartile group have a lower probability of missing data. For example, in the 10% missing data scenario, 15% of observations were randomly deleted from group 1, 10% from group 2, and 5% from group 3, ensuring that the mean missing data proportion was 10%. For MNAR, as used in previous studies (Rose et al., 2010; Sulis & Porcu, 2017), the ability scores θ were considered. The ability scores were estimated and each sample sizes are divided into three groups based on the ability scores' quartiles (.00 - .33, .33 - .66, .66 - 1.00). Missing data were generated such that test takers with higher ability levels had a lower probability of missing data, while those with lower ability levels had a higher probability. For instance, in the 10% missing data scenario, 15% of observations were randomly deleted from the group with the lowest ability level, 10% from the middle group, and 5% from the highest ability group, ensuring a mean data proportion of 10%.

Analysis

Missing data sets were modeled using the tree model for omitted items in Figure 1 proposed by Debeer et al. (2017). This model includes two nodes: propensity to omit and proficiency. The original responses in the datasets were converted into mapping matrices based on this tree model using the dendrify2 function from the flirt package (Jeon, Rijmen, & Rabe-Hesketh, 2014). Subsequently, each node in the tree model was modeled with the 2PL model using the mirt function (Chalmers, 2012) in the wide data matrix format. Latent variables for the nodes were estimated using the Expected a Posteriori (EAP) method. The EAP method, proposed by Bock and Aitkin (1981), calculates the mean of the ability parameter distribution given the observed response pattern. Unlike Maximum Likelihood (ML) estimations, EAP estimations can be computed even when a test taker answers all items correctly or incorrectly (Bock & Mislevy, 1982; De Ayala et al., 2001). Since the missing data mechanisms were determined based on the correlation between the propensity to omit and proficiency, the mean Pearson correlation coefficient was calculated for all conditions in the study. If the mean correlation between these two latent variables is greater than 0.4, it is considered a violation of ignorability as in the literature (Debeer et al., 2017; Glas et al., 2015; Holman & Glas, 2005; Huang, 2020).

Results

In this study, the effectiveness of IRTree in detecting missing data mechanisms was tested under different simulation conditions. The correlations between the propensity to omit and proficiency were calculated, and the mean of these correlations was then computed. Figure 2 illustrates the mean correlations for MCAR mechanism across the various conditions.

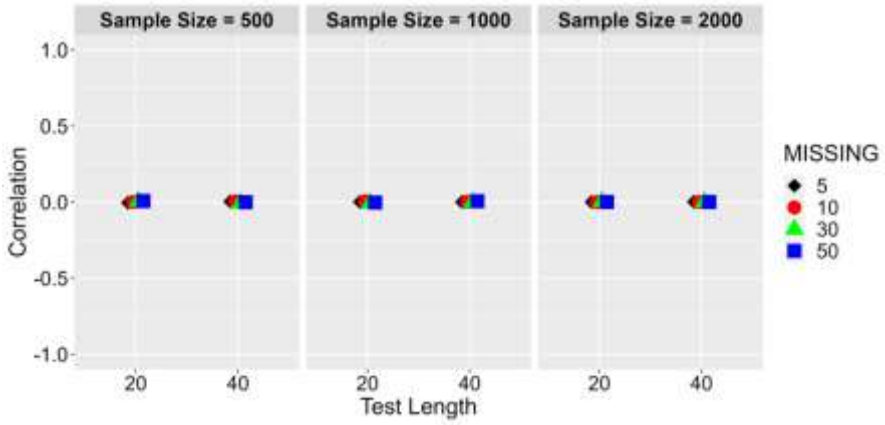

Figure 2. Mean correlations between propensity to omit and proficiency for MCAR

As shown in Figure 2, for MCAR, the mean correlations between the propensity to omit and proficiency are 0.0 for each condition. Mean correlations remained consistent across varying test lengths, sample sizes, and missing proportions. This indicates that under all conditions, there is no relationship between test takers' proficiency and propensity to omit, confirming that the missing data do not contain information about the measured latent trait and that the missingness is unrelated to both observed and unobserved variables.

The missing data for MCAR were generated by randomly deleting specified proportions of responses from the datasets. As a result, no relationship was expected between test takers' proficiency and their propensity to omit. Based on the results obtained through the IRTree framework, it can be concluded that the MCAR mechanism was correctly detected. Figure 3 illustrates the mean correlations across the conditions for MAR.

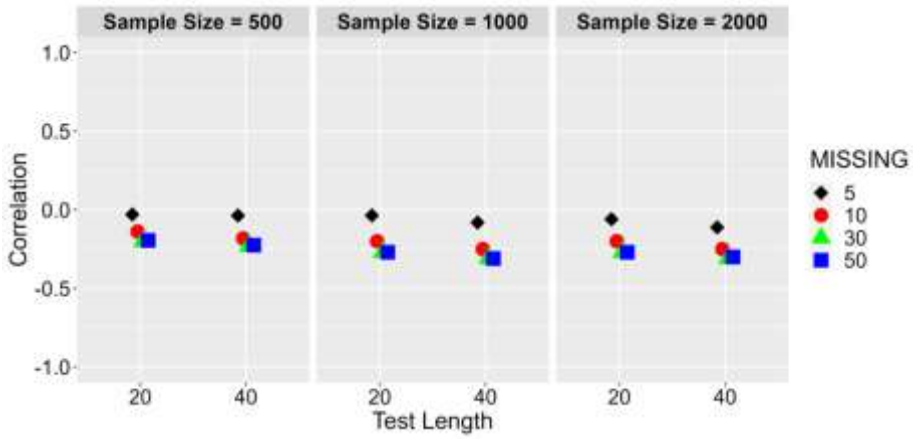

Figure 3. Mean correlations between propensity to omit and proficiency for MAR

In Figure 3, the mean correlations between the propensity to omit and proficiency are negative and up to -0.3. The negative mean correlations suggest that as test takers' proficiency increases, their propensity to omit decreases. However, the weakness of the relationship indicates that there is no violation of ignorability and that the missingness is not strongly related to proficiency.

Moreover, it was observed that an increase in the missing data proportion strengthens the relationship between the propensity to omit and proficiency. At this point, with a 5% missing data proportion, the missing data mechanism appears to align with MCAR, particularly in cases with smaller sample sizes and shorter tests. Therefore, for the relationship between propensity to omit and proficiency to adequately reflect the MAR assumption, and for the IRTree to effectively capture this relationship, the dataset should ideally contain at least 10% omitted responses. The increase in test length led to a slight rise in mean correlations, while increasing the sample size from 500 to 1000 resulted in higher mean correlations. However, increasing the sample size from 1000 to 2000 did not have an impact on the correlations, particularly when the missing data proportion exceeded 5%.

To generate missing data for MAR, a covariate variable with a low correlation to the total scores was created, and missing data were generated in increasing proportions based on the quartiles of this covariate variable. In this scenario, the probability of missing data does not depend on the test takers' proficiency, but rather on another (covariate) variable in the data set. Consequently, in the IRTree modeling, the low mean correlations between propensity to omit and proficiency successfully reflect this situation. Figure 4 illustrates the mean correlations across the conditions for MNAR.

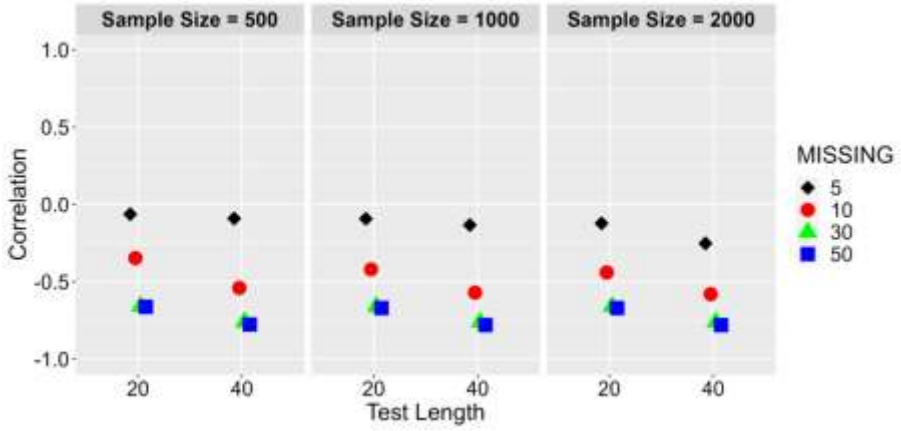

Figure 4. Mean correlations between propensity to omit and proficiency for MNAR

As shown in Figure 4, the mean correlations are negative and up to -0.8. However, when the missing data proportion is 5%, the relationship between propensity to omit and proficiency approaches 0.0, indicating either an MCAR or MAR mechanism. Nevertheless, as the missing data proportion increases, the mean correlations rise from moderate to high, indicating that the propensity to omit decreases as test takers' proficiency increases. This is consistent with the expectation that test takers with higher ability levels would omit fewer items, reflecting the MNAR data generation process. Therefore, the missing data contain information about proficiency and cannot be ignored.

The increase in the missing data proportion leads to a negative increase in the mean correlations, strengthening the MNAR mechanism. At this point, for the MNAR mechanism to be effectively detected using the IRTree, it is recommended that the dataset contain at least 10% omitted responses. Additionally, the mean correlations increased negatively with the increase in test length. While the increase in sample size resulted in higher mean correlations at 5% and 10% missing data proportions, changes in sample size did not cause a noticeable difference in mean correlations at 30% and 50% missing data proportions. At this point, after reaching a 30% missing data proportion, further increases in sample size did not affect the mean correlations.

While generating missing data for MNAR, the probability of missing data was associated with the ability of the test takers. High-ability test takers were less likely to omit items, whereas low-ability test takers were more likely to do so. Consequently, the missing data contained information about the latent trait being measured. In the IRTree modeling, the negative, moderate to high correlations between missing propensity and proficiency, especially in cases with 10% or more omitted responses, successfully reflect this situation.

Discussion

To ensure accurate assessments and evaluations in educational settings, it is important to effectively handle missing data. Understanding the missing data mechanism present in the dataset is crucial to using appropriate handling methods that avoid bias and error in parameter estimations. The literature shows that various methods can identify MCAR, but distinguishing between MAR and MNAR is challenging (Enders, 2010; McKnight et al., 2007). Consequently, researchers often cannot determine whether the missing data meet the MAR or MNAR assumption, potentially leading to biased parameter estimates. This study aimed to evaluate the efficacy of the IRTree framework in detecting missing data mechanisms under different test conditions. Specifically, it focused on items that respondents omitted for various reasons in dichotomously scored tests. Using a Monte Carlo simulation, datasets with missing data under MCAR, MAR, and MNAR were modeled with the IRTree model for omitted items. The mean Pearson correlation coefficients between the propensity to omit and proficiency were examined to detect the missing data mechanisms.

The IRTree modeling revealed no correlation between propensity to omit and proficiency under the MCAR mechanism, with mean correlations of 0.0 as observed in the literature (Glas & Pimentel, 2008; Köhler et al., 2017). For MAR, the mean correlations were negative, reaching up to -0.3, indicating that highly proficient test takers were less likely to omit items, though this relationship is weak. According to Holman and Glas (2005), nonignorability becomes evident when the correlation exceeds 0.4. Since the observed correlations in MAR remain below this threshold, the data can be considered ignorable. For MNAR, mean correlations were as high as -0.8, especially at the highest missing data proportion, indicating that the missing data process contains information about the proficiency. As test takers' proficiency increases, their propensity to omit decreases. This result is consistent with the finding that as the correlation between propensity to omit and proficiency increases, the level of nonignorability also rises (Glas & Pimentel, 2008; Pohl et al., 2014).

Varying missing data proportions, test lengths, and sample sizes did not affect the relationship between propensity to omit and proficiency for the MCAR mechanism, and the IRTree was able to detect MCAR under all conditions. For MAR and MNAR, however, the relationship between propensity to omit and proficiency increased with the rising proportion of missing data, with a more pronounced impact in

MNAR. Additionally, for the relationship between propensity to omit and proficiency to reflect MAR and MNAR, the dataset should ideally contain at least 10% omitted responses. Increasing the number of items raised the mean correlations in both MAR and MNAR mechanisms, with a more pronounced impact in MNAR. For MAR and MNAR, increasing the sample size from 500 to 1000 led to an increase in the mean correlations with the effect being more pronounced for MNAR. While further increasing the sample size from 1000 to 2000 did not result in a substantial change. Consequently, increases in the proportion of omitted responses, test length and sample size strengthened the relationship between the propensity to omit and proficiency, especially making the MNAR mechanism more pronounced and easier to detect.

Overall, the study demonstrated that the IRTree framework could accurately detect missing data mechanisms across various scenarios with different sample sizes, test lengths, and missing proportions in dichotomously scored tests. This finding supports the utility of IRTree in identifying missing data mechanisms, as suggested by previous studies (Debeer et al., 2017; Jeon & De Boeck, 2016). Through IRTree analyses, it is possible to determine whether the measured latent traits are related to the process that leads to missing data, enabling more informed and rational decisions on how to handle omitted items in real-world educational assessments. For example, if omitting behavior has a high correlation with the measured latent trait, in such a case, the Selection Model, Pattern Mixture Model or IRTree can be used to deal with nonignorable missing data will need to be employed (Debeer et al., 2017; Enders, 2010; Holman and Glas, 2005). This will increase the accuracy of the results by preventing biased parameter estimates. Accurate and precise estimates will enhance the validity and reliability of assessments, preventing incorrect decisions about test takers. Especially in tests that measure students' achievements or attitudes, using IRTree to determine whether omission behavior is related to the measured trait will guide educators in their decision-making processes. Additionally, in large-scale educational assessments such as PISA, TIMMS, or PIRLS, which guide countries' educational policies, IRTree results can provide a clearer understanding of the nature of missing data.

This study focused on omitted responses in dichotomously scored tests. Future research can extend this work by investigating the effectiveness of the IRTree framework in detecting missing data mechanisms in polytomously scored tests. Additionally, future studies should test the efficacy of IRTree under different simulation scenarios. Different IRTtree models can be developed by considering various processes leading to missing data. While this study used the 2PL model for the nodes of the IRTree model, future research can explore the use of different IRT models.

Declarations

Conflict of Interest: The author reports there are no competing interests to declare.

Ethical Approval: I declare that all ethical guidelines for the author have been followed. Ethical approval is not required as data has been simulated in this study.

Funding: The author received no financial support for the research, authorship, and/or publication of this article.

References

Alagöz, Ö. E. C., & Meiser, T. (2023). Investigating heterogeneity in response strategies: A mixture multidimensional IRTree approach. *Educational and Psychological Measurement, 84*(5), 957-993. https://doi.org/10.1177/00131644231206765

Alarcon, G. M., Lee, M. A., & Johnson, D. (2023). A Monte Carlo study of IRTree models' ability to recover item parameters. *Frontiers In Psychology, 14*, 1003756. https://doi.org/10.3389/fpsyg.2023.1003756

Allison, P. D. (2002). *Missing data.* Sage Publications.

Baker, F. B. (2001). *The basics of item response theory*. ERIC Clearinghouse on Assessment and Evaluation.

Bock, R. D., & Aitkin M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443-459. https://doi.org/10.1007/BF02293801

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*(4), 431-444. https://doi.org/10.1177/014662168200600405

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-460). MA: Addison-Wesley.

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods, 17*(4), 665-678. https://doi.org/10.1037/a0028111

Böckenholt, U. (2017). Measuring response styles in Likert items. *Psychological methods, 22*(1), 69–83. https://doi.org/10.1037/met0000106

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Cheema, J. R. (2014). A review of missing data handling methods in education research. *Review of Educational Research, 84*(4), 487-508. https://doi.org/10.3102/0034654314532697

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001) A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*(4), 330-51.

Damiani, V. (2016). Large-scale assessments and educational policies in Italy. *Research Papers in Education, 31*(5), 529–541.

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*(3), 213–234. https://doi.org/10.1111/j.1745-3984.2001.tb01124.x

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree based item response models of the GLMM family. *Journal of Statistical Software, 48*, 1-28. https://doi.org/10.18637/jss.v048.c01

Debeer, D., Janssen, R., & De Boeck, P. (2017). Modeling skipped and not-reached items using IRTrees. *Journal of Educational Measurement, 54*(3), 333-363. https://doi.org/10.1111/jedm.12147

DeMars, C. (2010). *Item response theory: Understanding statistics measurement.* Oxford University Press.

Dibek, M. I. (2019). Examination of the extreme response style of students using IRTree: The case of TIMMS 2015. *International Journal of Assessment Tools in Education, 6*, 300-313. https://doi.org/10.21449/ijate.534118

Enders, C. K. (2010). *Applied missing data analysis.* The Guilford Press.

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice, 35*(2), 36-49. https://doi.org/10.1111/emip.12111

Glas, C. A. W., & Pimentel, J. L. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement, 48*(6), 907-922. https://doi.org/10.1177/0013164408315262

Glas, C. A. W., Pimentel, J. L., & Lamers, S. M. A. (2015). Nonignorable data in IRT models: Polytomous models with covariates. *Psychological Test and Assessment Modeling, 57*(4), 523-541.

Graham, J. W. (2012). *Missing data analysis and design.* Springer.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*. California: Sage Publications.

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*(2), 101-125. https://doi.org/10.1177/014662169602000201

Holman, R., & Glas, C. A. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Pyschology, 58*, 1-17. https://doi.org/10.1111/j.2044-8317.2005.tb00312.x

Huang, H. Y. (2020). A mixture IRTree model for performance decline and nonignorable missing data. *Educational and Psychological Measurement, 80*(6), 1168-1195. https://doi.org/10.1177/0013164420914711

Huisman, M. (2000). Imputation of missing item responses: Some simple techniques. *Quality & Quantity, 34*, 331–351. https://doi.org/10.1023/A:1004782230065

Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods, 48*, 1070-1085. https://doi.org/10.3758/s13428-015-0631-y

Jeon, M., De Boeck, P., & van der Linden, W. (2017). Modeling answer change behavior: An application of a generalized item response tree model. *Journal of Educational and Behavioral Statistics, 42*(4), 467-490. https://doi.org/10.3102/1076998616688015

Jeon, M., Rijmen, F. & Rabe-Hesketh, S. (2014). Flexible item response theory modeling with FLIRT. Applied Psychological Measurement, 38, 404-405. https://doi.org/10.1177/0146621614524982

Jin, K.-Y., Wu, Y.-J., & Chen, H.-F. (2022). A new multiprocess IRT model with ideal points for likert-type items. *Journal of Educational and Behavioral Statistics, 47*(3), 297-321. https://doi.org/10.3102/10769986211057160

Köhler, C., Pohl, S., & Carstensen, C. (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. https://doi.org/10.1111/jedm.12154

Leventhal, B. C. (2019). Extreme response style: A simulation study comparison of three multidimensional item response models. *Applied Psychological Measurement, 43*(4), 322–335. https://doi.org/10.1177/0146621618789392

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* John Wiley & Sons.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*, 1198–1202.

Little, T. D., Lang, K. M., Wu, W., & Rhemtulla, M. (2016). Developmental psychopathology. In D. Cicchetti (Ed.), *Missing Data* (pp. 760-797). John Wiley & Sons.

Martens, K., Niemann, D., & Teltemann, J. (2016). Effects of international assessments in education – a multidisciplinary review. *European Educational Research Journal, 15*(5), 516-522. https://doi.org/10.1177/1474904116668886

McKnight, P. E., McKnight, K. M., Sidani, S. & Figueredo, A. J. (2007). *Missing data: A gentle introduction.* Guilford Press.

Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational research methods, 17*(4), 372-411. https://doi.org/10.1177/1094428114548590

Park, M., & Wu, A. D. (2019). Item response tree models to investigate acquiescence and extreme response styles in Likert-type rating scales. Educational and Psychological Measurement, 79(5), 911–930. https://doi.org/10.1177/0013164419829855

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, *74*(4), 525-556. https://doi.org/10.3102/00346543074004525

Pigott, T. D. (2010). A review of methods for missing data. *Educational Research and Evaluation: An International Journal on Theory and Practice, 7*(4), 353-383. https://doi.org/10.1076/edre.7.4.353.8937

Plieninger, H. (2021). Developing and applying Ir-Tree models: Guidelines, caveats, and an extension to multiple groups. *Organizational Research Methods, 24*(3), 654-670. https://doi.org/10.1177/1094428120911096

Pohl, S., Gräfe, L. & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models. Educational and Psychological Measurement, 74(3), 423–452. https://doi.org/10.1177/0013164413504926

Quirk, V. L., & Kern, J. L. (2023). Using IRTree models to promote selection validity in the presence of extreme response styles. *Journal of Intelligence, 11*(11), 216. https://doi.org/10.3390/jintelligence11110216

Rose, N., von Davier, M., & Nagengast, B. (2015). Modeling omitted and not-reached items in IRT models. *Psychometrika, 82*, 795-819. https://doi.org/10.1007/s11336-016-9544-7

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.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology, 47*(3), 537–560. https://doi.org/10.1111/j.1744-6570.1994.tb01736.x.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592. https://doi.org/10.1093/biomet/63.3.581

Spratto, E. M., Leventhal, B. C., & Bandalos, D. L. (2021). Seeing the forest and the trees: Comparison of two IRTree models to investigate the impact of full versus endpoint-only response option labeling. *Educational and Psychological Measurement, 81*(1), 39-60. https://doi.org/10.1177/0013164420918655

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. https://doi.org/10.1007/s00357-017-9220-3

Tabachnick, B. G., & Fidell L. S. (2007). *Using multivariate statistics.* Allyn and Bacon.