

Effect of missing values on the matching item in the graded model

Haitham Fuad Jamil Darweesh¹, Haitham Mohammad Ali Zureigat²

¹Department of Class Teacher, Faculty of Educational Sciences and Arts, United Nations Relief and Works Agency, Amman, Jordan

²Educational Measurement and Evaluation, Jordanian Ministry of Education, Jerash Education Directorate, Jerash, Jordan

Article Info

Article history:

Received Dec 6, 2023

Revised Apr 14, 2024

Accepted Apr 23, 2024

Keywords:

Graded model

Item matching

Item response theory

Missing values

Simple regression

ABSTRACT

The study aimed to examine the effect of the percentage of missing values on the matching of the items of the graded model. To achieve this, the experimental approach was followed in conducting the study, as hypothetical data were used to simulate the experimental conditions represented by the percentage of missing values. It has three levels (5%, 10%, and 20%). Moreover, the compensation method is represented by simple regression. The study found that there were no statistically significant differences in the percentage of items matching the graded model according to the missing percentages (5% and 10%). Moreover, there were statistically significant differences in the percentage of items matching the graded model according to the missing percentages (20%). Therefore, we must use the different nonparametric item response theory (IRT) models because of the good information, it provides to researchers about the test and its items without missing items suitable for this purpose and prejudice—the cumulative response of the subjects.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Haitham Mohammad Ali Zureigat

Educational Measurement and Evaluation, Jordanian Ministry of Education, Jerash Education Directorate

Jerash, Jordan

Email: haiytham.zreqat@gmail.com

1. INTRODUCTION

Accuracy of measurement and obtaining representative data, whether for societies or behavioral samples, are among the most important basic pillars in judging the quality of the researcher's scientific production [1]. Moreover, one of the most important issues that researchers in all fields may face is the problem of missing values, which prompted researchers to think about solving this problem [2]. Even the results of studies and research become more accurate by handling missing values with a procedure that ensures that it does not affect the good quality of the data matching the models and the validity and reliability of the results through handling missing values [3]. Therefore, in their experiments research and tests, the researchers sought to use statistical methods to obtain more results that are accurate by handling those missing values [4]. By following several methods to handle these missing values, such as regression, mean, median, and other methods; however, these methods are not without problems because they depend on observed values, which may lead to the emergence of some type of correlation between the test items [5]. In this regard, we find that almost no research is devoid of missing values [6]. This problem faces most researchers, the size of which increases as the size and nature of the missing values increase, which directly affects the results of research inferences and thus affects the results of the study and the ability to generalize the results from the research sample to the study population [3]. Regarding the quality of the scale items according to item response theory (IRT), the criterion for judging them lies in verifying four basic assumptions.

Unidimensionality is the assumption based on the existence of one factor that underlies performance on the scale, and this factor is the measured trait or ability [6]. Unidimensionality is achieved in a measure if the discrepancy between the subjects in performance on the items of the scale is due to only one factor, which is the measured trait or ability [4]. Unidimensionality items are those that evaluate the same characteristic and are homogeneous among themselves, as any question that must be answered follows the same behavioral procedures and processes, and the only difference between them is the level of difficulty [7]. However, many tests usually include many areas, and they share these domains and have basic characteristics in that they all measure the feature that is being measured [8]. In theory, there are four components or dimensions to arithmetic problems: addition, subtraction, multiplication, and division. However, we assume that these four dimensions measure arithmetic proficiency as one subject and it is difficult to find the expense of not combining them [6].

Local independence is the basic assumption and the cornerstone of all models of IRT [9]. Violating it leads to psychological errors in estimating individuals' abilities and the criteria of the elements (difficulty, discrimination, and guessing), which can be followed by the number of incorrect decisions of an individual on a test item that is independent of the outcome of his answer [7]. For any item in the test, in other words, the result of one item question is not affected positively or negatively by the result of any other question [5]. The assumption of the item characteristic curve, which reflects the real relationship between the ability variable and the observed variables (the response to an item), is known as the monotonicity relationship, and an assumption related to the characteristics of the items and their relationship to the performance of the subjects (freedom from speed) [8]. As for missing values and methods for dealing with them, which may lead to less efficient estimates, some rely on the deletion and some rely on compensatory value procedures [9]. In light of many estimation methods according to IRT models, many have emerged for us. A research issue is related to the accuracy of vertebral parameter estimation and power parameters [10].

It is worth noting that scientist Yates was the first to follow the method of deletion for cases that included missing values and to be satisfied with the sample size remaining after deletion [11]. Then the method of handling missing values by relying on statistical analysis related to variance and covariance appeared at the hands of the scientist Bartlett in 1939 [12]. After that, efforts continued to propose various methods for handling the problem of missing values, depending on the mechanism of loss, whether it was a completely random loss in which the missing values are independent of the rest of the other values [13]. Alternatively, random loss in which the missing values are related to other values of other variables and not related to the missing value itself, and finally the non-random loss, results from the missing value itself and is unrelated to other values [14]. With the advancement of IRT and its special methods for estimating the properties of psychometric scales, relying on its estimation of the parameters of the items: thresholds (difficulty), discrimination, guessing, and ability of the subjects [15]. Through mathematical models linking these properties, these models distinguish the theoretical item's response that differs from the rest of the measurement theories [16]. IRT models, both dichotomous and polytomous, are various probabilistic mathematical functions. The mathematical formula of the model varies depending on the number of features of the item that make up its mathematical structure [17]. These models aim to determine the relationship between the probabilities of an individual answering an item [4]. Correct answer and explain the underlying ability that causes this performance. These models include the Rasch model, the one-parameter logistic model, the two-parameter logistic model, the three-parameter logistic model, the partial scores model, and the graded response model (GRM) [4].

Because of the widespread use of numerical rating scales, such as Likert scales, in assessing many human traits, especially emotional and cognitive ones, the interest of theorists in item response has focused on deriving appropriate mathematical models for this scale such as the partial credit model (PCM) [18]. Moreover, the GRM is one of the models of the multi-graded IRT [19]. It resulted from the generalization of Birnbaum's two-parameter model using this model, extracting the greatest amount of information regarding the level of ability or trait to be measured using a fixed set of vocabulary [20]. By estimating one parameter to distinguish the item (a_i), and the item difficulty index (b_i), called the parameter difficulty threshold; their number is one less than the number of grading levels (four thresholds in the case of a five-point Likert scale) for each response section [21].

Test and scale developers strive seriously to build high-quality scales and tests with appropriate psychometric properties that lead to a high degree of measurement accuracy. However, one of the obstacles that emerges later after applying these tests is the loss of data collected, because of the lack of full response by the subjects. For any reason, this requires replacing these missing values to provide the possibility of analyzing them through various analysis programs and producing results. However, the various compensation processes for missing data are not without problems, as they, especially in parametric models, may lead to an increase in type I errors [22]. In addition, in graded models with more stringent assumptions, it is necessary to preserve the cumulative response of the subjects. Which may be distorted because of the presence of

missing values, even if one of the compensation methods replaces them [23]. Based on these issues, the purpose of this paper is to determine the effect of the percentage of missing values (5%, 10%, and 20%) on the matching of items to the graded model after handling it by regression analysis method. Furthermore, there were several terminologies of study:

- Estimation accuracy: a statistical indicator that expresses the quality of a parameter estimate and is measured for the value of the standard error of the estimate [24].
- Missing values: some respondents left some items in the test without answering them [25].
- Estimation methods: it is the statistical method or method used to estimate the parameter (difficulty, ability) [26].
- Graded-response model: a non-linear relationship between an individual's ability level and the probability of his response at a certain level of grading, which is referred to as the Threshold, which represents the trait level necessary for the individual's response to exceed the different threshold with a probability of 0.5, and a number of grading levels. The scale is equal to the number of thresholds plus one [27].

2. RESEARCH METHOD

The experimental approach was followed in conducting the research, as hypothetical data was used that simulated the experimental conditions represented by the percentage of missing values and has three levels (5%, 10%, and 20%) and the compensation method was represented by simple regression. The use of generated data has been resorted to because it allows for the possibility of missing responses from individuals, and this is difficult to provide in realistic studies because subjects respond to all items by guessing, even if they do not know the correct answer. The generated data also provides many standard conditions that are difficult to obtain when using realistic data, including an appropriate distribution of the abilities of the subjects, appropriate distributions of the parameters of the items used in the study, as well as the appropriate loss in responses in specific proportions.

2.1. Respondent

Responses for this study were obtained by following the procedures:

- Generating a hypothetical sample of respondents with a size of (1,000) individuals, guided by several studies [4], [28], which confirmed that this number of examiners is sufficient to obtain appropriate estimates of the parameters of the vertebrae, with the least error in the accuracy of the estimate. Whose ability follows a normal distribution with an arithmetic mean (0) and a standard deviation (1), using (WinGen) software.
- Generate a virtual test it consists of (50) items whose items follow the GRM using (WinGen) software.
- Generate responses for individuals in step: to the test and to form a matrix of hypothetical responses, using (WinGen) software.
- Randomly missing values in the response matrix at different rates (5%, 10%, and 20%) using Excel.
- Handling the missing values in the two tests after performing the missingness using a simple regression method using (SPSS) software.

2.2. Verifying the assumptions of the response theory of the item

2.2.1. Verify the assumption of unidimensionality

This assumption was verified by conducting an exploratory factor analysis test based on principal component analysis, to find the eigenvalue and the percentages of explained variance for both the first and second component, for the original data that was generated as: i) it is clear from Table 1 that the ratio of the Eigenvalue of the first factor to the ratio of the Eigenvalue of the second factor reached (10), which is greater than (2). Moreover, the percentage of variance explained by the first component is greater than (20%), which are indicator that confirms the unidimensional assumption is fulfilled [29]; and ii) in addition, this can be observed through the following graphical representation of the underlying roots (scree plot) from Figure 1 that there is a sharp inflection when there are two factors, which indicates the presence of only one dominant characteristic in the scale.

Table 1. Exploratory factor analysis for study data

Source	Eigen value	Percentages of explained variance (%)
First component	11.890	23.779
Second component	1.189	2.378

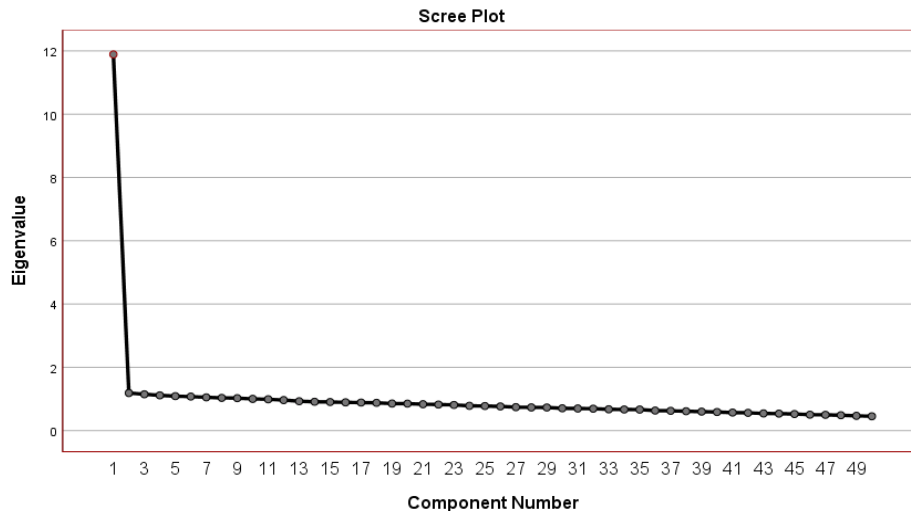


Figure 1. Scree plot for study data

2.2.2. Verifying the assumption of local independence

The assumption of positional independence was verified by entering data into the (R-studio) program in the form of a text file for the test consisting of (50) items. The data of which were generated through six experimental conditions on a sample of (1,000) respondents for one time. To reveal pair the related items by calculating the index (Q3) based on the correlation of their residual.

Moreover, what is a reference to it, that the local independence is not exposed with the related between the items that the test or the scale formed when it is applied to the sample individuals. In addition, this is a requirement to obtain a test of high accuracy in the measurement of the feature, which means the need for items that are consistent with each other in measuring the trait. The benefit of using IRT models is achieved by providing a good match between the mathematical model and the responses of the subjects on the scale, and this requires evaluating the goodness of fit through: evaluating the verification of the assumptions of the mathematical model in the data, and the goodness of fit of the model's expectations.

The idea of good matching of items is based on comparing the mathematical model's prediction of the respondent's score with the apparent response. This occurs at the level of the scale items together or at the level of each item. Many methods are based on examining the fit of items, including Chi-square, Chi-square odds ratio, and standardized residuals index.

The (Q3) test showed that the total number of pairs equals (1,225) pairs, of which only two pairs are not independent, with a percentage of (0.16%) of the total pairs. This shows that the number of pairs of vertebrae that achieved the assumption of local independence is four times higher than the number of pairs that achieved positional dependency, which amounted to (611.3). This is an indication that the assumption of local independence has been fulfilled [30].

2.2.3. Verifying the conformity of the item according to the graded response model

The matching of items to the GRM was verified based on the probability value of the item matching index (S-X2), which is one of the most important indicators of matching in the case of data with more than 200 respondents [3]. It is clear from Table 1 that the number of items conforming to the graded model is four times greater than the number of items that do not conform to the model. This confirms that the assumption of local independence has been met.

It is clear from Table 2 that two items do not conform to the GRM, as the probability value of the generalized item conformity index (S-X2) was less than (0.05), and they are items no. 14, 16, and thus it can be said the percentage of matching items reached (96%), which is the percentage that will be relied upon later to examine the differences in the effect of the percentage of missing values on the matching of items according to the graded model. Obtaining this high percentage of paragraphs matching the graded model results from the use of generated data, within a set of selected experimental conditions, which is something that cannot be obtained by collecting data from studies that are applied in reality. This allows us to obtain standard results that can be compared after performing data loss operations, and then treating them statistically through the simple regression method.

Table 2. Matching values for the scale items before the missing value

Item no	S_X2	df.S_X2	RMSEA.S_X2	p.S_X2	Matching
x1	276.2785	272	0.003968	0.416429	Matching
x2	284.5387	299	0	0.716949	Matching
x3	176.0604	184	0	0.649918	Matching
x4	218.1462	211	0.005823	0.353314	Matching
x5	238.7191	209	0.011931	0.07749	Matching
x6	264.964	276	0	0.6728	Matching
x7	197.3342	207	0	0.673555	Matching
x8	209.8006	218	0	0.642697	Matching
x9	212.5075	202	0.007216	0.292187	Matching
x10	129.0676	153	0	0.920453	Matching
x11	239.8659	232	0.005826	0.347534	Matching
x12	272.4429	244	0.010802	0.101941	Matching
x13	282.7876	285	0	0.525898	Matching
x14	312.6927	269	0.012751	0.034471	Not matching
x15	270.9058	254	0.008162	0.222591	Matching
x16	271.9202	233	0.012931	0.040771	Not matching
x17	268.3386	237	0.011505	0.079156	Matching
x18	281.1455	269	0.006723	0.293017	Matching
x19	193.493	189	0.004878	0.396074	Matching
x20	241.7469	241	0.001761	0.474345	Matching
x21	209.3983	236	0	0.893018	Matching
x22	288.7023	292	0	0.543542	Matching
x23	216.1159	282	0	0.998636	Matching
x24	262.7431	260	0.00325	0.440743	Matching
x25	229.9161	259	0	0.903118	Matching
x26	303.8193	282	0.008801	0.177749	Matching
x27	257.656	250	0.005537	0.356177	Matching
x28	208.0234	182	0.011964	0.090371	Matching
x29	264.0471	244	0.009069	0.180425	Matching
x30	202.6492	218	0	0.764559	Matching
x31	286.5125	271	0.00757	0.247455	Matching
x32	190.0199	203	0	0.734075	Matching
x33	158.0213	153	0.005732	0.373732	Matching
x34	261.5634	248	0.007399	0.264982	Matching
x35	307.018	298	0.005504	0.34717	Matching
x36	271.6856	250	0.009318	0.165347	Matching
x37	230.9501	224	0.005573	0.360752	Matching
x38	287.8113	283	0.004125	0.409401	Matching
x39	117.5492	111	0.007685	0.31713	Matching
x40	273.9105	293	0	0.781919	Matching
x41	277.9233	259	0.008552	0.200136	Matching
x42	211.0586	221	0	0.673035	Matching
x43	187.0855	204	0	0.796299	Matching
x44	171.8309	170	0.003283	0.446277	Matching
x45	201.5333	214	0	0.719826	Matching
x46	205.1809	184	0.010734	0.135919	Matching
x47	262.8818	244	0.008801	0.193895	Matching
x48	224.5918	216	0.00631	0.329971	Matching
x49	198.3814	171	0.01266	0.074392	Matching
x50	218.4064	256	0	0.957344	Matching

2.3. Data analysis

To reach the results of the study, we relied on several preliminary statistical tests, which were exploratory factor analysis using the SPSS program, as well as verifying the assumption of local independence using the Q3 test. The conformity of the paragraphs to the GRM was also verified based on the probability value of the item conformity index (S-X2) through R software. Then the percentage of items, matching the GRM was calculated through R software depending on the percentage of the missing value and after processing this by simple regression method using the SPSS program.

3. RESULTS AND DISCUSSION

3.1. Results

Table 3 displays that the percentage of items matching the graded model reached (94%, 90%, and 78%) for the missing values percentages (5%, 10%, and 20%) respectively. As it appears to us that there are apparent differences in the matching percentages. In addition, to verify the significance of these differences, the (Z) test was relied upon to examine the differences between the proportions, through the following mathematical as in (1).

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p} \cdot \hat{q} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad (1)$$

Table 3. Conformity values for the scale items after handling missing values

Ratio missing values (%)	Total number of items	Number of matching items	Number of mismatched items	Percentage of matching items (%)
5	50	47	3	94
10	50	45	5	90
20	50	39	11	78

Table 4 shows that there are no differences in the percentage of vertebrae matching the graded model according to the missing value rates (5% and 10%). This result can be inferred from its Z-test value, which reached (0.46 and 1.76), respectively. It is less than the tabular value of (Z) at the level of significance ($\alpha/2$)=(0.025), which is equal to (1.96). It is also clear from the previous table that there are differences in the proportion of matching items at the missing value rate (20%), as the calculated (Z) value for them reached (2.68), which is greater than (1.96).

Table 4. Z-test to examine the differences between the matching percentages depending on the percentages of the missing value

Source	Z value at missing value (5%)	Z value at missing value (10%)	Z value at missing value (20%)
Original data before missing value and handling	0.46	1.76	2.68
Significance	Not statistically significant	Not statistically significant	Statistically significant

3.2. Discussion

The results showed that there were no statistically significant differences in the proportion of vertebrae matching the graded model according to the missing value rates (5% and 10%). That is, the aforementioned missing value rates did not affect the probability value of the items matching index (S-X2). It also did not have a clear effect on the cumulative graded response as one of the requirements of these models. The number of non-matching items was relatively small and did not exceed the permissible error rate in this type of human study. The results showed that there were statistically significant differences regarding the presence of differences in the proportion of matching items at the missing value rate (20%). It is noted that the percentage of vertebrae that do not conform to the stepwise model has a direct relationship with the percentage of missing values. This can be explained by the imbalance caused by compensating for missing values in the cumulative graded response. Compensation for missing values is based on observed responses to the item and does not take into account the cumulateness necessary for the stepwise model. This is considered one of the most important requirements of these models, and hence the increase in the percentage of missing values shows a lack of cumulative response patterns in more items. This contradicts the expected pattern according to the stepwise model and is the direct cause of the item mismatch.

Thus, the problem of missing values raises many problems, perhaps the most important of which, according to several studies [30], [31] is the possibility of missing values reducing the strength of the statistical test results (test power). A defect appeared in the cumulative graded response because of the defect in the sample size and test length when items and individuals that did not conform to the model were deleted [25]. Other problems lie in the bias in estimating the parameters of the scale, especially when estimating validity and reliability, which increases the value of the standard error in the accuracy of estimating the parameters of the items [1]. Also, the problem of missing individuals' responses to several test items has an expected impact on the accuracy of estimating the parameters of the items and the power parameter, because of the defect that occurs in the pattern of response when those values are substituted [3]. The lower percentage of data loss maintains the fulfilment of assumptions and the cumulative response to the item better than the higher percentages of data loss, according to a study [21].

Moreover, the increase in the number of items that do not conform to the graduated model can be attributed to the difficulty and strictness of the assumptions that items conform to the model [27]. The logistical form of the item response function is the most stringent [20]. In addition, there are many reasons behind the items not matching the observed data, such as the presence of more than one dimension behind the responses of the sample members, and that the number of features for the items in question in the approved model is not sufficient [17]. Moreover, the description curve of the items in question is not increasing, or

there is a possibility that the population in question in which the parameters of the items were estimated includes a subpopulation that is not homogeneous with the overall population [28]. From a practical standpoint, the presence of a large number of items that do not conform to the model. Due to the presence of a loss in the collected data, leads to inaccurate estimates of students' abilities and the parameters of the items thus making inaccurate decisions regarding students.

4. CONCLUSION

The study concluded that as the percentage of missing value increased, the percentage of non-compliant vertebrae became clear and statistically significant, as the percentage of missing value (20%) is considered a high percentage and exceeds the statistically permissible percentages. This is what calls for relying on nonparametric models to estimate the parameters of items and the abilities of individuals. When there are high rates of missing value, as these models are less stringent in verifying their assumptions and therefore a large number of items and individuals that do not conform to the model will not be deleted. Which makes it more the ability to deal with the cumulative response without disrupting the style of the model used.





REFERENCES

- [1] S. Wang, W. Li, L. Hu, J. Cheng, H. Yang, and Y. Liu, "NAGuideR: performing and prioritizing missing value imputations for consistent bottom-up proteomic analyses," *Nucleic Acids Research*, vol. 48, no. 14, p. e83, Aug. 2020, doi: 10.1093/nar/gkaa498.
- [2] F. Yang *et al.*, "Missing value estimation methods research for arrhythmia classification using the modified kernel difference-weighted KNN algorithms," *BioMed Research International*, vol. 2020, pp. 1–9, Jun. 2020, doi: 10.1155/2020/7141725.
- [3] W.-C. Lin and C.-F. Tsai, "Missing value imputation: a review and analysis of the literature (2006–2017)," *Artificial Intelligence Review*, vol. 53, no. 2, pp. 1487–1509, Feb. 2020, doi: 10.1007/s10462-019-09709-4.
- [4] S. Monroe, *Item response theory*. Routledge, 2022, doi: 10.4324/9781138609877-REE61-1.
- [5] F. B. Baker and S.-H. Kim, "Item characteristic curve models," in *The basics of item response theory using R*, Cham: Springer International Publishing, 2017, pp. 17–34, doi: 10.1007/978-3-319-54205-8_2.
- [6] J. Liu, H. Lin, B. Hu, Z. Zhou, E. Agyeiwaah, and Y. Xu, "Advancing the understanding of the resident pro-tourism behavior scale: an integration of item response theory and classical test theory," *Journal of Business Research*, vol. 141, pp. 113–125, Mar. 2022, doi: 10.1016/j.jbusres.2021.12.013.
- [7] M. C. Edwards, C. R. Houts, and L. Cai, "A diagnostic procedure to detect departures from local independence in item response theory models," *Psychological Methods*, vol. 23, no. 1, pp. 138–149, Mar. 2018, doi: 10.1037/met0000121.
- [8] J. A. Rios and J. Soland, "Parameter estimation accuracy of the effort-moderated item response theory model under multiple assumption violations," *Educational and Psychological Measurement*, vol. 81, no. 3, pp. 569–594, Jun. 2021, doi: 10.1177/0013164420949896.
- [9] M. Wiley and J. F. Wiley, "Missing data," in *Advanced R statistical programming and data models*, Berkeley, CA: Apress, 2019, pp. 383–433, doi: 10.1007/978-1-4842-2872-2_9.
- [10] G. Spirova, T. Eftimov, and B. K. Seljak, "Evaluating missing value imputation methods for food composition databases," *Food and Chemical Toxicology*, vol. 141, p. 111368, Jul. 2020, doi: 10.1016/j.fct.2020.111368.
- [11] A. Shehu, N. S. Dauran, and A. G. Usman, "Estimation of missing value in Sudoku square design," *Asian Journal of Probability and Statistics*, vol. 24, no. 4, pp. 11–16, Sep. 2023, doi: 10.9734/ajpas/2023/v24i4530.
- [12] T. Su, Y. Shi, J. Yu, C. Yue, and F. Zhou, "Nonlinear compensation algorithm for multidimensional temporal data: a missing value imputation for the power grid applications," *Knowledge-Based Systems*, vol. 215, 2021, doi: 10.1016/j.knsys.2021.106743.
- [13] J. R. van Ginkel, M. Linting, R. C. A. Rippe, and A. van der Voort, "Rebutting existing misconceptions about multiple imputation as a method for handling missing data," *Journal of Personality Assessment*, vol. 102, no. 3, pp. 297–308, May 2020, doi: 10.1080/00223891.2018.1530680.
- [14] F. Tang and H. Ishwaran, "Random Forest missing data algorithms," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 10, no. 6, pp. 363–377, Dec. 2017, doi: 10.1002/sam.11348.
- [15] S. K. Johnson, E. L. de Moor, J. M. Chung, L. den Boer, and T. A. Klimstra, "An item response theory analysis of two frequently used measures of identity development," *Psychological Assessment*, vol. 34, no. 11, pp. 1047–1061, Nov. 2022, doi: 10.1037/pas0001171.
- [16] A. Ali, N. A. Emran, and S. A. Asmai, "Missing values compensation in duplicates detection using hot deck method," *Journal of Big Data*, vol. 8, no. 1, p. 112, Dec. 2021, doi: 10.1186/s40537-021-00502-1.
- [17] L. Stanke and O. Bulut, "Explanatory item response models for polytomous item responses," *International Journal of Assessment Tools in Education*, vol. 6, no. 2, pp. 259–278, Jul. 2019, doi: 10.21449/ijate.515085.
- [18] A. S. Edwards, K. E. Edwards, and B. C. Wesolowski, "The psychometric evaluation of a wind band performance rubric using the multifaceted Rasch partial credit measurement model," *Research Studies in Music Education*, vol. 41, no. 3, pp. 343–367, Oct. 2019, doi: 10.1177/1321103X18773103.
- [19] A. F. Kiliç, "Can factor scores be used instead of total score and ability estimation?" *International Journal of Assessment Tools in Education*, vol. 6, no. 1, pp. 25–35, Mar. 2019, doi: 10.21449/ijate.442542.
- [20] R. Wang, X. Xu, and B. Gu, "Statistical analysis of two-parameter generalized Birnbaum-Saunders Cauchy distribution," *Journal of Applied Analysis & Computation*, vol. 10, no. 4, pp. 1614–1635, 2020, doi: 10.11948/20190293.
- [21] A. N. Rani and A. S. Davamani, "Classification on missing data for multiple imputations," *International Journal of Trend in Scientific Research and Development*, vol. 2, no. 3, pp. 745–749, Apr. 2018, doi: 10.31142/ijtsrd9566.
- [22] S. G. Meintanis, B. Milošević, and M. Obradović, "Goodness-of-fit tests in conditional duration models," *Statistical Papers*, vol. 61, no. 1, pp. 123–140, Feb. 2020, doi: 10.1007/s00362-017-0930-8.
- [23] A. M. Sefidian and N. Daneshpour, "Missing value imputation using a novel grey based fuzzy c-means, mutual information based feature selection, and regression model," *Expert Systems with Applications*, vol. 115, pp. 68–94, Jan. 2019, doi: 10.1016/j.eswa.2018.07.057.





- [24] C. J. Harrison *et al.*, “Item response theory assumptions were adequately met by the Oxford hip and knee scores,” *Journal of Clinical Epidemiology*, vol. 158, pp. 166–176, Jun. 2023, doi: 10.1016/j.jclinepi.2023.04.008.
- [25] W.-C. Lin, C.-F. Tsai, and J. R. Zhong, “Deep learning for missing value imputation of continuous data and the effect of data discretization,” *Knowledge-Based Systems*, vol. 239, p. 108079, Mar. 2022, doi: 10.1016/j.knosys.2021.108079.
- [26] Y. Xia and Y. Yang, “RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: the story they tell depends on the estimation methods,” *Behavior Research Methods*, vol. 51, no. 1, pp. 409–428, Feb. 2019, doi: 10.3758/s13428-018-1055-2.
- [27] O. J. Kehinde, S. Dai, and B. French, “Item parameter estimations for multidimensional graded response model under complex structures,” *Frontiers in Education*, vol. 7, p. 947581, Sep. 2022, doi: 10.3389/educ.2022.947581.
- [28] Y. A. Al Mahrouq, “Effect of item difficulty and sample size on the accuracy of equating by using item response theory,” *Journal of Educational and Psychological Studies [JEPS]*, vol. 10, no. 1, pp. 182–200, Jan. 2016, doi: 10.24200/jeps.vol10iss1pp182-200.
- [29] D. M. Keating and F. J. Boster, “Nonlinear unidimensionality in communication science: tests, examples, and implications,” *Communication Research Reports*, vol. 36, no. 1, pp. 67–77, Jan. 2019, doi: 10.1080/08824096.2018.1555524.
- [30] S. Su, C. Wang, and D. J. Weiss, “Performance of the $S - \chi^2$ statistic for the multidimensional graded response model,” *Educational and Psychological Measurement*, vol. 81, no. 3, pp. 491–522, Jun. 2021, doi: 10.1177/0013164420958060.
- [31] R. A. Hughes, J. Heron, J. A. C. Sterne, and K. Tilling, “Accounting for missing data in statistical analyses: multiple imputation is not always the answer,” *International Journal of Epidemiology*, vol. 48, no. 4, pp. 1294–1304, Aug. 2019, doi: 10.1093/ije/dyz032.

BIOGRAPHIES OF AUTHORS



Haitham Fuad Jamil Darweesh     is Assistant Professor in the Faculty of Educational Sciences and Art, United Nations Relief and Works Agency (UNRWA). He worked as a teacher in UNRWA schools from 2006 until 2015. He also worked as an educational expert in UNRWA schools. He also worked as a teacher at the UNRWA College of Educational Sciences as an Assistant professor since 2016 and still is. He has an interest in studying organizational cultures, teachers' professional criteria, and item response theory models. He can be contacted at email: dhythm909@gmail.com.



Haitham Mohammad Ali Zureigat     is a Doctoral in Educational measurement and evaluation from Yarmouk University. He has worked as a teacher of vocational and technical education in the Jordanian Ministry of Education since 2004. His research focuses on the impact of methods for handling missing values with different percentages on the item parameters accuracy estimated according to parametric and non-parametric models, as well as the professional development of students. He can be contacted at email: haiytham.zreqat@gmail.com.