Research Report

Methods for Imputing Scores When All Responses Are Missing for One or More Polytomous Items: Accuracy and Impact on Psychometric Property

ETS RR–23-07

Yanxuan Qu
Sandip Sinharay

December 2023

ETS

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Methods for Imputing Scores When All Responses Are Missing for One or More Polytomous Items: Accuracy and Impact on Psychometric Property

Yanxuan Qu, & Sandip Sinharay

ETS, Princeton, New Jersey, USA

Though a substantial amount of research exists on imputing missing scores in educational assessments, there is little research on cases where responses or scores to an item are missing for all test takers. In this paper, we tackled the problem of imputing missing scores for tests for which the responses to an item are missing for all test takers. We considered three missing-data imputation methods—the median method, the item response theory (IRT) method, and the two-way method—for imputing scores. We compared the performance of these three imputation methods with respect to their accuracy in estimating scaled scores and test reliability for the aforementioned problem. Real data were used in the comparison. All three methods performed well in imputing scaled scores with negligible imputation error: The IRT method and the median method provided slightly more accurate scaled scores. The two-way method provided the most accurate reliability estimates. Recommendations for practice are provided.

Keywords constructed-response item; imputation of missing data; two-way method

doi:10.1002/ets2.12369

In educational testing, responses of test takers to items are sometimes missing due to unforeseen events that are unrelated to examinee behavior. Examples of unforeseen events are answer sheets being lost during transit and technical problems such as power outage, internet failure, and speaker or microphone malfunction during or after a test administration (Sinharay, 2022).

Researchers have proposed various methods for imputation of missing data. Huisman and Molenaar (2001) and Sinharay (2021) compared different missing-data imputation methods in terms of their accuracy in estimating individual raw or scaled scores. Sinharay provided a comprehensive review and comparison of six imputation methods—(a) person mean imputation, (b) linking, (c) regression imputation, (d) item response theory (IRT) imputation, (e) multiple imputation using data augmentation and chained equations, and (f) data mining—in terms of their accuracy in imputing total scaled scores. In his study, 10–50% of the responses were missing for each item. He found that all the imputation approaches performed rather well, especially the data mining, multiple imputation, and regression methods. Although Sinharay provided guidance about the methods to use when each item has both missing and observed responses, his study did not address the problem of imputing scores for an item when responses are missing for all test takers.

The current study focused on one specific situation that test administrators may encounter in practice—the situation where all responses to a polytomous item are missing and item scores have to be imputed during each administration. This is likely to occur in several situations. For example, operational tests are occasionally redesigned or revised, and a redesign often involves the addition of new items to the test or replacing old items with new items. Consequently, test administrators may want to know how the revisions will change the psychometric properties (e.g., test reliability) of the test scores before administering the new version, or test administrators may need to create or replace a raw-to-scale conversion table for the test scores. Given that there are no empirical response data for the new items prior to administering the new test, the problem may be viewed as one where all responses to the new items are missing. In other situations, all the responses for an item may be missing due to technical difficulties or problems like erroneous item text or answer keys such that the responses to the item must be marked "missing" for all examinees. When equating is feasible, items with technical difficulties or content problems can be removed from the test without the need to impute scores, but when equating is

Corresponding author: Yanxuan Qu, E-mail: yqu@ets.org

Table 1 The Types and Score Ranges of the 13 Items on the Pseudotest

Item number	Item type	Score range
1 – 4	1	0 – 3
5 – 6	2	0 – 3
7 – 12	3	0 – 3
13	4	0 – 5

not feasible, removing an item from a test form will cause fairness issues. One simple approach to address these issues involves the imputation of scores for these new and possibly problematic items.

In this paper, we explored various approaches to impute polytomous item scores when the scores on an item were 100% missing (i.e., missing for all test takers). Four of the six imputation methods (regression, linking, data mining, and multiple imputation using chained equations) included in Sinharay (2021) cannot handle cases with 100% missing answers for an item; these methods were not considered. Consequently, we included the median method (which is used in operation for imputing missing values for some tests) and the IRT method that were considered in Sinharay. We also included the two-way imputation method (van Ginkel et al., 2010) that is simple, but has rarely been considered in educational measurement. We compared the three imputation methods not only with respect to their accuracy at individual test score level but also with respect to their impact on the accuracy of the test reliability estimates and interitem correlations. The comparison was performed using data from 67 operational forms of a large-scale test. The research question was "How accurate are the estimated scaled score and test reliability after applying each imputation method if scores on one item are 100% missing?"

Note that it is possible to apply approaches that do not involve any imputation (such as the Spearman – Brown formula) to our problem, but we did not consider those approaches given the nature of the test that was the source of the data sets. If the items on a redesigned or revised test are all dichotomous and are quite homogeneous, the Spearman – Brown formula can provide accurate estimates of the corresponding test reliability; however, the items in these data were not dichotomous and were of different types with different score categories (as will be described shortly), so the Spearman-Brown formula was not used.

Data

The data used in the comparison of the imputation methods are real responses from 67 pseudoforms of a large-scale test. Each pseudoform has 13 CR items after removing one item from a real test. These items vary in item types and score categories, as shown in Table 1. Items 1 – 12 have a score range of 0 – 3. Item 13 has a score range of 0 – 5. Items with the same type are more similar to each other than items with different types. The scores on these 67 pseudoforms were not equated. A single raw-to-scale conversion was applied to all pseudoforms to calculate scale scores for all test takers. These scale scores were considered as test takers' actual scale scores in this study.

For each pseudoform, we assumed that scores on one item were missing, and we imputed the missing item scores by the three aforementioned methods. Data before imputation were used as criteria to evaluate the performance of the three imputation methods.

Analyses

In this study, for all the pseudoforms, we assumed that 100% of the scores on Item 6 were missing and imputed the scores on Item 6 for all test takers using three different imputation methods. We then calculated the total scaled scores based on the imputed scores on Item 6 plus the actual scores on the other 12 items. We called this scaled score the *imputed scaled score*. The imputed scaled score was compared to the actual scaled score for each test taker over the 67 pseudoforms. Item 6 and Item 5 are the same item type. They both have four possible score categories: 0, 1, 2, or 3. Test takers with blank responses or off-topic responses were scored as 0.

Following Sinharay (2021), we used the following three measures to compare the accuracy of the imputation methods in estimating the scaled scores: percent of zero-scaled score changes, standardized bias, and standardized root mean squared difference (RMSD). Standardized bias is the average of the differences between imputed scaled scores and actual scaled

scores expressed in units of the overall standard deviation (*SD*) of the scaled scores. To get the overall *SD*, the standard deviation of the actual scaled scores was first calculated for each form. The overall *SD* was then computed as the average of the 67 standard deviations across different forms. Standardized RMSD is the square root of the average of the squared differences between the imputed and actual scaled scores. Let $\widehat{T}_j$ and T_j denote the imputed and actual scaled scores of test taker j, and N is the total number of test takers in the final analysis sample for each imputation method. Then,

$$\text{Standardized RMSD} = \frac{1}{SD}\sqrt{\frac{\sum_j\left(\widehat{T}_j - T_j\right)^2}{N}}.$$

We summed over the 67 pseudoforms while computing each of the accuracy measures. As mentioned in Sinharay (2021), the standardized RMSD can be considered an effect size, roughly representing the average absolute error in imputing the scaled scores. Thus, a standardized RMSD smaller than 0.2 can be considered as representing a negligible score change. A standardized RMSD between 0.2 and 0.5 can be considered as representing a small to moderate score change.

We also calculated the reliability (using the coefficient alpha) for the test scores with and without imputation for Item 6. Test reliability without imputation was computed from the actual scores on all 13 items. Test reliability with imputation was computed from the imputed scores on Item 6 and actual scores on the other 12 items. In addition, we looked at interitem correlations involving Item 5 or 6 and other items with the same score ranges with and without imputation.

In regard to imputation methods, as mentioned earlier, we started with the pseudotest forms with complete data and then introduced missingness by dropping Item 6. We imputed the missing item scores using three different methods. We then evaluated the measures (reliability, accuracy of scaled scores, etc.) for the pseudoforms after imputation based on different imputation methods using information from the complete data as the criterion.

The three methods included in our study are (a) imputation by median, (b) imputation based on IRT, and (c) two-way imputation with error (abbreviated as TW^*_{ij} in van Ginkel et al., 2010).

The first method uses the median of the first five items as the imputed value for Item 6 because Items 1–6 have the same score ranges.[1]

For the IRT-based imputation method, the generalized partial credit model (GPCM) was fitted to each form (excluding scores on Item 6) and theta values for each test taker within each form were computed; then the estimated item parameters for Item 5 and the estimated theta values were used to simulate item responses on Item 6 for each form. The basis of this strategy of simulation is the fact that Item 6 (the new item) is of the same type and has similar item score mean and standard deviation as Item 5 (see Tables 1 and A1). The IRT parameter estimation[2] and response data generation were performed using the *mirt* package in R (Chalmers, 2012). For each form, the scores on Item 6 were simulated 20 times. The final imputed score for each test taker on Item 6 was the rounded average of Item 6 scores over 20 simulations. After we fit the GPCM to each form, we also checked item fit for Item 5. If the signed chi-squared test ($S\text{-}X^2$, Orlando & Thissen, 2000; Kang & Chen, 2008) is significant, then data from this form were discarded. The posterior mean of the examinee ability was used as the theta (ability) estimate; the use of other estimates such as maximum likelihood estimate or Warm's likelihood estimate did not change the conclusions.

The two-way method (van Ginkel et al., 2010) imputes TW^*_{ij}, the score of test taker i on item j, based on a two-way analysis of variance (ANOVA) model. Assume that there is an item score matrix with N persons and J items. Let PM_i denote the mean of person I's available item scores, let IM_j denote the mean score of item j based on all scores available for this item, and let OM be the overall mean of all available item scores in the $N \times J$ data matrix. It is assumed that $TW^*_{ij} = PM_i + IM_j - OM + \varepsilon_{ij} = TW_{ij} + \varepsilon_{ij}$, where $TW_{ij} = PM_i + IM_j - OM$. To calculate TW_{ij}, we assumed that Item 6 (the new item) has the same difficulty level as Item 5, that is, $IM_6 = IM_5$. This is because Item 6 and Item 5 were of the same item type, and their difficulty levels across the 67 pseudoforms were very close in operation (Table A1 shows a summary of item score means and standard deviations across the 67 pseudoforms with real data). The error term ε_{ij} is drawn from a normal distribution with mean 0 and variance S^2_ε, which is the error variance in the observed data (or nonmissing data). The variance S^2_ε is calculated by $S^2_\varepsilon = \sum\sum\left(X_{ij} - TW_{ij}\right)^2/M$, where X_{ij} represents each observed (nonmissing) item score and M is the number of observed scores minus 1. For each combination of examinee and item score, TW_{ij} was calculated. The TW_{ij} scores are considered to be the expected scores of the ANOVA model. The error term S^2_ε quantifies how well the ANOVA model fits the data. Because the items in our test have different score ranges, we transformed the categorical scores into proportional scores for each item before applying the two-way method. After imputation, we transformed

proportional scores back to their original scales and used rounding and truncating to get the final imputed score for Item 6.

Note that although the first method used scores on Items 1 – 5 to impute the scores on Item 6, the IRT method and the two-way method mainly used data on Item 5 to impute scores on item 6.

The median method is a crude and simple method that does not take into account the difficulty of the items and is expected to perform the worst. The two-way method is also a simple method, but more sophisticated than the median method in roughly taking into account item difficulty, and it has performed well in existing studies (van Ginkel et al., 2010), so we expected the method to perform better than the median method. The IRT method has been found to perform well by Sinharay (2021) and Huisman and Molenaar (2001) and is based on IRT theory. If an IRT model fits the data, the IRT imputation method was expected to perform the best among the three methods in terms of scale score accuracy. We were not sure how each imputation method would perform in estimating test reliability and interitem correlations.

Results

During the imputation using the IRT method, three pseudoforms were excluded from our analyses because the GPCM model did not have a good fit on Item 5. Therefore, further analyses for the IRT imputation method were based on data from 64 pseudoforms ($N = 72{,}580$).

Table 2 shows the percentage distributions of scaled score changes (transformed into units of the *SD* of the scaled scores) with and without imputation using the three different methods over the 67 pseudoforms. Detailed information about scaled score changes can be found in Tables A2 – A4. All three methods had a high percentage of test takers with 0 scaled score change. The imputation by IRT method had the highest percentage of test takers with 0 scaled score change. The two-way method had the lowest percentage compared to the other two methods. Almost all test takers had a scaled score change within .42 *SD* for all three methods.

Table 3 shows the standardized bias and standardized RMSD, each computed from the data combined over the 67 pseudoforms, for the three imputation methods. On average, the imputed scores from all three methods were slightly smaller than the actual scaled scores, but the bias was negligible for all three methods. Relatively, the median method had larger bias than the other two methods. The standardized RMSD was smaller than 0.2 for the three imputation methods, indicating that the absolute error in imputing scaled scores was small for all three methods.

Table 4 shows the reliability estimates (Cronbach's alpha), over the 67 pseudoforms, with and without imputations for Item 6. The last column is the reliability without imputation and computed based on 13 items. The two-way method provided the most similar reliability estimate as the actual test reliability estimate. The median and the IRT method both overestimated the test reliability.

Table 2 Percentage Distribution of Scaled Score Changes After Imputation

Scaled score change (in units of *SD*)	Median	IRT	Two-way
−0.83			0.01
−0.42	14.88	9.11	11.66
0	80.66	82.3	78.21
0.42	4.45	8.59	10.13
0.83	0	0	0
N	75,819	72,580	75,819

Abbreviation: IRT, item response theory.

Table 3 Average Scaled Score Changes Across 67 Pseudoforms After Imputation

Group	Median	IRT	Two-way
Standardized bias	−0.0434	−0.0022	−0.0064
Standardized RMSD	0.1833	0.1753	0.1946
N	75,819	72,580	75,819

Abbreviations: IRT, item response theory; RMSD, root mean squared difference.

Table 4 Comparison of Test Reliability With and Without Imputation

Comparison	With imputation			Without imputation
Coefficient alpha	Median	IRT	Two-way	Real data
Average	0.842	0.841	0.830	0.828
Minimum	0.781	0.778	0.761	0.747
Maximum	0.874	0.870	0.866	0.862

Abbreviation: IRT, item response theory.

Table 5 Average Interitem Correlations Over 67 Pseudoforms

Scores		Item 1	Item 2	Item 3	Item 4	Item 5	Item 6_Real
Item 5		0.242	0.262	0.245	0.261	1	0.308
Item 6_Actual		0.250	0.258	0.251	0.263	0.308	1
Item 6_Imputed	Median	0.653	0.666	0.649	0.659	0.423	0.302
Item 6_Imputed	IRT	0.485	0.503	0.482	0.500	0.367	0.304
Item 6_Imputed	Two-way	0.289	0.291	0.286	0.291	0.258	0.211

Table 5 presents some average interitem correlations with and without imputation. Here we only focused on the correlations between Items 5 and 6 and other items with same score categories. Table 5 shows that the correlations of Item 5 with Items 1 – 6 ranged between 0.242 and 0.308. The correlations between the imputed Item 6 scores and actual scores on Items 1 – 5 based on the two-way method were very similar to the correlations between actual scores of Item 6 (or Item 5) and actual scores on other items. The correlations between Item 6 scores imputed by the median and the IRT methods with Items 1 – 5 were larger (and ranged between 0.367 and 0.659) than the correlations between the actual Item 6 scores and Items 1 – 5 (that ranged between 0.250 and 0.308).

Conclusions and Practical Implications

We compared three data imputation methods in terms of the accuracy in estimating scaled score and reliability of test score. Starting with empirical data from 67 pseudoforms that did not include any missing item scores, we first assumed that the scores on one item (i.e., Item 6) were missing for all (100%) test takers and then imputed those scores using the median method, IRT method, and two-way method. Scaled scores and score reliability after imputation were compared against those based on actual data.

Our results indicate that all three methods performed well in imputing scaled scores when scores on only one item are missing for 100% examinees. The imputation error for each method was found to be small. The IRT method worked best in providing accurate imputed scaled scores with negligible bias and a small RMSD. The median method performed worse than the IRT method. Scaled scores imputed by the median method had larger bias and a larger RMSD than for the IRT method. The two-way method also had slightly larger bias and RMSD than the IRT method, but it provided the most accurate reliability estimate. Item scores imputed using the IRT method and the median method had much higher correlations with the other items, leading the reliability estimate based on imputed scores to be higher than the actual reliability.

The results are consistent with our expectations about the accuracy of the three imputation methods in terms of reporting scaled scores. The IRT imputation method provided the most accurate scaled scores. However, item scores imputed by the IRT method did not have reasonable correlations with other items. In addition, test reliability estimate based on scores imputed by the IRT method was not the most accurate.

For the given situation, that is, when scores were missing completely for all test takers only on Item 6, we did not find a method that was uniformly better than the other methods in all the investigated aspects. Our results suggest that when choosing an imputation method to use in operation, the investigator should choose a method according to the purpose. If the purpose is to evaluate test reliability before administering a redesigned test with new items added, then one can simulate scores on the newly added items using the two-way imputation method. If one is going to report missing scores in operation, one may use the IRT method if IRT model fit is not a big concern.

It is likely that the performance of the imputation methods will depend on test content specifications, and the portion of the test with missing scores. The test in our study had 13 constructed-response (CR) items with different item types

and different score categories. If we had used data from a test with all multiple-choice items, or CR items with same score categories, we might have obtained different results. In addition, in this study, we only considered the case when scores on only one item were completely missing. So, we only imputed scores on one item based on scores on the other 12 items. All three imputation methods performed quite well. If scores on more items had to be imputed, then these three imputation methods might not have worked as well as in the current case, and they might work more differently from each other.

Note that the type of missing data that we consider in this paper is a special case of data that are missing completely at random because the probability of a missing score of an examinee on an item is 1, which means that the probability does not depend on missing or observed data (where the data are item scores). Therefore, approaches that are based on the modeling of missing not at random data (Holman & Glas, 2005) were not employed.

We plan to conduct further studies to investigate the impact of test characteristics and the amount of missing information on the performance of these three and potentially other imputation methods. In this study, we estimated score reliability using only coefficient alpha. We plan to evaluate the impact of imputation methods on other types of score reliability estimates in the future.

Notes

1 Another strategy would be to impute Item 6 score by Item 5 score, but such a strategy was deemed not good for estimating reliability—so median of the first five item scores was used to impute the Item 6 score in this method.
2 Note that the various forms do not have any items in common—so no attempt was made to equate the forms.

References

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

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), 1–17. https://doi.org/10.1111/j.2044-8317.2005.tb00312.x

Huisman, M., & Molenaar, I. W. (2001). Imputation of missing scale data with item response models. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 221–244). Springer. https://doi.org/10.1007/978-1-4613-0169-1_13

Kang, T., & Chen, Troy T. (2008). An investigation of the performance of the generalized S-X^2 item-fit index for polytomous IRT models. *Journal of Educational Measurement*, *45*(4), 391–406. https://doi.org/10.1111/j.1745-3984.2008.00071.x

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50–64. https://doi.org/10.1177/01466216000241003

Sinharay, S. (2021). Score reporting for examinees with incomplete data on large-scale educational assessments. *Educational Measurement: Issues and Practice*, *40*(1), 79–91. https://doi.org/10.1111/emip.12396

Sinharay, S. (2022). Reporting proficiency levels for examinees with incomplete data. *Journal of Educational and Behavioral Statistics*, *47*(3), 263–296. https://doi.org/10.3102/10769986211051379

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*(1), 17–30. https://doi.org/10.1027/1614-2241/a000003

Appendix

Table A1 Summary of Item Score Means and Standard Deviations (*SD*)

Statistic		Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12	Item 13
Mean	Avg.	2.26	2.24	2.27	2.23	2.48	2.49	2.53	2.46	2.32	2.60	2.39	2.44	2.73
	Min.	2.07	2.05	2.13	2.06	2.28	2.27	2.01	1.93	1.94	2.22	1.60	2.14	2.11
	Max.	2.43	2.41	2.51	2.39	2.65	2.69	2.82	2.85	2.60	2.85	2.81	2.69	2.98
SD	Avg.	0.51	0.51	0.52	0.51	0.53	0.53	0.57	0.57	0.56	0.57	0.71	0.56	0.63
	Min.	0.44	0.46	0.46	0.38	0.49	0.48	0.41	0.39	0.47	0.39	0.44	0.46	0.53
	Max.	0.56	0.55	0.56	0.57	0.57	0.56	0.81	0.83	0.70	0.85	0.94	0.82	0.87

Table A2 Percentage of Scaled Score Changes (in Units of *SD*) at Each Scaled Score Level—the Median Method ($N = 75{,}819$)

Scaled score change	Scaled score levels																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
−0.42			0.0	0.0	0.0	0.0	0.1	0.1	0.2	0.3	0.5	1.2	2.4	3.5	3.9	2.1	0.2	0.1	0.0		
0	0.0	0.0	0.0	0.1	0.1	0.2	0.3	1.0	1.4	3.1	5.9	14.6	11.5	12.4	10.1	7.9	6.3	2.7	2.2	0.3	0.4
0.42		0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.2	0.5	0.7	1.1	0.9	0.3	0.1	0.0	0.0	
0.83		0.0							0.0							0.0					

Table A3 Percentage of Scaled Score Changes (in Units of *SD*) at Each Scaled Score Level—the Item Response Theory Method ($N = 72{,}580$)

Scaled score change	Scaled score levels																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
−0.42			0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.4	1.1	2.0	2.5	2.0	0.7	0.1	0.0	0.0		
0	0.0	0.0	0.0	0.1	0.1	0.2	0.3	1.1	1.5	3.2	5.9	14.0	10.8	11.8	11.5	9.4	6.5	2.9	2.3	0.3	0.4
0.42		0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.2	0.7	1.5	2.3	1.9	0.9	0.3	0.1	0.0	0.0	
0.83																0.0					

Table A4 Percentage of Scaled Score Changes (in Units of *SD*) at Each Scaled Score Level—the Two-Way Method ($N = 75{,}819$)

Scaled score change	Scaled score levels																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
−0.83										0.0		0.0									
−0.42			0.0	0.0	0.0	0.0	0.1	0.1	0.2	0.4	0.7	1.3	1.8	2.2	2.1	1.5	0.8	0.3	0.1	0.0	0.0
0	0.0	0.0	0.0	0.1	0.1	0.2	0.3	1.1	1.3	2.7	5.1	13.2	10.5	12.2	11.6	8.7	5.8	2.6	2.1	0.3	0.4
0.42	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.4	0.8	1.5	2.1	2.2	1.5	0.7	0.3	0.1	0.0	0.0	

Suggested citation:

Qu, Y., & Sinharay, S. (2023). *Methods for imputing scores when all responses are missing for one or more polytomous items: Accuracy and impact on psychometric property* (Research Report No. RR-23-07). ETS. https://doi.org/10.1002/ets2.12369

Action Editor: Gautam Puhan

Reviewers: Hongwen Guo and Jonathan Weeks