Investigation of Missing Responses in Q-Matrix Validation

Applied Psychological Measurement
2018, Vol. 42(8) 660–676
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0146621618762742
journals.sagepub.com/home/apm

Shenghai Dai¹ , Dubravka Svetina² and Cong Chen³

Abstract

Missing data can be a serious issue for practitioners and researchers who are tasked with Q-matrix validation analysis in implementation of cognitive diagnostic models. The article investigates the impact of missing responses, and four common approaches (treat as incorrect, logistic regression, listwise deletion, and expectation–maximization [EM] imputation) for dealing with them, on the performance of two major Q-matrix validation methods (the EM-based δ -method and the nonparametric Q-matrix refinement method) across multiple factors. Results of the simulation study show that both validation methods perform better when missing responses are imputed using EM imputation or logistic regression instead of being treated as incorrect and using listwise deletion. The nonparametric Q-matrix validation method outperforms the EM-based δ -method in most conditions. Higher missing rates yield poorer performance of both methods. Number of attributes and items have an impact on performance of both methods as well. Results of a real data example are also discussed in the study.

Keywords

missing responses, Q-matrix validation, Q-matrix misspecification, missing imputation, cognitive diagnostic models

Missing responses and how to treat them often present a common problem in the context of measurement. For decades, research had been conducted to study the impact of missing responses and a variety of approaches had been developed to deal with this problem, most of which were within the framework of item response theory (IRT). However, fewer studies were conducted so far to examine the phenomenon of missingness in implementation of cognitive diagnostic models (CDMs) and Q-matrix validation procedures.

Recently, CDMs have gained a lot of attention because of their advantage for providing students with fine-grained diagnostic information about students' mastery of measured attributes (DiBello, Roussos, & Stout, 2007; Leighton & Gierl, 2007; Xu & von Davier, 2006). In the implementation of a CDM, a Q-matrix is constructed to identify the interaction between cognitive attributes and test items. It is defined as a matrix having rows as items and columns as

¹Washington State University, Pullman, USA

²Indiana University, Bloomington, USA

³University of Illinois at Urbana–Champaign, USA

Corresponding Author:

Shenghai Dai, Educational Psychology Program, Washington State University, Cleveland Hall 354, P.O. Box 642136, Pullman, WA 99164-2136, USA.

Email: s.dai@wsu.edu

attributes, where “1” in a cell indicates mastery of a certain attribute being required to respond to that item correctly, and “0” otherwise (Henson & Douglas, 2005). The Q-matrix is the key factor in determining the quality of cognitive diagnostic information (DeCarlo, 2012; Jang, 2009). If a Q-matrix is misspecified, diagnostic classification based on it as well as other parameter estimates may lead to inaccurate inference. In light of this, researchers who have been working on CDMs began to recognize the consequences of Q-matrix misspecification and the potential of Q-matrix validation (Chiu, 2013; de la Torre, 2008; de la Torre & Chiu, 2015; Rupp & Templin, 2008). Currently, in implementation of the CDMs, most literature had used existing assessments and retrofit Q-matrix in the hope of extracting more information not captured by traditional approaches (DiBello et al., 2007; Jang, 2009). The presence of missing responses in these settings is usually inevitable and should not be ignored (Rose, von Davier, & Xu, 2010).

Missingness can be a serious issue for practitioners and researchers who are tasked with Q-matrix validation in the framework of CDMs. To the authors’ knowledge, no study to date examined missingness in the context of CDMs, especially in implementation of Q-matrix validation procedures. To address this concern, the current article investigates the impact of missing responses, and four common approaches for dealing with it, on performance of two major Q-matrix validation methods in the context of CDMs.

Background

Overview of Missingness in Educational Measurement

Rubin’s (1976) three mechanisms are commonly used to distinct different types of missingness. According to Rubin, missing responses can be missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). If the missingness does not have any systematic cause nor depend on any other information, then the data are considered MCAR and ignorable (Finch, 2008; Schafer & Graham, 2002). When data are MAR, the probability of an item being missing depends on some measurable characteristic of the individual but not on the item itself (Cheema, 2014; Finch, 2008; Little & Rubin, 1989). When data are MNAR, the probability of a missing response on an item only depends on that item itself.

In large-scale assessments, it is typical to utilize what is known as the balanced-incomplete-block (BIB) design, within which examinees are administered only a subset (e.g., booklet) of items. Missing responses to items that are not administered to examinees are considered missing-by-design. For items that are presented to students, missing responses are categorized into omitted or not-reached responses depending on whether missing data occur subsequent to the last question the student answered or not. In practice, researchers who work with large-scale assessment data typically utilize booklet-level response data (e.g., de la Torre, 2008; Lee, Park, & Taylan, 2011), rendering missing-by-design data irrelevant. In such applications, when encountered, missingness refers to either omitted or not-reached responses. In the current study, the authors also approach missingness as an issue of either missing by omitting an item or not reached. Previous studies suggested that both types of missing responses were nonignorable and they may result in biased item and ability parameter estimates (De Ayala, Plake, & Impara, 2001; Lord, 1974, 1980; Mislevy & Wu, 1988, 1996).

Researchers identified several reasons for item responses to be missing. Some scholars suggested that missing responses can be assumed MAR because they are related to certain individual characteristics. For example, Lord (1980) suggested that if an examinee knew how ability was estimated in IRT, he or she could obtain a high(er) ability estimate simply by answering only items that he or she had confidence in answering correctly. Finch (2008), on the contrary,

held the perspective that the probability of missingness was related inversely to students' proficiency. Some other researchers suggested that the likelihood of an item to be missing was more likely dependent on that item itself, in which case missingness could be MNAR. In examining variation in missing responses on National Assessment of Educational Progress (NAEP), Brown, Dai, and Svetina (2014) showed that nearly all predictors explained negligible amounts of variance in omitted scores, with an exception of item format, which accounted for 21.3% to 52.6% of the variance of omitted responses in the 2009 NAEP math data.

In the context of measurement, a large number of different options for dealing with missingness have been proposed in the literature (see Finch, 2008; Peugh & Enders, 2004; and Schafer & Graham, 2002, for a more comprehensive review). Standard procedures in the majority of large-scale assessments, such as NAEP, are to replace omitted responses with incorrect scores or fractionally correct at the value of the reciprocal of the number of response alternatives and to treat not-reached responses as not-presented data (Rogers & Stoeckel, 2011). Other procedures, such as the expectation-maximization (EM) algorithm (Little & Rubin, 2002), response function (RF) imputation (Sijtsma & Van der Ark, 2003), and logistic regression (LR) imputation (Finch, 2010), have been proposed and applied in assessment settings as well. These missing data methods have been evaluated using both simulated and real assessment data in the framework of IRT (e.g., Brown, Svetina & Dai, 2014; De Ayala et al., 2001; Finch, 2008). In the context of CDMs and Q-matrix validation, to the authors' knowledge, no study has been conducted to investigate the impact of missing responses. In the limited research on CDMs where omitted responses were present, researchers usually treated omissions as incorrect (e.g., Jang, 2009; Lee et al., 2011).

Q-Matrix Validation Methods

Research on Q-matrix misspecification suggests that validating a Q-matrix can be accomplished using several methods. Methods addressed in literature include parametric sequential EM-based δ -method (de la Torre, 2008), nonparametric Q-matrix refinement method (Chiu, 2013), Bayesian approach (DeCarlo, 2012), and data-driven approach (Liu, Xu, & Ying, 2012). In the current study, the authors focused on the two approaches, proposed by de la Torre and Chiu, respectively, primarily for the following three reasons. First, both methods aim to refine an existing Q-matrix in which elements are predefined. Second, both methods were introduced in the context of the deterministic inputs, noisy "and" gate (DINA) model (Junker & Sijtsma, 2001), one of the most commonly used CDMs for Q-matrix validation, making comparisons feasible and reasonable. Third, both methods showed promise in studied conditions as potential approaches in Q-matrix validation (Chiu, 2013; de la Torre, 2008).

The EM-based δ method. De la Torre (2008) proposed the sequential EM-based δ method to validate a Q-matrix used in the DINA model analysis. It is a parametric method for it relies on the estimation of item parameters from the DINA model. The parameter δ is denoted as the difference in the probabilities of a correct response to a specific item between examinees who possess all the required skills and examinees who do not. δ can be thought of as an item discrimination index for which items with higher δ values are assumed to differentiate between examinees more effectively than those with low δ values. According to de la Torre's rationale, for the DINA model, the q-vector for a specific item is considered to be the correct q-vector if it maximizes the item discrimination index. To search for a valid Q-matrix that maximizes δ more efficiently, a sequential search algorithm is used. Results from both simulated and real data showed that the EM-based method was able to detect misspecified q-vectors and retain those that were correctly specified (see de la Torre, 2008, for more details).

The nonparametric Q-matrix refinement method. Chiu (2013) developed the nonparametric Q-matrix refinement method for identifying and correcting the misspecified q-entries of a Q-matrix. Using the weighted Hamming distance, Chiu's method operates by minimizing the residual sum of squares (RSS) computed from the observed response and the ideal responses to each test item. The "nonparametric" aspect of her method is reflected in how it relies on the estimation of attribute profiles using the nonparametric classification method (Chiu & Douglas, 2013). It uses RSS as a loss function to identify the misspecification by assuming that when all examinees are correctly classified, the correct q-vector for a specific item is expected to have the lowest RSS among all alternative q-vectors. Thus, a correct Q-matrix is assumed to minimize the overall RSS of the test given the independence between the RSS of each item. Chiu evaluated performance of this method with simulation studies and a real data example. Results showed that it was able to detect and recover misspecified q-entries efficiently and effectively (see Chiu, 2013, for more details).

The Simulation Study

Studied Factors

To investigate the impact of missing responses on performance of two Q-validation methods, a simulation study was conducted using following manipulated factors and their levels.

Mechanism and rate of missingness. Two missing data mechanisms, MAR and MNAR, were considered in this study. Missing rate (MR) was set at three levels (10%, 20%, and 30%) based on the literature in missing data (Finch, 2008). Estimates of both validation methods for complete data (i.e., MR = 0%) were used as the baseline for comparison.

Missing imputation approaches. Four imputation approaches were used: EM, treating missingness as incorrect (IN), LR, and listwise deletion (LW). The EM imputation was selected given its good performance in the literature (Finch, 2008; Peugh & Enders, 2004). IN was considered because it is one of the most common approaches when missing with missing data in implementation of CDMs. This method assumes that the student did not respond to an item because he or she made an accurate appraisal of his or her ability to answer the item correctly, and concluded he or she would get the item incorrect. LR imputation was chosen because item response data in measurement are usually dichotomous (e.g., representing correct and incorrect responses). The LW was used as the baseline of comparison.

Q-matrix structure (number of attributes and items). Four Q-matrices that correspond to tests of 20 or 40 items ($J = 20, 40$), and three or five attributes ($K = 3, 5$), respectively, were simulated. Q-matrices that corresponded to tests of 20 items were the same as those used in Chiu (2013, p. 608). For tests with 40 items, the corresponding Q-matrices were generated by doubling the Q-matrix of 20 items. All four Q-matrices were treated as correctly specified.

Q-matrix misspecification rate. Misspecification rates of Q-matrix entries were set at 5%, 10%, and 20%, respectively. To generate Q-matrices with different misspecification rates, q-entries in the correct Q-matrix were selected at a specific probability and then replaced with opposite values (0's were replaced as 1's and vice versa). In the current study, 12 misspecified Q-matrices were generated ($2 [\text{item}] \times 2 [\text{attribute}] \times 3 [\text{misspecification rate}]$), alongside the four correctly specified ones.

Sample size. Sample sizes of 500, 1,000, and 2,000 were used for each condition.

In addition to manipulated factors stated above, the source of examinees' attribute patterns and item parameters (slipping and guessing) were also required to simulate data sets. In this study, both item parameters were generated from a uniform distribution $U(0, 0.2)$. Examinees' attribute patterns were generated from a multivariate normal distribution.

A fully crossed design yielded a total number of 1,800 conditions, 72 of which used the complete data while the other 1,728 used the imputed data, corresponding to the six manipulated factors stated above. Each condition was replicated 100 times.

Data Generation

The DINA model was first used to simulate complete dichotomous responses using information specified above. To generate missing responses of MAR, a continuous hypothetical variable was generated from a standard normal distribution and used as the proxy that was assumed to be inversely related with the probability of missing data. To generate data of MNAR, the probability of reporting a missing response was directly related to whether the simulee got the item correct or not in the original complete data (De Ayala et al., 2001; Finch, 2008).

Analysis

Analyses were conducted using R software (R Core Team, 2017) and SAS version 9.4. The LR imputation was conducted using R-mice package (van Buuren & Groothuis-Oudshoorn, 2015), and EM imputation was conducted using SAS PROC MI. Q-matrix validation procedures were implemented using R Package for cognitive diagnosis model (R-CDM; Robitzsch, Kiefer, George, & Uenlue, 2017) and R Package for Nonparametric Methods for Cognitive Diagnosis (R-NPCD; Zheng & Chiu, 2016) packages, respectively. Results were evaluated using two criteria: (a) the proportion of misspecified q-entries (mean recovery rate [MRR]) and (b) the mean misrecovery rate (MMR). MRR is operationalized as the mean proportion of misspecified q-entries that were identified and corrected using the validation method. MMR refers to the mean proportion of correct q-entries in the original Q-matrix that are wrongly identified in the validation procedure. One-way ANOVA was used to examine main effects of manipulated factors. In addition, to further investigate the impact of missing data as well as the way they are treated in Q-matrix validation procedures, the authors also examined results of both attribute and item parameter estimations that the validation procedures rely on. The authors noted that the EM-based δ method selects the optimal Q-matrix that maximizes the item discrimination index, whereas the nonparametric procedure refines the Q-matrix based on the distance between observed and ideal responses using the nonparametric classification method. Therefore, the authors used the root mean squared error (RMSE) to examine the estimation of item discrimination indices of the DINA model for the EM-based method, and the attribute-wise classification accuracy (ACA) index to evaluate the attribute classification results for the nonparametric procedure, respectively.

Results

ANOVA results suggested the impact of design factors on both MRRs and MMRs shared similar patterns across levels sample sizes and test lengths with only trivial to small effect sizes (partial $\eta^2 < .06$), with the methods performed better in conditions with larger sample sizes and more items. Therefore, the authors report results associated with $N = 2,000$ and $J = 40$, only. The accompanied, complete tabulated results are available upon request.

MRR results. Figure 1 presents the MRRs of both Q-validation methods across factors when data were complete and MAR. The figure contains six graphs (two columns and three rows) with columns 1 and 2 representing conditions with $K = 3$ and 5 while rows A, B, and C representing different degrees of Q-matrix misspecification. In particular, within each graph, dotted and solid lines represent results associated with the EM-based method and the nonparametric procedure, respectively; lines with square, circle, triangle, and asterisks marks denoted results associated with EM, IN, LR, and LW imputation methods.

Results indicated that when data were complete, both validation methods performed equally robust when there were three attributes, with nearly all MRRs being 100%. When there were five attributes, however, MRRs of both methods decreased, especially in the presence of higher Q-matrix misspecification rates. In addition, when $K = 5$, MRRs of the nonparametric procedure were always larger than those of the EM-based method.

When data were MAR, results showed that the EM imputation outperformed the other three imputation methods across all conditions for both validation procedures, as indicated by higher lines with square marks. In addition, MRRs of both validation methods only decreased slightly as the missing rate increased when EM imputed data were used. The LR imputation performed slightly worse than EM, especially in conditions with three attributes and a low rate of Q-matrix misspecification. The difference between the two imputation methods, however, became more prominent in the presence of a larger number of attributes and higher misspecification rate. When missing responses were treated as IN, results showed that higher missing rates yielded decreased MRRs across most of the conditions for both validation methods, especially in conditions of five attributes. The LW imputation performed noticeable worse than the other methods in nearly all conditions, as indicated by lower lines with asterisks in all graphs.

Based on results of MRRs in Figure 1, the authors noted that the nonparametric method performed superior to the EM-based method in almost all conditions, especially when there were five attributes, which was visualized by higher solid lines (as opposed to dotted lines) in the figure. Furthermore, performances of both validation methods using EM and LR imputed data (lines with square and triangle marks) were the most robust among all conditions with highest average MRRs. In addition, as it can be seen from the figure that higher missing rates yielded lower MRRs, regardless of the misspecification rate.

Figure 2 reports the results when data were MNAR. Similar patterns were detected to those in Figure 1. For each of the validation methods, EM imputation yielded higher MRRs than the other three imputation procedures did. Differences of MRRs across missing rates were small when missing data were imputed using EM. LR performed similarly to EM in most conditions but slightly worse when there were larger number of attributes and higher percent of Q-matrix misspecification. This finding seemed reasonable because the data were simulated in such a way that incorrect responses had a higher probability to be missing. However, IN still performed worse than EM and LR. When missing responses were treated using LW, both validation methods yielded lower MRRs than they did in comparable conditions when data were assumed MAR. LW performed the worst among selected methods.

In addition, similar conclusions with respect to the performance of both Q-matrix validation procedures were found. As noted in Figures 1 and 2, the presence of larger number of attributes in a Q-matrix, higher misspecification, and missing rates resulted in larger MRRs. Furthermore, the authors noted that the nonparametric method performed superior to the EM-based method in almost all conditions, especially when there were five attributes.

Table 1 shows patterns based on ANOVA to visualize the change of MRRs. All design factors showed statistically significant effect on MRRs for both validation procedures with only several exceptions. In the presence of missing responses, the impact of missing rates over MRRs were not significant for both validation procedures when missing data were imputed

Figure 1. Average MRRs of Q-validation methods and missing data imputation methods when MAR. Note. MRR = mean recovery rate; MAR = missing at random; EM = expectation-maximization imputation; IN = treating missingness as incorrect; LR = logistic regression imputation; LW = listwise deletion; QM = Q-matrix misspecification rate; K = number of attributes.

Figure 2. Average MRRs of Q-validation methods and missing data imputation methods when missing not at random.
Note. MRR = mean recovery rate; EM = expectation-maximization imputation; IN = treating missingness as incorrect; LR = logistic regression imputation; LW = listwise deletion; QM = Q-matrix misspecification rate; K = number of attributes.

Table 1. Change of MRRs of Both Parametric and Nonparametric Q-Matrix Validation Methods.

Studied factors	Missing treatment	EM-based δ method			Nonparametric method		
		No missing	MAR	MNAR	No missing	MAR	MNAR
MR $\uparrow$	EM	/	▼	▼	/	▼	▼
	IN		▼***	▼***		▼***	▼***
	LR		▼***	▼**		▼**	▼*
	LW		▼***	▼***		▼***	▼***
QM $\uparrow$	EM	▼***	▼***	▼**	▼**	▼**	▼**
	IN		▼***	▼***		▼**	▼***
	LR		▼***	▼***		▼***	▼***
	LW		▼***	▼**		▼*	▼
K $\uparrow$	EM	▼***	▼***	▼***	▼***	▼***	▼***
	IN		▼***	▼***		▼***	▼***
	LR		▼***	▼***		▼***	▼***
	LW		▼***	▼**		▼***	▼***
J $\uparrow$	EM	▲***	▲***	▲***	▲**	▲***	▲**
	IN		▲***	▲***		▲	▲*
	LR		▲***	▲***		▲***	▲***
	LW		▲	▲		▲***	▲***
N $\uparrow$	EM	▲***	▲**	▲***	▲***	▲***	▲***
	IN		▲	▲		▲*	▲**
	LR		▲**	▲*		▲***	▲***
	LW		▲*	▲		▲**	▲*

Note. MRRs = mean recovery rates; EM = expectation–maximization imputation; MAR = missing at random; MNAR = missing not at random; $\uparrow$ = increase in magnitude; MR = missing rate; IN = treating missing responses as incorrect; LR = logistic regression missing imputation; LW = listwise deletion; ▲ = increase in MRRs; ▼ = decrease in MRRs; QM = Q-matrix misspecification rate; K = number of attributes; J = number of items; N = sample size.

* $p < .05$. ** $p < .01$. *** $p < .001$.

using EM, regardless of the missing mechanisms. When missing data were imputed using the other three methods, however, the effect of the missing rate was significant. The effect of sample size was not significant when the EM-based method was utilized and missingness was treated using IN.

MMR results. Figure 3 presents patterns of average MMRs for the two validation methods. When examining MMRs, the authors noted similar patterns for each Q-matrix validation procedure when missing data were treated using EM, IN, and LR. Regardless of other factors, both validation procedures yielded almost perfect MMRs (i.e., 0%) when there were only three attributes. However, MMR values increased rapidly when a Q-matrix had larger number of attributes and higher misspecification. Furthermore, the EM-based validation procedure resulted in higher MMRs than the nonparametric procedure did. When EM, LR, and IN were applied to missing data, the nonparametric procedure performed excellently with MMRs of nearly all conditions being close to zero, as suggested by solid lines with square, triangle, and circle marks that were overlapping and at zero. The EM-based procedure performed differently across conditions. In similar conditions, the nonparametric procedure performed superior to the EM-based procedure, especially in the presence of higher misspecification, as indicated by higher solid lines than dotted lines. The LW resulted in noticeable larger MMRs than other imputation methods, especially when there are a larger number of attributes and higher missing rates.

Figure 3. Average MMRs of Q-validation methods and missing data imputation methods for both MAR and MNAR.

Note. MMRs = mean misrecovery rates; MAR = missing at random; MNAR = missing not at random; EM = expectation-maximization imputation; IN = treating missingness as incorrect; LR = logistic regression imputation; LW = listwise deletion; QM = Q-matrix misspecification rate; K = number of attributes.

Impact of missingness on attribute and item parameter estimations. To further investigate the impact of missing data as well as the way they are treated in Q-matrix validation procedures, the authors also examined results of both attribute and item parameter estimations that the validation procedures rely on. Figure 4 shows results of attribute estimates of the nonparametric procedure when data were MAR in terms of ACAs. The authors noted both similarities and discrepancies when comparing the performance of selected missing data imputation methods in estimating attribute profiles to that in validating Q-matrices. On one hand, EM and LR shared similar patterns. They performed similarly and both were superior to the other two methods with LR performed slightly worse than EM. On the other hand, both IN and LW showed different patterns across contexts. The performance of IN was more easily impacted by the missing rates when there were three attributes in attribute estimations than it did in Q-matrix validation, especially when $MR \leq 20\%$. Discrepancy of the performance of LW in two contexts was visually noticeable, too. LW performed at an acceptable level in estimating attribute profiles but not in validating Q-matrices. Similar patterns were detected when data were MNAR.

Figure 5 presents the RMSEs of estimated item discrimination indices from the DINA model when data were MAR. The authors note both similarities and discrepancies when comparing the performance of selected missing data imputation methods between tasks of item parameter estimation and Q-matrix validation. With respect to the accuracy of the estimation of item discrimination, EM and LR performed almost identical and both are equally robust across all conditions. The performance of LR, however, was more easily impacted by higher misspecification and larger number of attributes. Different patterns of the results of IN and LW across contexts were detected, too. Similar patterns were detected when data were MNAR.

Real Data Example

Real data example involved the fraction-subtraction data collected and analyzed by Tatsuoaka (1990). This data set has been widely used and studied in implementation of CDMs (Chiu, 2013; de la Torre, 2008; de la Torre & Douglas, 2004; Mislevy, 1996; Tatsuoaka, 2002). The data consisted of dichotomous responses of 536 middle school students to 15 items that measured five attributes (see Table 4). Details about the Q-matrix can be found in de la Torre (2008). To examine performance of both validation methods, in addition to the original (complete) item response, data sets with different mechanisms of missingness (i.e., MAR and MNAR) and missing rate (i.e., 10%, 20%, and 30%) were generated using the same procedures in the simulation study. EM, IN, and LR imputation were used to impute missing responses. LW was not considered in the real data example given its poor performance in the simulation study.

Table 2 presents modified q-entries that were proposed by both validation methods. When there was no missingness, the EM-based method suggested that the current Q-matrix was reasonable and should be retained by detecting zero q-entries that were misspecified (as shown by “/” in the table). This result is aligned with de la Torre (2008). The nonparametric method proposed that there were three q-entries that were associated with the second and fourth attributes in the Q-matrix that needed to be reconsidered. Result showed that the proposed q-entries to be modified were different in different missing rates and when missing responses were imputed differently. The EM-based method proposed 1 (q_{122}), 1 (q_{93}), and 2 (q_{42} , and q_{132}) q-entries that might be misspecified with IN, LR, and EM imputation, respectively, when data were of MAR, while the nonparametric proposed 6, 5, and 6 misspecified q-entries, respectively. Among these proposed q-entries, only three q-entries (q_{42} , q_{122} , and q_{132}) were suggested by both methods. In MNAR, the EM-based method proposed only one misspecified q-entry (q_{122}), whereas the nonparametric method proposed 11 across three missing treatment methods.

Figure 4. ACA rates for the nonparametric Q-matrix refine when missing at random.
Note. Only a limited number of replications returned the results for LW when MR = 30% due to lack of a sufficient sample size. ACA = attribute classification accuracy; EM = expectation-maximization imputation; IN = treating missingness as incorrect; LR = logistic regression imputation; LW = listwise deletion; MR = missing rate; QM = Q-matrix misspecification rate; K = number of attributes.

Figure 5. RMSE of the item discrimination index for the EM-based method when missing at random. Note. RMSE = root mean squared error; EM = expectation-maximization imputation; IN = treating missingness as incorrect; LR = logistic regression imputation; LW = listwise deletion; QM = Q-matrix misspecification rate; K = number of attributes.

Table 2. Modified Q-Entries of Fraction-Subtraction Data Suggested by Two Q-Matrix Validation Methods.

Treatment & rate of missingness	Q-validation methods	Missing mechanism		
		MAR	MAR	MNAR
No missingness	EM- δ	/		
IN	Nonparametric	q42, q122, q14		
	EM- δ	/		/
	Nonparametric	q42, q122, q14		q42, q122, q13, q14
	EM- δ	/		/
LR	Nonparametric	q132		q13, q14
	EM- δ	q122		q122
	Nonparametric	q62, q72		q72, q152, q13
	EM- δ	/		/
	Nonparametric	q42, q122, q14		q132, q13
	EM- δ	/		/
	Nonparametric	q42, q122, q13, q14		q22, q42, q122, q14
	EM- δ	q93		/
EM	Nonparametric	q22, q42, q122, q14		q12, q42, q82, q122, q132, q23
	EM- δ	/		/
	Nonparametric	q22, q122, q14		q42, q13, q132
	EM- δ	q132		/
	Nonparametric	q22		q22, q42, q122, q13
	EM- δ	q42, q132		/
	Nonparametric	q22, q42, q122, q13, q132		q12, q13, q14

Note. “/” = no q-entries being misspecified; q₄₂ = q-entry that represents the interaction between the fourth item and the second attribute; q₁₂₂ = q-entry of the 12th item and the second attribute, and so on. Five attributes associated with the Q-matrix are (a) performing basic fraction-subtraction operation, (b) simplifying/reducing, (c) separating whole number from fraction, (d) borrowing one from whole number to fraction, and (e) converting whole number to fraction. MAR = missing at random; MNAR = missing not at random; EM = expectation-maximization imputation; IN = treating missing responses as incorrect; LR = logistic regression missing imputation.

Summary and Conclusion

A misspecified Q-matrix may result in high misclassification rates and problematic diagnostic inferences in implementation of CDMs. Hence, there is an increasing need for Q-matrix validation approach that can reduce both the resource cost and the subjectivity of judgment in the Q-matrix development. Several methods, such as the EM-based δ and the nonparametric method, have demonstrated the potential for obtaining an adequate Q-matrix. To the authors' knowledge, however, their performances were not examined thus far in the literature in the presence of missing responses. The aim of the study was to compare performance of two validation methods in Q-matrix validation when missingness occurred.

The real data example showed that the selection of methods to handle missing data may have an impact on the accuracy in Q-matrix validation procedures. Results of the simulation study revealed that the nonparametric validation method outperformed the EM-based method in nearly all conditions in terms of both MRRs and MMRs. Both EM and LR imputation methods were superior to IN and LW in dealing with missing responses. EM imputation yielded slightly better results than LR in nearly all conditions and it was more stable than other methods in the presence of various rates of missing data (this was especially true for the EM-based validation method). Higher missing rates yielded poorer performance of both methods especially when data were treated using IN and LW. Number of attributes and items had an impact on performance of both methods as well. In most conditions, the EM-based method performed worse when there were larger number of attributes and higher misspecification in a Q-matrix.

Although Q-matrix validation procedures rely on estimations of either item parameters from a CDM (e.g., the EM-based method) or attribute profiles (e.g., the nonparametric refinement method), the authors noted both similarities and discrepancies of the impact of missing responses between tasks of Q-matrix validation and parameter estimation. Similarities lied in the performance of EM and LR, and they both performed robust across contexts. Both IN and LW showed different patterns in different tasks. For example, the performance of LW was acceptable when estimating students' attribute profiles using the nonparametric classification methods but not robust when validating a Q-matrix using the nonparametric validation procedure.

This study represents an initial step in examining performance of Q-matrix validation approaches under the presence of missingness. The authors recognize that as with any simulation study, generalizability is bound by the context of the study design and that limitations exist. They discuss limitations within the context of their study as well as point to potential future directions. First, the study was conducted in the context of the DINA model. To thoroughly examine the performance of Q-matrix validation methods, a method that can be implemented in a more generalized form of a CDM, such as the log-linear cognitive diagnosis model (LCDM), should be considered. For example, de la Torre and Chiu (2015) proposed a discrimination index that can be applied with a wider class of CDMs, the generalized DINA model, to validate the Q-matrices. Second, multiple missing data treating procedures exist besides the three implemented imputation approaches. The impact of different missingness treating approaches (e.g., multiple imputations) should also be considered when comparing the performance of Q-matrix validation methods in presence of missing data. In addition, the performance the both validation methods is examined in this study assuming attributes being noncompensatory. Under the framework of CDMs, however, the relationship of attributes contains both compensatory and noncompensatory in nature. Results may vary if the attributes were compensatory. These limitations should be taken into consideration and potentially addressed in the future studies. Although these limitations exist, information based on the simulation study seems promising in how researchers could potentially deal with missingness in the studied context.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Dai Shenghai <https://orcid.org/0000-0002-7829-0621>

Supplemental Material

Supplemental material is available for this article online.

References

- Brown, N. J. S., Dai, S., & Svetina, D. (2014, April). *Predictors of omitted responses on the 2009 National Assessment of Educational Progress (NAEP) mathematics assessment*. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.
- Brown, N. J. S., Svetina, D., & Dai, S. (2014, April). *Impact of methods of scoring omitted responses on achievement gaps*. Paper presented at the annual meeting of the National Council on Measurement in Education, Philadelphia, PA.
- Cheema, J. R. (2014). Some general guidelines for choosing missing data handling methods in educational research. *Journal of Modern Applied Statistical Methods*, 13(2), 53-75.
- Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, 37, 598-618.
- Chiu, C.-Y., & Douglas, J. A. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal item response patterns. *Journal of Classification*, 30, 225-250.
- De Ayala, R. J., Plake, B. S., & Impara, J. C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement*, 38, 213-234.
- DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement*, 36, 447-468. doi:10.1177/0146621612449069
- de la & Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, 45, 343-362.
- de la Torre, J., & Chiu, C.-Y. (2015). A general method of empirical Q-matrix validation. *Psychometrika*, 81, 253-273.
- de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69, 333-353.
- DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. *Handbook of Statistics Psychometrics*, 26, 979-1030.
- Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, 45, 225-245.
- Finch, H. (2010). Imputation methods for missing categorical questionnaire data: A comparison of approaches. *Journal of Data Science*, 8, 361-378.
- Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement*, 29, 262-277.
- Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for fusion model application to language assessment. *Language Testing*, 26, 31-73.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25, 258-272.

- Lee, Y.-S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing, 11*, 144-177.
- Leighton, J. P., & Gierl, M. J. (2007). Why cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 3-18). New York, NY: Cambridge University Press.
- Little, R. J., & Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods & Research, 18*, 292-326.
- Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: John Wiley.
- Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement, 36*, 548-564.
- Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39*, 247-264.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*, 379-416.
- Mislevy, R. J., & Wu, P. K. (1988). Inferring examinee ability when some item responses are missing. *ETS Research Report Series, 1988*(2), i-75. doi:10.1002/j.2330-8516.1988.tb00304.x
- Mislevy, R. J., & Wu, P.-K. (1996). Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing. *ETS Research Report Series, 1996*(2), i-36.
- 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*, 525-556.
- R Core Team. (2017). R: A language and environment for statistical computing (Version 3.4.1) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2017). CDM: Cognitive diagnosis modeling (R Package Version 5.4-0). Retrieved from <http://CRAN.R-project.org/package=CDM>
- Rogers, A. M., & Stoeckel, J. J. (2011). *NAEP 2009 mathematics, reading, science, and grade 12 restricted-use data files data companion* (NCES 2011-475). Washington, DC: Institute of Education Sciences, U.S. Department of Education.
- Rose, N., von Davier, M., & Xu, X. (2010). Modeling nonignorable missing data with item response theory (IRT). *ETS Research Report Series, 2010*(1), i-53.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581-592.
- Rupp, A. A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*, 78-96.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147-177.
- Sijtsma, K., & Van der Ark, L. A. (2003). Investigation and treatment of missing item scores in test and questionnaire data. *Multivariate Behavioral Research, 38*, 505-528.
- Tatsuoka, K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold & M. Safto. (Eds.), *Monitoring skills and knowledge acquisition* (pp. 453-488). Hillsdale, NJ: Lawrence Erlbaum.
- Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 51*, 337-350.
- van Buuren, S., & Groothuis-Oudshoorn, K. (2015). Mice: Multivariate imputation by chained equations (R Package Version 2.2.5). Retrieved from <http://CRAN.R-project.org/package=mice>
- Xu, X., & von Davier, M. (2006). Cognitive diagnosis for NAEP proficiency data. *ETS Research Report Series, 2006*(1), i-25.
- Zheng, Y., & Chiu, C.-Y. (2016). NPCD: Nonparametric methods for cognitive diagnosis (R Package Version 1.0-10). Retrieved from <http://CRAN.R-project.org/package=NPCD>