Article

Applied Psychological Measurement
2017, Vol. 41(4) 277–293
© The Author(s) 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621616686021
journals.sagepub.com/home/apm
SAGE

A Residual-Based Approach to Validate Q-Matrix Specifications

Jinsong Chen[1]

Abstract

Q-matrix validation is of increasing concern due to the significance and subjective tendency of Q-matrix construction in the modeling process. This research proposes a residual-based approach to empirically validate Q-matrix specification based on a combination of fit measures. The approach separates Q-matrix validation into four logical steps, including the test-level evaluation, possible distinction between attribute-level and item-level misspecifications, identification of the hit item, and fit information to aid in item adjustment. Through simulation studies and real-life examples, it is shown that the misspecified items can be detected as the hit item and adjusted sequentially when the misspecification occurs at the item level or at random. Adjustment can be based on the maximum reduction of the test-level measures. When adjustment of individual items tends to be useless, attribute-level misspecification is of concern. The approach can accommodate a variety of cognitive diagnosis models (CDMs) and be extended to cover other response formats.

Keywords

cognitive diagnosis model, Q-matrix, validation, fit measure, residual based

Introduction

Cognitive diagnosis models (CDMs) have received increased attention in educational measurement. Unlike conventional test frameworks such as classical test theory or item response theory which focus on assessing examinees' overall ability, CDMs have been developed primarily to diagnose the strengths and weaknesses of examinees across a set of attributes. CDM-based assessments can provide finer-grained, domain-specific diagnostic information that can be used for different formative purposes, such as facilitating a more precise measurement of students' learning status and required remedies or aiding in the design of better instruction. Recently, various types of CDMs that can be applied across a wide range of practical settings had been developed. Among others, these included (a) highly constrained models such as the *deterministic inputs, noisy ''and''* (DINA; Junker & Sijtsma, 2001) *gate* and the *deterministic inputs, noisy ''or''* (DINO; Templin & Henson, 2006) *gate*; (b) additive models such as the linear logistic model (LLM; Maris, 1999), the reduced reparameterized unified model (R-RUM; DiBello, Roussos, & Stout, 2007; Hartz, 2002), and the *additive* CDM (A-CDM; de la Torre, 2011); and

[1] Sun Yat-Sen University, Guangzhou, China

Corresponding Author:
Jinsong Chen, Department of Psychology, Sun Yat-Sen University, No. 135, Xingangxi Road, Guangzhou 510275, China.
Email: jinsong.chen@live.com

(c) saturated models such as the log-linear CDM (LCDM; Henson, Templin, & Willse, 2009) and the generalized DINA (G-DINA; de la Torre, 2011).

A common and critical process of adopting CDMs for diagnostic purpose is to specify the item–attribute relationships by domain experts in the Q-matrix (K. K. Tatsuoka, 1983). Without the Q-matrix, substantive knowledge cannot be integrated into the modeling process to provide meaningful diagnostic information. Considering the significance of the Q-matrix and the subjective tendency of Q-matrix specifications by domain experts, Q-matrix validation is of increasing concern among methodologists (e.g., Chiu, 2013; de la Torre, 2008; de la Torre & Chiu, 2015; DeCarlo, 2012; Liu, Xu, & Ying, 2012; Rupp & Templin, 2008). Specifically, different validation methods or procedures that can apply to different CDMs were provided. As a representative, the general validation method based on the discrimination index (de la Torre, 2008) and the G-DINA model can cover a wide range of CDMs (de la Torre & Chiu, 2015). Relying on the item-level index and an exhaustive search algorithm, the general method can identify multiple inappropriate q-vectors and suggest substitutions simultaneously. The essence of the method is to maximize the variation of latent group probabilities based on specific cutoffs. Although the discrimination-based general method was promising, the conditions of the simulation studies were limited and the applicability of the method in different practical situations was unclear (de la Torre & Chiu, 2015). More importantly, one might obtain different results based on different cutoffs, and there is no equivocal rule to determine the optimal value. Besides, it is questionable if methods based on a single index can effectively cover the complexity of Q-matrix misspecifications across various settings.

This research proposes a residual-based approach to empirically validate Q-matrix specification with a combination of fit measures. The approach builds on the absolute fit statistics based on the residuals between the second moments of the observed and expected response patterns (J. Chen, de la Torre, & Zhang, 2013). But the fit measures rely on different mechanisms to exploit the residuals. There are four logical steps of Q-matrix validation under the approach: (a) test-level evaluation of the Q-matrix in absolute sense, (b) possible distinction between attribute-level and item-level misspecifications, (c) identification of misspecified items one by one and based on specific sequence, and (d) fit information to aid item adjustment in the right track. The approach can accommodate various reduced and saturated CDMs for dichotomous and polytomous attributes (e.g., J. Chen & de la Torre, 2013; von Davier, 2008). Moreover, it can be extended to cover other response formats such as polytomous or nominal.

The rest of this article will first introduce the theoretical background of the Q-matrix, its misspecification, the proposed approach, and the related fit measures. The designs and results of several simulation studies will then be presented. After that, two real-life examples will be used to explore the effectiveness of the fit measures in practice. Practical recommendations on how to adopt the approach will also be given. Finally, the article will end with some further discussions.

Theoretical Framework

Q-matrix and Its Misspecifications

Let q_{jk} denote the element in row j and column k of a $J \times K$ Q-matrix, where J and K represent the number of items and attributes, respectively. The entry q_{jk} is specified as 1 if mastery of attribute k is required to answer item j correctly, and as 0 otherwise. The jth row of the Q-matrix (i.e., $\boldsymbol{q}_j$) is called the jth q-vector, which gives the attribute specification of item j. $K_j^* = \sum_{k=1}^{K} q_{jk}$ is used to denote the number of required attributes for item j. For notational convenience, let the first K_j^* attributes be required for item j. The required attributes for item

j can be represented by the reduced vector $\boldsymbol{\alpha}_{lj}^{*} = (\alpha_{l1}, \cdots, \alpha_{lK_j^*})'$, where $l = 1, \ldots, 2^{K_j^*}$. By adopting the concept of required attributes (de la Torre, 2011), the model can be simplified because the number of attribute vectors to be considered for item j reduces from 2^K to $2^{K_j^*}$. The probability that examinees with reduced vector $\boldsymbol{\alpha}_{lj}^{*}$ will answer item j correctly is denoted as $P(X_j = 1|\boldsymbol{\alpha}_{lj}^{*}) = P(\boldsymbol{\alpha}_{lj}^{*})$, which contains the item parameters. Readers can refer to the literature (e.g., de la Torre, 2009, 2011; Henson et al., 2009; von Davier, 2008) for model formulation and parameter estimation of various saturated and reduced CDMs.

Q-matrix misspecification can refer to the misspecification of the attributes (i.e., column vectors) or the items (i.e., row vectors or q-vectors) in the Q-matrix. If the misspecification is on the attribute level (e.g., ill-defined attributes), adjustment of individual items can be inefficient or difficult, and it is better to consider test-level adjustment (e.g., redefining the attributes and/or reconstructing the Q-matrix). On the item level, one can distinguish among three types of Q-matrix misspecification: underspecified (i.e., some 1s have been incorrectly specified as 0s in the q-vectors), overspecified (i.e., some 0s have been incorrectly specified as 1s), or both under- and overspecified (i.e., some 1s have been incorrectly specified as 0s, and some 0s have been incorrectly specified as 1s).

Although any Q-matrix misspecification will result in model–data misfit, it might not be detectable if the Q-matrix is incomplete or unidentified, or the conditions are extreme such as small sample size or test length. Readers can refer to the literature about Q-matrix completeness (e.g., Chiu, Douglas, & Li, 2009; Köhn & Chiu, 2016) or Q-matrix identifiability (e.g., Y. Chen, Liu, Xu, & Ying, 2015; Xu & Zhang, 2016) across different situations. In addition, factors such as the compensatory nature of the CDM (J. Chen et al., 2013) and item quality (Rupp & Templin, 2008) can interfere with the misspecification to confound the source of misfit. Consequently, evaluation of the validation conditions should precede the validation of the Q-matrix. More about condition evaluation will be discussed later.

The Residual-Based Approach to Q-Matrix Validation

Under this approach, there are different logical steps of Q-matrix validation, which corresponds to different types of fit information. First, one needs absolute fit information to evaluate if the Q-matrix is rejected at the test level, and no further action is required if the answer is no. Second, in case of misspecification, it is better to distinguish between the cases of a few and many misspecified items. In the latter case, it is more likely that the misspecifications occur due to attribute-level reasons rather than at random, and it is better to consider test-level rather than item-level adjustment in practice. However, one needs to understand that any fit information to distinguish between the two cases is deemed suggestive due to the relative difference between ''a few'' and ''many'' misspecified items. More about how to utilize such information in practice will be discussed in the Section ''Practical Recommendations.''

Third, in case there are only a few misspecified items, this research proposes to identify the misspecified items one by one and based on specific sequence. Simultaneous identification of multiple misspecified items is not suggested as interference among items can make some misspecified items ''appear'' correctly specified or vice versa. Moreover, it will need universal cutoff to determine what items are misspecified or not, which is challenging. Instead, it would be better to identify and adjust the item that is most likely misspecified, given that such item-level fit information is available. Given that the item can be fixed, one can proceed to detect and adjust the next item that is most likely misspecified similarly, until no misspecification can be found. Fourth, for the above strategy to be successful, one needs fit information that can aid item adjustment (e.g., adding and/or dropping attributes) in the right track. Specifically, the fit information before and after item adjustment should change consistently (e.g., reduction), given that any misspecified item is correctly adjusted. Moreover, if the misspecified item is correctly adjusted, the change magnitude should be the largest among all possible adjustments of the same item. However, it is challenging to demonstrate that correct adjustment is always matched with maximum change theoretically or in simulation studies. Instead, this research will rely on real-life examples for such purpose, although it can be a topic in future research.

This research will investigate a combination of fit measures for the above steps. All proposed fit measures are based on the residuals between the observed and expected response patterns. Specifically, all measures involve the residual between the observed and predicted Fisher-transformed correlation of item pairs (referred to as *r* or correlation based), and the residual between the observed and predicted log-odds ratio (LOR) of item pairs (referred to as *l* or LOR based). Because the predicted responses are model based, the residuals are associated with the maximum likelihood based on the likelihood function of the model and should approach 0 when the model fits the data well. But different measures rely on different mechanisms to utilize the residuals on the test- or item level, and accordingly are expected to function differently.

Let $\boldsymbol{X}_j$ and $\tilde{\boldsymbol{X}}_j$ denote the observed and predicted response vectors for item j, respectively. For item j, it is given as

$$r_{jj'} = \left|Z\left[Cor\left(\boldsymbol{X}_j, \boldsymbol{X}_{j'}\right)\right] - Z\left[Cor\left(\tilde{\boldsymbol{X}}_j, \tilde{\boldsymbol{X}}_{j'}\right)\right]\right|. \tag{1}$$

$$l_{jj'} = \left|\log(N_{11}N_{00}/N_{01}/N_{10}) - \log\left(\tilde{N}_{11}\tilde{N}_{00}/\tilde{N}_{01}/\tilde{N}_{10}\right)\right|, \tag{2}$$

where $\tilde{N}$ is the predicted sample size, $j \neq j'$, $Z[Cor(\cdot)]$ is the Fisher transformation of Pearson's correlation, and $N_{yy'}$ and $\tilde{N}_{yy'}$ are the number of observed and predicted examinees, respectively, who scored y on item j and y' on item j'. The approximate standard errors (SEs) of *r* and *l* can be computed as

$$SE\left(r_{jj'}\right) = [N - 3]^{1/2}, \tag{3}$$

$$SE\left(l_{jj'}\right) = \left[\frac{\tilde{N}\left(1/\tilde{N}_{11} + 1/\tilde{N}_{00} + 1/\tilde{N}_{01} + 1/\tilde{N}_{10}\right)}{N}\right]^{1/2}. \tag{4}$$

With these SEs, the *z*-scores of *r* and *l* can be derived for further usage. With a large $\tilde{N}$, the expected response patterns can be simulated stably by integrating across the estimated item parameters and the posterior distribution of the attribute patterns. Nevertheless, randomness is inevitable due to simulation. For both the *r* or *l* statistics and a test with J items, there are $(J - 1)$ and $J(J - 1)/2$ item pairs of z-scores to be evaluated at the item and test levels, respectively.

Investigated in J. Chen et al.'s (2013) work, the first two fit measures adopt the maximum *z*-score of each statistics (denoted as *zr* and *zl* for *r* and *l*, respectively) for test-level misspecification, which can be used to provide absolute fit information in the first step. With the fit measures and Bonferroni correction to keep the Type I error normal, the null hypothesis of no misspecification can be rejected based on specific significance level (i.e., in absolute sense). It was found that both *zr* and *zl* were similar in detecting test-level misspecification, with very high statistical power and conservative Type I error (J. Chen et al., 2013). Moreover, their performance was stable across different conditions such as sample sizes, test lengths, or true or fitting models. However, both were insensitive to purely overspecified Q-matrices unless highly constrained models were used.

For fit information in the second step, it is proposed to use the root mean square (RMS) value of the *z*-scores at the test level, which are

$$sr = \left[\frac{2\sum_{j=1}^{J}\sum_{j'=1}^{j-1}\left(\frac{r_{jj'}}{SE\left(r_{jj'}\right)}\right)^{2}}{J(J-1)}\right]^{1/2}, \tag{5}$$

$$sl = \left[\frac{2\sum_{j=1}^{J}\sum_{j'=1}^{j-1}\left(\frac{l_{jj'}}{SE\left(l_{jj'}\right)}\right)^{2}}{J(J-1)}\right]^{1/2}, \tag{6}$$

for *r* and *l*, respectively. *sr* and *sl* are test-level measures comparable with *zr* and *zl*, respectively, but likely less sensitive to individual misspecified items due to averaging across all item pairs. This research will investigate if they can provide useful fit information to signal a large number of misspecified items based on specific cutoff value.

For fit information in the third step, it is proposed to use the RMS value of the *z*-scores at the item level, which are

$$sr_j = \left[\frac{\sum_{j'\neq j}\left(\frac{r_{jj'}}{SE\left(r_{jj'}\right)}\right)}{J-1}\right]^{1/2}, \tag{7}$$

$$sl_j = \left[\frac{\sum_{j'\neq j}\left(\frac{l_{jj'}}{SE\left(l_{jj'}\right)}\right)}{J-1}\right]^{1/2}, \tag{8}$$

for *r* and *l*, respectively. sr_j and sl_j are item-level measures that average all *z*-scores of residuals related to specific item. The items with the maximum values for sr_j and sl_j (denoted as *mr* and *ml*, respectively) can be considered as most likely misspecified, and will be called the hit items for convenience of discussion. When there are many misspecified items however, it is unclear if the hit item will be a correctly specified item due to interference. This research will investigate if, and the conditions under which, the misspecified items can be detected sequentially as the hit item by the *mr* and *ml*.

For fit information in the fourth step, one can consider the changes of *sr* and *sl* (denoted as Δsr and Δsl, respectively) and the changes of *mr* and *ml* (denoted as Δmr and Δml, respectively) before and after the adjustment of the hit item. This research will investigate which of the above fit measures are more effective to aid item adjustment in the right track.

Simulation Studies

Different simulation studies were used to investigate how the residual-based approach can perform to validate the Q-matrix. The true Q-matrix Q0 is shown in Table 1, and different misspecified Q-matrices were used across studies as shown in Table 2. Purely overspecification was excluded as the fit measures were deemed inappropriate unless highly constrained models were fitted. The true CDMs considered were the DINA model and LLM, which can represent the cases of highly constrained (e.g., conjunctive) and less constrained CDMs, respectively. The

Table 1. True Q-Matrix Q0 for J = 20.

	Attribute				
Item	α_1	α_2	α_3	α_4	α_5
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4	0	0	0	1	0
5	0	0	0	0	1
6	1	1	0	0	0
7	1	0	0	0	1
8	0	1	1	0	0
9	0	0	1	1	0
10	0	0	0	1	1
11	1	1	1	0	0
12	1	1	0	0	1
13	1	0	0	1	1
14	0	1	1	1	0
15	0	0	1	1	1
16	1	0	1	0	0
17	1	0	0	1	0
18	0	1	0	1	0
19	0	1	0	0	1
20	0	0	1	0	1

Note. The true Q-matrix for J = 40 was duplicated.

test lengths were set to J = 20 and 40, and sample size was fixed at N = 1,000. The saturated G-DINA model was used as the fitting CDM. The number of attributes K was fixed to 5. The multivariate normal threshold method (Chiu et al., 2009) was adopted to simulate attribute patterns and correlations. Underlying the discrete patterns of the K attribute, a multivariate normal distribution, $MVN(\mathbf{0}, \mathbf{\Sigma})$, of K continuous latent variables with all variances and covariance in $\mathbf{\Sigma}$ equal to 1.0 and R, respectively, was assumed. In this case, R is the correlation of the latent variables underlying the attributes and was set to .5. For the item parameters, $P(\mathbf{0})$ and $P(\mathbf{1})$ were randomly generated from $Unif(0.0, 0.3)$ and $Unif(0.7, 1.0)$, respectively. For the LLM, all main effects (i.e., δ_{jk}) were set to be equal, so that each required attribute contributed equally to the examinees' success probability. Finally, all item parameters were randomly generated in each replication. Such a simulation design was closer to practical situations compared with previous studies (e.g., J. Chen et al., 2013). Each simulation cell was replicated 200 times, and the estimation code was written in Ox (Doornik, 2003).

It was found that the performance of the fit measures was generally better when the true model was the DINA model rather than the LLM, but the differences were not substantial in most cases. To simplify analysis, results from the two true CDMs were averaged.

Study 1: The Scenario of Single Misspecified Item

Design. This study considered the cases where no more than single item was misspecified, which can serve as the baseline performance of the approach. As shown in Table 2, the five misspecified Q-matrices (i.e., Q1a ~ Q1e) covered all scenarios that the one-, two-, and three-attribute item can be misspecified (except for purely overspecification). In addition to the

Table 2. Misspecified Q-Matrices and Items in Three Studies.

S	Q	I	*q*-Vector					T	RA
1	Q1a	5	$\underline{1}$	0	0	0	$\underline{0}$	b	1
1	Q1b	6	1	$\underline{0}$	0	0	0	u	2
1	Q1c	6	1	$\underline{0}$	$\underline{1}$	0	0	b	2
1, 2	Q1d	11	1	1	$\underline{0}$	0	0	u	3
1	Q1e	11	1	1	$\underline{0}$	$\underline{1}$	0	b	3
2	Q2	10	0	0	$\underline{1}$	$\underline{0}$	1	b	2
		11	1	1	$\underline{0}$	0	0	u	3
2	Q3	6	1	$\underline{0}$	1	0	0	u	2
		10	0	0	$\underline{1}$	$\underline{0}$	1	b	2
		11	1	1	$\underline{0}$	0	0	u	3
2	Q4	5	$\underline{1}$	0	0	0	$\underline{0}$	b	1
		6	1	$\underline{0}$	$\underline{1}$	0	0	b	2
		10	0	0	$\underline{1}$	$\underline{0}$	1	b	2
		11	1	1	$\underline{0}$	0	0	u	3
3	Q5	7	1	0	0	0	$\underline{0}$	u	2
		10	0	0	$\underline{1}$	$\underline{0}$	1	b	2
		12	$\underline{0}$	1	0	0	1	u	3
		13	1	$\underline{1}$	0	1	$\underline{0}$	b	3
		15	0	0	1	$\underline{0}$	1	u	3
		19	0	$\underline{0}$	0	0	1	u	3
		20	$\underline{1}$	0	$\underline{0}$	0	1	b	3
3	Q6	7	1	0	0	0	$\underline{0}$	u	2
		10	0	0	0	1	$\underline{0}$	u	2
		12	1	1	0	0	$\underline{0}$	u	3
		13	1	0	0	1	$\underline{0}$	u	3
		15	0	0	1	1	$\underline{0}$	u	3
		19	0	1	0	0	$\underline{0}$	u	3
		20	0	0	1	0	$\underline{0}$	u	3

Note. S = study; Q = Q-matrix; I = item; T = types of misspecification; RA = number of required attributes; misspecified entries were underscored; u = underspecification; b = both under- and overspecification.

values of the fit measures, the percentage the model was rejected with *zr* or *zl* at 5% significance level (%rejection) and the percentage the hit item was misspecified (%hit) were of concern.

Result. The performance of different fit measures can be found in Table 3. The difference between the correlation-based and LOR-based measures was trivial. The low Type I error rates (i.e., rejection% for Q0) suggested the conservative nature of the *zr* and *zl* statistics. The rejection% from Q1a to Q1e showed that the more required attributes the misspecified item had, the less power either statistics can offer. When misspecification occurred however, the hit items were most likely misspecified based on either *mr* or *ml*. The values of the *sr, sl, mr*, and *ml* statistics were the smallest for Q0 and tended to increase with less required attributes. It implied that the four statistics always reduced if the misspecified item was correctly adjusted, and that misspecified item with less required attribute was more detectable.

Study 2: The Scenario of Multiple Misspecified Items

Design. As shown in Table 2, four hierarchically misspecified Q-matrices were designed to evaluate if multiple misspecified items can be detected as the hit item and possibly adjusted

Table 3. Simulation Results for Study 1.

		Correlation based				LOR based			
J	Q-matrix	%Rejection	%Hit	*sr*	*mr*	%Rejection	%Hit	*sl*	*ml*
20	Q0	0	0	0.83	1.11	0	0	0.81	1.09
	Q1a	100	100	1.34	3.40	100	100	1.31	3.35
	Q1b	83	91	1.09	2.04	79	92	1.06	1.98
	Q1c	81	91	1.05	1.88	77	91	1.02	1.84
	Q1d	61	90	0.97	1.62	56	87	0.95	1.58
	Q1e	54	85	0.95	1.52	48	83	0.93	1.48
40	Q0	0	0	0.84	1.05	0	0	0.81	1.03
	Q1a	100	100	1.49	3.50	100	100	1.48	3.49
	Q1b	100	100	1.12	2.83	100	100	1.09	2.76
	Q1c	95	100	1.08	2.05	92	100	1.05	2.01
	Q1d	91	100	1.01	1.65	88	100	0.99	1.64
	Q1e	82	100	0.98	1.56	81	100	0.96	1.53

Note. LOR = log-odds ratio; %rejection = % the model was rejected with *zr* or *zl* at 5% significance level; %hit = % the hit item was misspecified.

Table 4. Simulation Results for Study 2, Part I.

		Correlation based				LOR based			
J	Q-matrix	%Rejection	%Hit	*sr*	*mr*	%Rejection	%Hit	*Sl*	*ml*
20	Q1d	61	91	0.97	1.59	56	92	0.95	1.56
	Q2	98	95	1.13	1.85	98	95	1.11	1.82
	Q3	99	97	1.30	2.25	99	97	1.27	2.19
	Q4	100	100	1.62	3.53	100	100	1.59	3.51
40	Q1d	91	100	0.94	1.65	88	100	0.92	1.64
	Q2	100	100	1.12	1.97	100	100	1.09	1.95
	Q3	100	100	1.33	2.78	100	100	1.30	2.74
	Q4	100	100	1.76	3.62	100	100	1.74	3.56

Note. Values averaged across three true CDMs. LOR = log-odds ratio; %rejection = % the model was rejected with *zr* or *zl* at 5% significance level; %hit = % the hit item was misspecified; CDMs = cognitive diagnosis models.

sequentially. There existed one to four misspecified items from Q1d to Q4, which were nested within each other. As implied in Study 1, Items 5, 6, 10, and 11 were expected to be the hit item for Q4, Q3, Q2, and Q1d, respectively. Accordingly, the four Q-matrices constituted a most likely sequence that the four misspecified items in Q4 would be detected as the hit item one by one, given that they will be corrected sequentially. Moreover, one can evaluate if the *sr, sl, mr*, and *ml* statistics change consistently along the sequence, and accordingly can be useful to aid item adjustment in the right track.

Result. Part I of the performance of different fit measures can be found in Table 4. Similarly, the difference between the correlation-based and LOR-based measures was trivial. The %hit, %rejection, and the values of all fit measures decreased from Q4 to Q1d consistently. Note that the trends of change were consistent across test lengths. This suggested that the reduction of the *sr, sl, mr*, or *ml* statistics can aid adjustment of the hit item in the right track. The trends of the %rejection suggested that the power of the *zr* and *zl* statistics reduced as the test length got

Table 5. Simulation Results for Study 2, Part II.

		Correlation based				LOR based			
J	Q-matrix	IHM	IH%	STE (*sr*)	STE (*mr*)	IHM	IH%	STE (*zl*)	STE (*ml*)
20	Q1d	11	91	.07	.13	11	92	0.07	0.13
	Q2	10	66	.06	.12	10	68	0.06	0.12
	Q3	6	72	.06	.16	6	70	0.06	0.17
	Q4	5	100	.06	.08	5	100	0.06	0.08
40	Q1d	11	100	.04	.12	11	100	0.04	0.12
	Q2	10	96	.04	.11	10	96	0.05	0.11
	Q3	6	86	.04	.10	6	87	0.05	0.11
	Q4	5	100	.04	.10	5	100	0.04	0.10

Note. Values averaged across three true CDMs. LOR = log-odds ratio; IHM = item hit mostly; IH% = % the IHM was hit; STE = *SD*-to-estimate ratio; CDMs = cognitive diagnosis models.

smaller, and can be especially low when there existed only one multiple-attribute-misspecified item (i.e., Q1b). The %hit suggested that the hit item can be a correct item up to 28% of times in the worst case. But it was found that once the model was rejected based on *zr* or *zl*, the hit item was always misspecified.

Part II of the performance of different fit measures can be found in Table 5. In all cases, the items hit mostly were as expected (i.e., Items 5, 6, 10, and 11 for Q4, Q3, Q2, and Q1d, respectively). However, other items can be hit up to ~50% of times when there were multiple misspecified items, which evidenced the interference among the items. Fortunately, the hit item was always misspecified when the model was rejected based on *zr* or *zl*. In brief, it meant that correction of the hit item based on reduction of one of the four statistics would result in hitting the next misspecified item, until the Q-matrix was NOT rejected based on *zr* or *zl*. To evaluate the stability of the four statistics, the standard deviation (*SD*)-to-estimate ratios were compared. As shown in Table 5, the ratios for *sr* or *sl* were always smaller than those for *mr* or *ml*, which suggested that the changes of former two statistics were more stable than those of the latter two, and hence were preferred for item adjustment.

Study 3: The Scenario of Many Misspecified Items

Design. There can be various scenarios of many misspecified items. With two Q-matrices (Q5 and Q6 in Table 2), two cases were investigated: random misspecification versus serious underspecification on specific attribute. They were investigated to understand if the item-level measures would hit correctly specified rather than misspecified items, and how the test-level measures would perform in case it did. In both cases, the same seven items were misspecified, which occupied 35% of all items for $J = 20$. In the random cases, each attribute was misspecified (under- or both) an equal number of times, whereas all underspecification occurred on the last attribute in the other case.

Result. The performance of the fit measures can be found in Table 6. For the case of random misspecification, the hit item was misspecified at least 99% of times, and the models were always rejected at 5% significance level. It suggested that the misspecified items can be still detected sequentially as the hit item in this case. For the case of serious underspecification on specific attribute, the models were always rejected at 5% significance level. However, the low %hit when $J = 20$ suggested that the procedure of sequential detection with the hit item would

Table 6. Simulation Results for Study 3.

		Correlation based				LOR based			
J	Q-matrix	%Hit	*mr*	*sr*	*P*(*sr* > 1.5)	%Hit	*ml*	*sl*	*P*(*sl* > 1.5)
20	Q5	99	2.52	1.57	.72	99	2.48	1.54	.66
	Q6	15	3.17	1.62	.85	15	3.15	1.59	.78
40	Q5	100	3.39	1.65	.94	100	3.33	1.61	.89
	Q6	67	3.47	1.77	1.00	67	3.40	1.74	1.00

Note. The %rejection was 100% for all models. LOR = log-odds ratio; %hit = % the hit item was misspecified.

be most likely failed in this case. Although the %hit raised largely as the test length was increased to 40, there was still a considerable chance of hitting a correct rather than misspecified item.

No clear distinction based on any of the fit measures can be made between the cases of random misspecification and serious underspecification on specific attribute. Between the cases of a few and of many misspecified items, however, one can find a rather clear cutoff based on either *sr* or *sl*. Both statistics were most likely smaller than 1.5 in the former cases (Table 3 or 4) but larger than 1.5 in the latter cases (Table 6). Accordingly, the criterion that *sr* or $sl > 1.5$ can be regarded as a concern of many misspecified items. The concern can be further treated as a suggestive signal of attribute-level rather than item-level treatment, if one believes that a case of many misspecified items should not occur at random.

Real-Life Examples

In this section, empirical data from two real-life examples were analyzed to evaluate how the above fit measures and simulation results can be used to validate the Q-matrix in practice. The saturated G-DINA model was used in both examples.

Example 1: Fraction Subtraction

The data used were a subset of the data originally described and analyzed by K. K. Tatsuoka (1990), and more recently by C. Tatsuoka (2002), de la Torre and Douglas (2004), and DeCarlo (2011). The data consisted of responses of 536 middle school students to 20 fraction subtraction items measuring eight attributes. The original Q-matrix and items can be found in Table 7.

J. Chen et al. (2013) found that the Q-matrix was rejected at 5% significance level with the saturated CDM, and Item 8 (i.e., $2/3 - 2/3$) was problematic as the only required attribute is not necessary to answer the item correctly. As shown in Table 8, Item 8 was also the hit item based on *mr* or *ml*. One can rely on the maximum change of *sr* or *sl* (i.e., Δsr or Δsl) to adjust the item. Rather than adding any required attribute, it turned out that both *sr* and *sl* were reduced the most by removing the item. After Item 8 was removed, the modified Q-matrix was significant at around 5% level based on *zr* or *zl*, and Item 1 was hit (second row in Table 8). As the simulation demonstrated however, the power for mild misspecification was low in this case (i.e., $J = 20$). The best adjustment for Item 1, based on the maximum reduction of *sr* or *sl*, was to specify α_5. After that, Items 9 and 6 were hit based on *mr* and *ml*, respectively. Involvement of subject experts is necessary to justify the adjustments of Item 1.

Table 7. Items and Q-Matrix for the Fraction Subtraction Data.

		Attributes									
Item number	Text	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	*P*(**0**)	*P*(**1**)
1	5/3 − 3/4	0	0	0	1	0	1	1	0	.02	1.00
2	3/4 − 3/8	0	0	0	1	0	0	1	0	.02	.97
3	5/6 − 1/9	0	0	0	1	0	0	1	0	.01	.89
4	3 1/2 − 2 3/2	0	1	1	0	1	0	1	0	.41	.88
5	4 3/5 − 3 4/10	0	1	0	1	0	0	1	1	.06	.88
6	6/7 − 4/7	0	0	0	0	0	0	1	0	.13	.95
7	3 − 2 1/5	1	1	0	0	0	0	1	0	.00	.80
8	2/3 − 2/3	0	0	0	0	0	0	1	0	.47	.81
9	3 7/8 − 2	0	1	0	0	0	0	0	0	.23	.78
10	4 4/12 − 2 7/12	0	1	0	0	1	0	1	1	.01	.85
11	4 1/3 − 2 4/3	0	1	0	0	1	0	1	0	.05	.94
12	11/8 − 1/8	0	0	0	0	0	0	1	1	.10	.96
13	3 3/8 − 2 5/6	0	1	0	1	1	0	1	0	.00	.67
14	3 4/5 − 3 2/5	0	1	0	0	0	0	1	0	.03	.96
15	2 − 1/3	1	0	0	0	0	0	1	0	.02	.88
16	4 5/7 − 1 4/7	0	1	0	0	0	0	1	0	.02	.91
17	7 3/5 − 4/5	0	1	0	0	1	0	1	0	.00	.87
18	4 1/10 − 2 8/10	0	1	0	0	1	1	1	0	.00	.86
19	4 − 1 4/3	1	1	1	0	1	0	1	0	.00	.82
20	4 1/3 − 1 5/3	0	1	1	0	1	0	1	0	.00	1.00

Note. α_1 = convert a whole number to a fraction; α_2 = separate a whole number from a fraction; α_3 = simplify before subtracting; α_4 = find a common denominator; α_5 = borrow from whole number part; α_6 = column borrow to subtract the second numerator from the first; α_7 = subtract numerators; α_8 = reduce answers to simplest form.

Table 8. Q-Matrix Validation Results for the Fraction Subtraction Data.

J	z_c	*zr*	*zl*	*mr*	*ml*	*sr*	*sl*	Hit	BA
20	3.65	10.57	11.97	4.89	5.70	1.97	2.18	8	Remove
19	3.62	3.67	3.45	1.76	1.69	1.27	1.26	1	α_5: 0 → 1
19	3.62	2.84	3.50	1.51	1.53	1.19	1.17	9/6	NA

Note. Saturated CDM was fitted. z_c = critical z-score at 5% significance level; BA = best adjustment based on the maximum reduction of *sr* or *sl*; CDM = cognitive diagnosis model; NA = not applicable.

Example 2: Program for International Student Assessment (PISA) Reading

The PISA 2000 reading assessment data (Organisation for Economic Co-Operation and Development [OECD], 1999, 2006a) with the released items (OECD, 2006b) were used. The subset of the data consisted of responses of 1,029 examinees to 20 English reading items. An initial Q-matrix can be specified with the five processes (aspects) of reading under the PISA assessment framework (OECD, 1999; 2006a), as shown in Table 9.

Although the initial Q-matrix can be used to help redefine a new set of attributes with appropriate Q-matrix (e.g., H. Chen & Chen, 2015), item-level adjustment was deemed inappropriate as the assessment was not designed for diagnostic purpose with the initial Q-matrix. To some extent, the data and the initial Q-matrix with only single-attribute items are similar to the case of serious underspecification in Study 3 (e.g., similar sample size, test length, number of

Table 9. Items and Initial Q-Matrix for the PISA Data.

Item number	Code	α_1	α_2	α_3	α_4	α_5
1	R040Q02	0	0	1	0	0
2	R040Q03A	1	0	0	0	0
3	R040Q03B	0	0	0	0	1
4	R040Q04	0	1	0	0	0
5	R040Q06	0	0	1	0	0
6	R077Q02	1	0	0	0	0
8	R077Q04	0	0	1	0	0
10	R077Q06	0	0	0	1	0
11	R088Q01	0	1	0	0	0
15	R088Q07	0	0	0	0	1
16	R110Q01	0	1	0	0	0
17	R110Q04	1	0	0	0	0
18	R110Q05	1	0	0	0	0
19	R110Q06	0	0	1	0	0
20	R216Q01	0	1	0	0	0
21	R216Q02	0	0	0	0	1
22	R216Q03T	0	0	1	0	0
23	R216Q04	0	0	1	0	0
24	R216Q06	0	0	1	0	0
25	R236Q01	0	0	1	0	0

Note. α_1 = retrieving information; α_2 = forming a broad general understanding; α_3 = developing an interpretation; α_4 = reflecting on and evaluating the content of a text; α_5 = reflecting on and evaluating the form of a text. PISA = Program for International Student Assessment.

attributes, and likely attribute-level underspecification). Here, the initial Q-matrix is used to illustrate the case of attribute-level misspecification. Similarly, one can treat the hit item as misspecified and adjust it sequentially based on maximum reduction of *sr* or *sl*. As shown in Table 10, the modified Q-matrix was still rejected at 5% significance level even after five rounds of adjustment. Moreover, both *sr* and *sl* were always larger than 1.5, which implied that the concern of many misspecified items persisted. So did the signal to conduct attribute-level rather than item-level adjustment, if it is believed that random misspecification cannot explain the possible existence of many misspecified items in this case.

Practical Recommendations

Based on the above simulation studies and empirical examples, a set of procedures is proposed along the four steps of Q-matrix validation under the residual-based approach, as illustrated in Figure 1 and below:

1. One needs to evaluate the completeness and identifiability of the Q-matrix and if the conditions such as the sample size and test length are good enough for Q-matrix validation. When the Q-matrix is incomplete or unidentifiable, the validation of Q-matrix misspecification can be questionable. When the sample size or test length is too small, the power to reject any misspecified Q-matrix might be too small, and further action is preferred.
2. Given that the conditions are acceptable, one can proceed to evaluate if the Q-matrix is acceptable at the test level based on the *zr* or *zl* statistics.

Table 10. Initial Q-Matrix Validation Results for the PISA Data.

J	z_c	*zr*	*zl*	*mr*	*ml*	*sr*	*sl*	Hit	BA
20	3.65	10.25	9.01	3.00	2.89	2.18	2.10	19	Remove
19	3.62	8.23	8.69	2.98	3.02	2.04	1.99	2	Remove
18	3.59	4.75	4.55	2.80	2.80	2.01	1.94	8	α_5: 0 → 1
18	3.59	4.57	3.90	2.23	2.04	1.81	1.72	6	α_2: 0 → 1
18	3.59	4.78	3.77	2.42	2.09	1.80	1.70	11	α_2: 0 → 1
18	3.59	4.26	3.74	2.07	1.98	1.71	1.63	17	NA

Note. Saturated CDM was fitted. z_c = critical z-score at 5% significance level; BA = best adjustment based on the maximum reduction of *sr* or *sl*. PISA = Program for International Student Assessment; NA = not applicable.

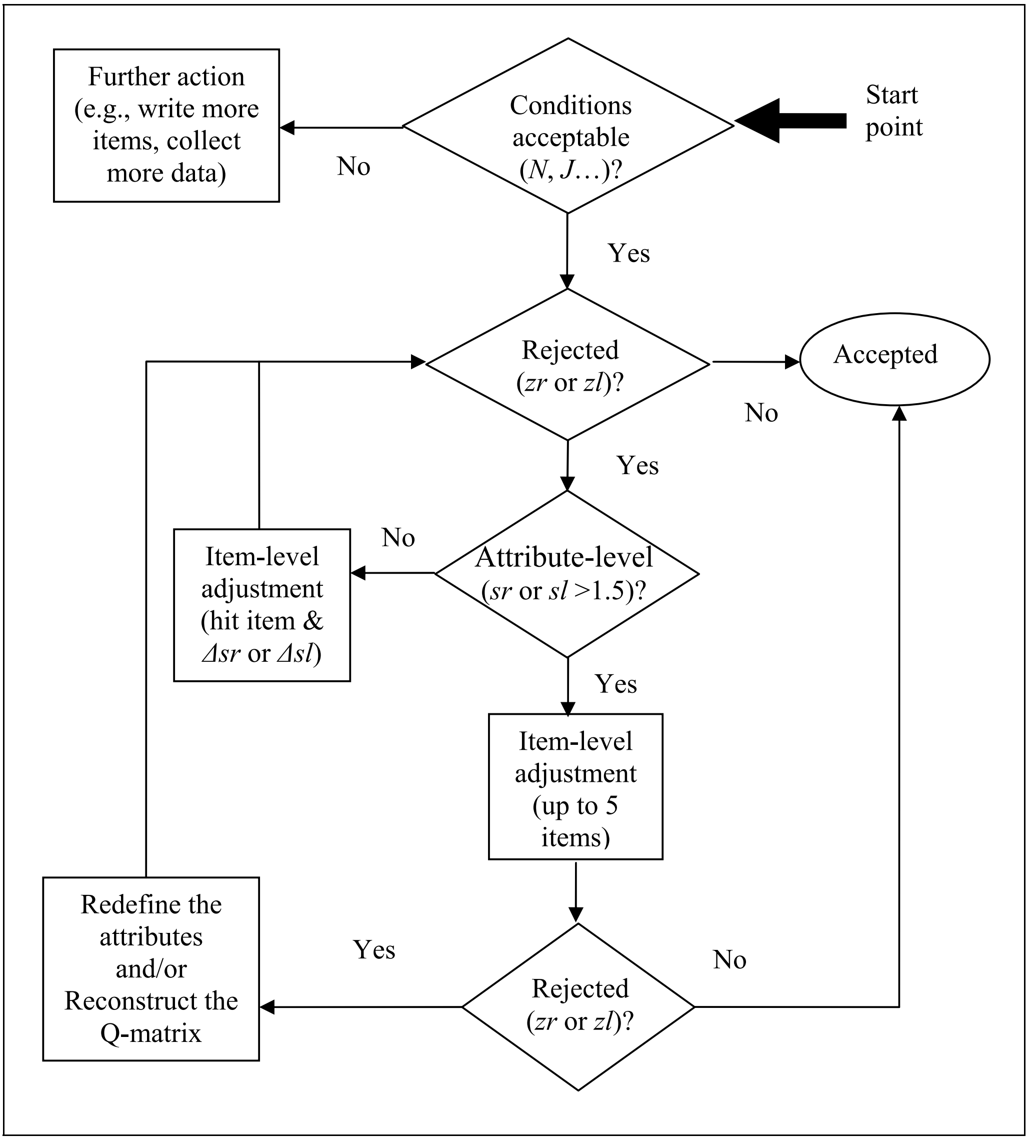
Further action (e.g., write more items, collect more data)
No
Conditions acceptable (N, J…)?
Start point
Yes
Rejected (zr or zl)?
Accepted
No
Yes
No
Item-level adjustment (hit item & Δsr or Δsl)
Attribute-level (sr or sl >1.5)?
Yes
Item-level adjustment (up to 5 items)
Redefine the attributes and/or Reconstruct the Q-matrix
Yes
Rejected (zr or zl)?
No

Figure 1. Suggested Q-matrix validation procedure.

3. If the Q-matrix is rejected (e.g., at 5% significance level), one can evaluate if the concern of attribute-level misspecification exists with the criterion *sr* or $sl > 1.5$. If the concern does not exist, one can identify the hit item based on *mr* or *ml*, and adjust the item (e.g., adding and/or dropping attributes) with the help of Δsr or Δsl. When the Δsr or Δsl is the largest among all possible adjustments, the item is most likely corrected. After that, one can return to the second step above.
4. Even if the concern exists, one should still adjust several (e.g., up to five) misspecified items as above (i.e., with the hit item based on *mr* or *ml* and adjustment based on Δsr or Δsl) due to the suggestive nature of the concern. When adjustments of individual misspecified items tend to be useless, one can redefine the attributes and/or respecify the Q-matrix, and then go back to the second step for another round of validation.

Discussion

Q-matrix validation is of increasing concern due to the significance and subjective tendency of Q-matrix construction in the modeling process. This research proposes a residual-based approach to empirically validate Q-matrix specification based on a combination of fit measures. The approach separates Q-matrix validation into four logical steps, including the test-level evaluation, possible distinction between attribute-level and item-level misspecifications, identification of the hit item, and fit information to aid in item adjustment. Through simulation studies and real-life examples, it is shown that the misspecified items can be detected as the hit item and adjusted sequentially when the misspecification occurs at the item level or at random. Adjustment can be based on the maximum reduction of test-level measures. When adjustment of individual items tends to be useless, attribute-level misspecification is of concern. As the fit measures were more sensitive to under- rather than overspecification, attribute-level adjustment suggests that more attributes are required, or the attributes should be redefined. The statistical power of the approach can be low when the test length is relatively short, and only one multiple-attribute item is misspecified. Good news is that, once the Q-matrix is rejected and attribute-level misspecification is not of concern, the hit item is always misspecified.

Although the DINA model, LLM, and G-DINA model were used in the simulation studies, there is no reason to prevent the applications of the residual-based approach to other reduced or saturated models. Moreover, the approach does not rely on the nature of the attribute and accordingly can be applied to polytomous attributes readily. The correlation-based and LOR-based measures perform similarly, but they can be extended for different purposes. Specifically, the former can be extended to address polytomous responses with an ordinal scale, whereas the latter might be revised for responses with a nominal scale. Besides, the simulation-based measures can be adapted for posterior predictive model checking in the Markov chain Monte Carlo estimation context. In addition to CDMs for dichotomous attributes, the approach can accommodate models for polytomous attributes.

The residual-based measures are insensitive to pure overspecification unless highly constrained models are involved. In this case however, one can argue that pure overspecification is of less concern. With a less constrained model (e.g., LLM), a purely overspecified Q-matrix has more item parameters but only produces slightly higher maximum likelihood than the true Q-matrix (J. J. Chen et al., 2013). Similarly, it was found that the item effects due to the overspecified attributes were trivial. As a result, the impact of pure overspecification on both model misfit and the residuals between the observed and predicted (i.e., model-based) response patterns was generally small.

Although both the applicability and generality appear to be promising, some other concerns or needs should be addressed in future research before the approach is full-fledged. First, more empirical studies are needed on the effectiveness of the fit measures and the procedures of item adjustment based on the maximum reduction of the test-level measures across different conditions such as different number of attributes or magnitudes of attribute relationships. In this regard, the reduction consistency of the fit measures across all possible sequences of hit items may also need to be verified. Second, the current way of item adjustment is time-consuming and tedious, as one needs to try different possible specifications on the hit item for maximum change in each round of adjustment manually. An efficient and exhaustive search algorithm similar to the discrimination-based general method (de la Torre & Chiu, 2015) is desirable. Third, although pure overspecification is of less concern, its detection and adjustment are always desirable. As one reviewer mentioned, overspecification is more likely to occur in practice due to the tendency of specifying more rather than less entries in the Q-matrix. Construction of new fit measures for such purpose, together with investigation of their performance and possible interactions with existing measures, is preferred in future research. Fourth, it would be meaningful to empirically compare the approach to validation methods based on different mechanism. Specifically, the general method based on the discrimination index (de la Torre & Chiu, 2015) can also cover a wide range of CDMs. Based on maximization of the variation of latent group probabilities, the general method can be sensitive to different types of misspecification but requires specific cutoff that needs to be established empirically. Accordingly, it is expected to perform differently from the residual-based approach. It would be useful to see if the two approaches can complement each other in future work. Fifth, there are other sources or factors of model–data misfit not related to the Q-matrix (e.g., nature of the latent variables or inappropriate attribute structures). It is useful to evaluate how the approach can be applied, likely with other fit methods, to provide a full picture of misfit information beyond Q-matrix misspecification. Finally, similar to other validation methods, the adjustment information based on the proposed approach is deemed suggestive and supplemental. Although it can increase the validity of inferences from the test statistically, its substantive meaningfulness remains to be examined by subject experts. In such regard, it is useful to evaluate how statistical and substantive discrepancies can be resolved in practice.

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) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Open Fund of the National Higher Education Quality Monitoring Data Center (Guangzhou), Grant No. L1601.

References

Chen, H., & Chen, J. (2015). Exploring reading comprehension skill relationships through the G-DINA model. *Educational Psychology: An International Journal of Experimental Educational Psychology*, *36*, 1049-1064. doi:10.1080/01443410.2015.1076764

Chen, J., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement*, *37*, 419-437. doi:10.1177/0146621613479818

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, *50*, 123-140.

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association*, *110*, 850-866.

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, *37*, 598-618.

Chiu, C.-Y., Douglas, J., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, *74*, 633-665.

de la Torre, J. (2008). An empirically-based method of Q-matrix validation for the DINA model: Development & applications. *Journal of Educational Measurement*, *45*, 343-362.

de la Torre, J. (2009). DINA model & parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*, 115-130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179-199.

de la Torre, J., & Chiu, C.-Y. (2015). A general method of empirical Q-matrix validation. *Psychometrika*, *81*, 253-273. doi:10.1007/s11336-015-9467-8

de la Torre, J., & Douglas, J. (2004). A higher-order latent trait model for cognitive diagnosis. *Psychometrika*, *69*, 333-353.

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, & the Q-matrix. *Applied Psychological Measurement*, *35*, 8-26.

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-Matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement*, *36*, 447-468.

DiBello, L., Roussos, L., & Stout, W. (2007). Cognitive diagnosis Part I. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics: Vol. 26. Psychometrics* (pp. 979-1030). Amsterdam, The Netherlands: Elsevier.

Doornik, J. A. (2003). Object-oriented matrix programming using Ox (Version 3.1) [Computer software]. London, England: Timberlake.

Hartz, S. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Unpublished doctoral dissertation). University of Illinois at Urbana–Champaign.

Henson, R. A., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*, 191-210.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with non-parametric item response theory. *Applied Psychological Measurement*, *25*, 258-272.

Köhn, H.-F., & Chiu, C.-Y. (2016). A procedure for assessing the completeness of the Q-matrices of cognitively diagnostic tests. *Psychometrika*, *82*, 112-132.

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement*, *36*, 548-564.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*, 187-212.

Organisation for Economic Co-Operation and Development. (1999). *Measuring student knowledge & skills: A new framework for assessment*. Paris, France: Author.

Organisation for Economic Co-Operation and Development. (2006a). *Assessing scientific, reading & mathematical literacy: A framework for PISA 2006*. Paris, France: Author.

Organisation for Economic Co-Operation and Development. (2006b). *PISA released items: Reading*. Retrieved from http://www.oecd.org/pisa/38709396.pdf

Rupp, A. A., & Templin, J. (2008). Effects of Q-matrix misspecification on parameter estimates and misclassification rates in the DINA model. *Educational and Psychological Measurement*, *68*, 78-98.

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.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345-354.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory & cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Safto (Eds.), *Diagnostic monitoring skills and knowledge acquisition* (pp. 453-488). Hillsdale, NJ: Lawrence Erlbaum.

Templin, J., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*, 287-305.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical & Statistical Psychology*, *61*, 287-307.

Xu, G., & Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika*, *81*, 625-649.