Journal of Educational Measurement
Winter 2008, Vol. 45, No. 4, pp. 343–362

An Empirically Based Method of Q-Matrix Validation for the DINA Model: Development and Applications

Jimmy de la Torre
Rutgers, The State University of New Jersey

Most model fit analyses in cognitive diagnosis assume that a Q matrix is correct after it has been constructed, without verifying its appropriateness. Consequently, any model misfit attributable to the Q matrix cannot be addressed and remedied. To address this concern, this paper proposes an empirically based method of validating a Q matrix used in conjunction with the DINA model. The proposed method can be implemented with other considerations such as substantive information about the items, or expert knowledge about the domain, to produce a more integrative framework of Q-matrix validation. The paper presents the theoretical foundation for the proposed method, develops an algorithm for its practical implementation, and provides real and simulated data applications to examine its viability. Relevant issues regarding the implementation of the method are discussed.

Cognitive diagnosis models are developed primarily for the purpose of identifying students' mastery or nonmastery of fine-grained attributes. These models allow assessments to pinpoint students' specific strengths and weaknesses. The specificity of the information derived from cognitive diagnosis models can be used to inform classroom instruction and student learning.

One of the commonly used cognitive diagnosis models is the *deterministic, inputs, noisy "and" gate* (DINA; Junker & Sijtsma, 2001) model. The DINA model is a discrete latent variable model that allows inferences about both the cognitive information of the items and the cognitive attributes of the examinees, and has been the foundation of several approaches to cognitive diagnosis and assessment (e.g., Doignon & Falmagne, 1999; Tatsuoka, 1995). The model is parsimonious in that it requires only two parameters (i.e., slip and guessing) for each item regardless of the number of underlying attributes being considered. Some applications and discussions of the DINA model can be found in de la Torre and Douglas (2004), and Junker and Sijtsma (2001). Although the term DINA model was not explicitly used, Doignon and Falmagne, Haertel (1989), Macready and Dayton (1977), and Tatsuoka (2002) provide additional applications of the model. Like many cognitive diagnosis models, implementation of the DINA model requires construction of a Q matrix (Tatsuoka, 1983) to describe how the test items are related to the attributes. In this matrix, each of the J items is listed in a separate row, and each of the K attributes in a separate column. The Q matrix plays an important role in test development in that it embodies the attribute blueprint or cognitive specifications for test construction (Leighton, Gierl, & Hunka, 2004). Therefore, the Q matrix, and consequently the test, can be designed to provide maximum information about some specific attribute patterns of interest.

Copyright © 2008 by the National Council on Measurement in Education

Although an integral part of the model, many model fit analyses in cognitive diagnosis assume that a Q matrix is correct after it has been constructed, without verifying its appropriateness. Consequently, any model misfit attributable to the Q matrix cannot be addressed and remedied. Thus, it can be said that model fit analysis that does not include verification of the correctness of the Q matrix can only provide a partial and incomplete picture of how a cognitive diagnosis model in its entirety fits the data.

To address this concern, this paper proposes an empirically based method of validating a Q matrix used in a cognitive diagnosis modeling analysis, in particular, the DINA model analysis. This method is intended to complement, not replace, current methods of checking model fit (e.g., residual analysis; de la Torre & Douglas, in press). In contrast to other methods of validating a Q matrix, the proposed method is based solely on the data at hand, and, in its current implementation, does not explicitly take into account substantive information about the items, or expert knowledge about the domain. However, as discussed below, the various considerations in validating a Q matrix can be assimilated and implemented in an integrative framework.

The Sequential EM-Based δ-Method

Rationale

Assume that a set of K latent attributes is relevant to the domain under investigation. Denote the 2^K binary vectors defined by the K attributes as $\boldsymbol{\alpha}_l$, $l = 0, 1, \ldots, 2^K - 1$, and let $\boldsymbol{\alpha}_0$ correspond to the null vector $(0, 0, \ldots\ 0)'$). For the DINA model, the q-vector for item j corresponding to $\boldsymbol{\alpha}_l$ is said to be correctly specified if it maximizes the difference of probabilities of correct response between examinees who have all the required attributes and those who do not. That is, $\boldsymbol{q}_j$ is the correct q-vector if

$$\boldsymbol{q}_j = \arg\max_{\alpha_l} [P(X_j = 1|\eta_{ll'} = 1) - P(X_j = 1|\eta_{ll'} = 0)] = \arg\max_{\alpha_l} [\delta_{jl}], \quad (1)$$

for l, $l' = 1, 2, \ldots, 2^K - 1$, where $\eta_{ll'} = \prod_{k=1}^{K} \alpha_{l'k}^{\alpha_{lk}}$. Because $P(X_j = 1|\eta_{ll'} = 1) = 1 - s_j$ and $P(X_j = 1|\eta_{ll'} = 0) = g_j$. Maximizing the difference in Equation 1 is equivalent to minimizing the sum of the slip and guessing parameters, s_j and g_j, of item j given the data. For this reason, the size of the slip and guessing parameters can be used, albeit informally, to establish the fit of the model to the data. However, it should be noted that obtaining small slip and guessing parameters is a sufficient, but not a necessary condition for establishing model-data fit. In some situations, items may have high guessing or slip parameters given a particular set of the attributes, but no additional model-data fit improvement can be expected unless a different set of attributes is employed. Finally, δ_j, defined above as the difference in the probabilities of correct responses between examinees in groups $\eta_j = 1$ and $\eta_j = 0$, can be regarded as the discrimination index of item j in that items which highly differentiate between these examinees have high δ_j, whereas those that do not have low δ_j. However, δ_j is not an inherent characteristic of the item, but rather a characteristic that can change as the q-vector of the item changes.

Exhaustive search algorithm. To illustrate this point, consider a domain involving $K = 5$ attributes, and an item that requires attributes 1 and 2. Additionally, set $s = g = .20$. Given in Table 1 are the 32 attribute patterns for this example, with the first 24 attribute patterns resulting in $\eta = 0$, and the remaining 8 attribute patterns in $\eta = 1$. The last three columns of Table 1 give the probabilities of correct response by group, and the difference between the two probabilities (i.e., δ) when the corresponding row is used as the q vector. For illustration purposes, these probabilities were computed assuming the patterns are equally likely. For instance, if pattern 17, $(1, 0, \ldots, 0)'$, is used as the q vector of the item, the first 16 attribute patterns will result in $\eta^* = 0$, whereas the last 16 patterns will result in $\eta^* = 1$. The mean of the true probability of the correct response given in column 3 for the first 16 patterns (all of which are .20) is .20, whereas the mean for the remaining patterns (half of which are .20, and the other half are .80) is .50. As can be seen from the table, the difference in the probabilities of correct response between $\eta^* = 0$ and $\eta^* = 1$ is largest when attribute pattern 25, the correct q vector, is used. Omitting one or both of the two required attributes dramatically increases the slip parameter. In addition to omitting required attributes, if unnecessary attributes are made requisite, the guessing parameter also increases. Lastly, including unnecessary attributes on top of the required attributes increases the guessing, but not the slip parameter.

An exhaustive search algorithm that involves computing δ_{jl} for each item is only practicable when K is reasonably small. As K increases, the number of attribute patterns increases exponentially, and the cost for implementing an exhaustive search becomes prohibitive. Thus, a more efficient algorithm that does not require computing δ_{jl} for the $2^K - 1$ possible q vectors needs to be developed.

Sequential search algorithm. Before describing an alternative to the exhaustive search algorithm, additional observations regarding the impact of q-vector misspecification on δ^* need to be made. We can note that the best separation between the true probabilities of correct response (.20 and .80 in the example above) is achieved when all the required attributes are correctly specified. However, when a required attribute is omitted, some of the response patterns whose $P(X = 1|\eta = 0) = .20$ are incorrectly classified with response patterns whose $P(X = 1|\eta = 1) = .80$, which results in $P(X = 1|\eta^* = 1) < .80$. On the other hand, if nonessential attributes are included, some of the response patterns whose $P(X = 1|\eta = 1) = .80$ are incorrectly classified with response patterns whose $P(X = 1|\eta = 0) = .20$, resulting in $P(X = 1|\eta^* = 0) > .20$. Both errors in specification cause shrinkage in δ^*. Finally, the problem is exacerbated when both errors are committed simultaneously. From this observation we can deduce that, for a fixed number of required attributes, the vector with the least number of misspecified attributes has the least amount of shrinkage relative to the optimum δ. Furthermore, if the impact of an attribute inclusion can be isolated δ^* can be used to infer whether or not the attribute is required for the item.

The alternative approach flows directly from these observations. The algorithm starts by comparing δ^* based on the single-attribute q vectors. The attribute prescribed in the q vector that results in the largest $\delta^{(1)}$, say $\alpha^{(1)}$, must be a required

TABLE 1
Probabilities of Correct Response for a Hypothetical Item in a Five-Attribute Domain (Attributes 1 and 2 are Required to Answer the Item Correctly)

	Under True Q Vector		Attributes					Probability of Correct Response		
Pattern	η	$P(X=1\|\eta)$	α_1	α_2	α_3	α_4	α_5	$\eta^*=0$	$\eta^*=1$	δ^*
1	0	0.20	0	0	0	0	0	–	–	–
2	0	0.20	0	0	0	0	1	0.35	0.35	0.00
3	0	0.20	0	0	0	1	0	0.35	0.35	0.00
4	0	0.20	0	0	0	1	1	0.35	0.35	0.00
5	0	0.20	0	0	1	0	0	0.35	0.35	0.00
6	0	0.20	0	0	1	0	1	0.35	0.35	0.00
7	0	0.20	0	0	1	1	0	0.35	0.35	0.00
8	0	0.20	0	0	1	1	1	0.35	0.35	0.00
9	0	0.20	0	1	0	0	0	0.20	0.50	0.30
10	0	0.20	0	1	0	0	1	0.30	0.50	0.20
11	0	0.20	0	1	0	1	0	0.30	0.50	0.20
12	0	0.20	0	1	0	1	1	0.33	0.50	0.17
13	0	0.20	0	1	1	0	0	0.30	0.50	0.20
14	0	0.20	0	1	1	0	1	0.33	0.50	0.17
15	0	0.20	0	1	1	1	0	0.33	0.50	0.17
16	0	0.20	0	1	1	1	1	0.34	0.50	0.16
17	0	0.20	1	0	0	0	0	0.20	0.50	0.30
18	0	0.20	1	0	0	0	1	0.30	0.50	0.20
19	0	0.20	1	0	0	1	0	0.30	0.50	0.20
20	0	0.20	1	0	0	1	1	0.33	0.50	0.17
21	0	0.20	1	0	1	0	0	0.30	0.50	0.20
22	0	0.20	1	0	1	0	1	0.33	0.50	0.17
23	0	0.20	1	0	1	1	0	0.33	0.50	0.17
24	0	0.20	1	0	1	1	1	0.34	0.50	0.16
25	1	0.80	**1**	**1**	**0**	**0**	**0**	**0.20**	**0.80**	**0.60**
26	1	0.80	1	1	0	0	1	0.29	0.80	0.51
27	1	0.80	1	1	0	1	0	0.29	0.80	0.51
28	1	0.80	1	1	0	1	1	0.32	0.80	0.48
29	1	0.80	1	1	1	0	0	0.29	0.80	0.51
30	1	0.80	1	1	1	0	1	0.32	0.80	0.48
31	1	0.80	1	1	1	1	0	0.32	0.80	0.48
32	1	0.80	1	1	1	1	1	0.34	0.80	0.46

attribute. The process continues by investigating q vectors requiring two attributes, with $\alpha^{(1)}$ being one of the two. The attributes in the q vector with the largest $\delta^{(2)}$, $\alpha^{(1)}$ and $\alpha^{(2)}$, are the most relevant attributes. However, unless these two attributes improve on the discrimination of $\alpha^{(1)}$ alone (i.e., $\delta^{(2)} > \delta^{(1)}$), $\alpha^{(2)}$ is deemed unnecessary for the item. In this case, the algorithm terminates and identifies $\alpha^{(1)}$ as the only required attribute. Otherwise, the algorithm proceeds in the same manner investigating q vectors with more required attributes. In general, the process is terminated after step s if $\delta^{(s)} < \delta^{(s-1)}$, or when $s = K$ and the q vector corresponding to max $(\delta^{(s-1)}, \delta^{(s)})$

TABLE 2
Relevant Probabilities of Correct Response Based on the δ-Method

		Sequential Steps				
Number of Attributes	η^*	1	2	3	4	5
One	0	0.20	0.20	0.35	0.35	0.35
	1	0.50	0.50	0.35	0.35	0.35
Two ($\alpha^{(1)} = \alpha_1$)	0	–	0.20	0.30	0.30	0.30
	1	–	0.80	0.50	0.50	0.50
Three ($\alpha^{(1)} = \alpha_1, \alpha^{(2)} = \alpha_2$)	0	–	–	0.29	0.29	0.29
	1	–	–	0.80	0.80	0.80

specifies the correct attributes for the item. Because the algorithm depends on δ to find a solution to the problem, the algorithm will be referred to as the delta (δ) method for validating a Q matrix. For K-attribute domains, the maximum number of δ^* that need to be estimated using the δ-method is $(K^2 + K)/2$, which can be significantly lower than the $2^K - 1$ using the exhaustive search algorithm. The exact number of δ^* that need to be computed using the δ-method is equal to $(K_j + 1)K - (K_j^2 + K_j)/2$, where K_j is the correct number of attributes required for item j.

The example above can be used to demonstrate how the δ-method operates. Table 2 gives the probabilities of correct response for q vectors with up to three attributes based on this algorithm. In the first step, δ^* of q vectors with one attribute are compared. Attributes 1 and 2 have the highest difference of .30, and either of the two attributes can be designated as $\alpha^{(1)}$. Suppose we choose $\alpha^{(1)} = \alpha_1$. The next step compares the q vectors that specify α_1 and another attribute. With $\delta^* = .60$, and $\delta^{(2)} > \delta^{(1)}$, α_2 is chosen as $\alpha^{(2)}$. The following step involves comparing the δ^* of q vectors that include α_1, α_2, and one of the remaining attributes. For this step, $\delta^* < \delta^{(2)}$, which indicates that attributes 3, 4, or 5 do not improve on the discrimination provided by α_1 and α_2. Hence, the algorithm is terminated, and identifies α_1 and α_2 as the required attributes for this item. It can be added that the guessing and slip parameters for this item are computed to be both .20, and the maximum $\delta = .60$. Finally, in this example where $K_j = 2$, the number of δ^* that needed to be examined to find the correct q vector is $(K_j + 1)K - (K_j^2 + K_j)/2 = 12$.

Implementation with Real Data

Complete recalibration solution. In applying the δ-method to real data, two critical differences need to be noted between the hypothetical item above and real test items. First, when real items are involved, δ_j cannot be computed directly because the true guessing and slip parameters, and the exact (multinomial) distribution of the attribute patterns are unknown. Second, a clear cut separation between the groups $\eta_j = 0$ and $\eta_j = 1$ cannot be expected. That is, even with the optimal separation of the attribute patterns using the correct q vector, the probabilities of correct response for the attribute patterns within the same group are not identical. Incidentally, a large variability in the probabilities of correct response within a group, and not the actual magnitude of s_j and g_j, can call into question the appropriateness of the DINA model.

A straightforward but computationally demanding solution that can bypass the first concern is to use a variation of the δ-method that involves a recalibration of the item parameters for each $\boldsymbol{q}_{jl}$ of interest. This solution requires calibrating the data $\sum_{j=1}^{J}[(K_j+1)K-(K_j^2+K_j)]/2$ times assuming the problem can be solved in one pass. This in itself is prohibitive. Consequently, this method becomes an impractical solution to finding the correct q-vectors, particularly when J or K is large.

EM-based solution. Using an implementation of the EM algorithm, de la Torre (in press) showed that the DINA model parameter estimates of item j are estimated based on $R_j^{(\eta_j)}$ and $N_j^{(\eta_j)}$, the posterior expectations of the number of correct responses and examinees, respectively, in group η_j, for $\eta_j = 0, 1$. In particular, it was shown that

$$\hat{g}_j = \frac{R_j^{(0)}}{N_j^{(0)}} \text{ and } \hat{s}_j = \frac{N_j^{(1)} - R_j^{(1)}}{N_j^{(1)}}.$$

Therefore, $\hat{g}_{jl}$ and $\hat{s}_{jl}$, the guessing and slip parameters of item j corresponding to any $\boldsymbol{q}_{jl}$, and hence, $\hat{\delta}_{jl}$, can be easily computed given the estimated posterior distribution $\hat{p}(\boldsymbol{\alpha}_i|X_i)$, $i = 1, \ldots, N$, and the data matrix $\mathbf{X}$. Without recalibrating the data multiple times, the EM-based solution to the δ-method dramatically reduces the computational requirements in estimating the relevant $\hat{\delta}_{jl}$.

We should not lose sight of the fact that $\hat{\delta}_{jl}$ is based on $\hat{p}(\boldsymbol{\alpha}_i|\boldsymbol{X}_i)$, and not $p(\boldsymbol{\alpha}_i \mid \boldsymbol{X}_i)$, and the extent to which the former can be used in the latter's stead depends on several factors that include the accuracy of the item parameter estimates. As shown earlier, q-vector misspecifications affect the quality of item parameter estimates. Because the item parameters are employed in computing the posterior distribution, q-vector misspecifications also can degrade the quality of the approximation to the posterior distribution.

The use of the estimated posterior distribution, plus the fact that even the correct q vector may not produce a clear cut separation between the groups $\eta_j = 0$ and $\eta_j = 1$ when dealing with real data, can cause $\hat{\delta}^{(K_j+1)} > \hat{\delta}^{(K_j)}$ (i.e., a recommended q vector may have more attributes than necessary). A possible solution to this problem is to use the criterion $\hat{\delta}_j^{(s)} - \hat{\delta}_j^{(s-1)} > \varepsilon$, where ε is a cut-off point representing the minimum increment in the discrimination index of the item resulting from an additional attribute for it to be deemed relevant. A small ε (e.g., $\varepsilon = 0$) is considered more liberal allowing more attributes to be specified, whereas a larger ε, say, $\varepsilon = .2$, is more stringent, making it difficult to build up the q vector. Using liberal and stringent criteria can result in over- and underspecified q vectors, respectively. To address this problem, several values of ε can be used, and their solutions compared in terms of how the different q vectors affect the entire test. For example, the mean slip and guessing parameters across all the items obtained using the different recommended Q matrices can be compared to determine which of the possible solutions is optimal. The comparison can be based on the updated estimates of the item parameters that can be obtained easily by running a few additional cycles of the EM algorithm based on the candidate Q matrices.

Finally, two important details about the Q matrix need to be noted in using the proposed method in practice. First, the Q matrix at hand represents just one of the possible Q matrices that can be used to fit the data. Therefore, it is possible that some Q matrices that provide equally good or better fits may exist. However, one needs to take into account the fact that as the collection of attributes or the Q-matrix specification changes, so do the definitions and interpretations of the attributes. Thus, the best fitting Q matrix may not necessarily be interpretable, hence, be of little practical value. Second, one should remember that enormous time and resources are typically invested in constructing a Q matrix. Thus, without strong evidence to suggest otherwise, one should err on being more conservative (i.e., be willing to acknowledge the correctness of the Q-matrix specification). With these in mind, one can be more judicious in implementing the proposed method, particularly when subjective judgments are required.

Simulation Study

Design and Analysis

To investigate whether the sequential EM-based δ-method can work under the ideal condition, simulated data with $K = 5$ and $J = 30$ were generated. The Q matrix for these data is given in Table 3, and was constructed to have an equal number of 1-, 2-, and 3-attribute items. Also, the attributes in this Q matrix appear an equal number of times. The attribute patterns were assigned equal probability, and a sample size of $N = 5{,}000$ was used in generating the data. Last, the slip and guessing parameters were set to .20 for all items.

In addition to the correct Q matrix, 10 Q matrices each with single misspecified q-vector and one Q matrix with three misspecified q vectors were also used to analyze the data. The misspecifications of the q vectors were applied to items 1, 11, and 21, which represent 1-, 2-, and 3-attribute items, respectively. The q-vector misspecifications are summarized in Table 4. The item parameters for each Q matrix were estimated using an empirical Bayesian implementation of the EM algorithm with a convergence criterion of .001. The codes used in implementing the EM algorithm and the sequential EM-based δ-method were written in Ox (Doornik, 2003).

Results

As expected, the parameter estimates showed little bias (i.e., the mean and maximum absolute biases were .01 and .04, respectively) when the correct Q matrix was used. The mean discrimination index was $\delta = .61$. In contrast, the parameter estimates for the items with misspecified q vectors, given in Table 5, show large biases, and shrunken δ. In addition, although not presented in the table, some correctly specified items were affected by the q-vector misspecification. However, the impact was not considerable so that, as a whole, the mean guessing and slip parameter estimates across the items with correct q vectors were close to the generating parameters. The estimates in Table 5 also show that the directions of the biases were consistent with the discussion above concerning the impact of the omission and inclusion of relevant and irrelevant attributes on item parameter estimates. Finally, the sum of

TABLE 3
Q Matrix for the Simulated Data

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	0	0	0	0
7	0	1	0	0	0
8	0	0	1	0	0
9	0	0	0	1	0
10	0	0	0	0	1
11	**1**	**1**	**0**	**0**	**0**
12	1	0	1	0	0
13	1	0	0	1	0
14	1	0	0	0	1
15	0	1	1	0	0
16	0	1	0	1	0
17	0	1	0	0	1
18	0	0	1	1	0
19	0	0	1	0	1
20	0	0	0	1	1
21	**1**	**1**	**1**	**0**	**0**
22	1	1	0	1	0
23	1	1	0	0	1
24	1	0	1	1	0
25	1	0	1	0	1
26	1	0	0	1	1
27	0	1	1	1	0
28	0	1	1	0	1
29	0	1	0	1	1
30	0	0	1	1	1

the mean guessing and slip parameters are also given in the table. For a systematic investigation of the impact of the q-vector misspecification on parameter estimation and attribute classification, refer to Rupp and Templin (2008).

In implementing the sequential EM-based δ-method, five cut-off points ($\varepsilon = .00$, .01, .05, .10, .20) were used to select the candidate q vectors, and five additional EM cycles were employed in comparing the resulting Q matrices. For each condition and ε the results given in Table 6 show the number of q vectors changed after running the procedure, the proposed q vector for the misspecified items, and the sum of mean guessing and slip parameter estimates. As expected, the number of attributes specified using a more liberal ε is greater than or equal to the number of attributes specified using a larger ε for all conditions. The results for $\varepsilon = .20$ indicate that this

TABLE 4
Summary of Q-Vector Misspecifications

Condition	Item Altered	Alterations Alterations	Number Added	Number Deleted	Net Change	Number of Alterations
0	0	None	0	0	0	0
1	1	$\alpha_1 \rightarrow 0, \alpha_2 \rightarrow 1$	1	1	0	2
2	1	$\alpha_2 \rightarrow 1$	1	0	1	1
3	11	$\alpha_1 \rightarrow 0, \alpha_3 \rightarrow 1$	1	1	0	2
4	11	$\alpha_1 \rightarrow 0$	0	1	−1	1
5	11	$\alpha_3 \rightarrow 1$	1	0	1	1
6	21	$\alpha_1 \rightarrow 0$	0	1	1	1
7	21	$\alpha_1 \rightarrow 0, \alpha_2 \rightarrow 0$	0	2	−2	2
8	21	$\alpha_1 \rightarrow 0, \alpha_4 \rightarrow 1$	1	1	0	2
9	21	$\alpha_1 \rightarrow 0, \alpha_2 \rightarrow 0$ $\alpha_4 \rightarrow 1$	1	2	−1	2
10	21	$\alpha_1 \rightarrow 0, \alpha_2 \rightarrow 0$ $\alpha_4 \rightarrow 1, \alpha_5 \rightarrow 1$	2	2	0	4
11	1	$\alpha_2 \rightarrow 1$	3	2	1	5
	11	$\alpha_2 \rightarrow 0, \alpha_3 \rightarrow 1$				
	21	$\alpha_1 \rightarrow 0, \alpha_4 \rightarrow 1$				

TABLE 5
Parameter Estimates of Items with Misspecified Q Vectors

		Estimate			
Condition	Item	g	s	δ	$\bar{\hat{g}} + \bar{\hat{s}}$
0	–	–	–	–	0.3924
1	1	0.1807	0.1984	0.6208	0.4139
2	1	0.4850	0.4878	0.0273	0.4008
3	11	0.3908	0.1918	0.4174	0.4058
4	11	0.3093	0.4789	0.2118	0.4019
5	11	0.1976	0.4686	0.3338	0.3968
6	21	0.2966	0.1867	0.5167	0.4028
7	21	0.1948	0.4688	0.3364	0.4083
8	21	0.1936	0.6362	0.1702	0.4044
9	21	0.2445	0.4704	0.2850	0.4104
10	21	0.2560	0.6554	0.0885	0.4105
	1	0.2654	0.6347	0.1000	
11	11	0.1807	0.1984	0.6208	0.4260
	21	0.4850	0.4878	0.0273	

criterion is too stringent, and produced new Q matrices with about a third of the vectors changed even for the condition where all the q vectors were correctly specified. Consequently, with more incorrect q vectors, the mean guessing and slip parameters across the 30 items based on the q vectors when $\varepsilon = .20$ were worse than those based

TABLE 6
Results of the δ-Method Analysis for the Simulated Data (Numbers in Parenthesis are Negative)

Condition	Misspecified Item	ε	# Vectors Changed	Q Vector for Item in Col. 2					$\bar{\hat{g}} + \bar{\hat{s}}$
				α_1	α_2	α_3	α_4	α_5	
		0.00	0	–	–	–	–	–	0.3927
		0.01	0	–	–	–	–	–	0.3927
0	–	0.05	0	–	–	–	–	–	0.3927
		0.10	0	–	–	–	–	–	0.3927
		0.20	10	–	–	–	–	–	0.5367
		0.00	1	X					0.3930
		0.01	1	X					0.3930
1	1	0.05	1	X					0.3930
		0.10	1	X					0.3930
		0.20	11	X					0.5371
		0.00	1	X					0.3929
		0.01	1	X					0.3929
2	1	0.05	1	X					0.3929
		0.10	1	X					0.3929
		0.20	11	X					0.5373
		0.00	1	X	X				0.3927
		0.01	1	X	X				0.3927
3	11	0.05	1	X	X				0.3927
		0.10	1	X	X				0.3927
		0.20	11	X	X				0.5367
		0.00	1	X	X				0.3927
		0.01	1	X	X				0.3927
4	11	0.05	1	X	X				0.3927
		0.10	1	X	X				0.3927
		0.20	10		X				0.5464
		0.00	0	X	X	X			0.3966
		0.01	0	X	X	X			0.3966
5	11	0.05	1	X	X				0.3928
		0.10	1	X	X				0.3928
		0.20	11	X	X				0.5369
		0.00	1	X	X	X	X	X	0.3933
		0.01	1	X	X	X	X	X	0.3933
6	21	0.05	1	X	X	X			0.3927
		0.10	1	X	X	X			0.3927

on the initial q vectors. For this reason, this value can be discounted as a reasonable cut-off point.

The remaining values of ε yielded identical and correct recommendations for conditions 0 through 4. Compared to the initial estimates, the new Q matrices provided smaller $\bar{\hat{g}} + \bar{\hat{s}}$. It is worth noting that even with very stringent ε, the proposed method correctly replaced the misspecified vector in conditions 1, 2, and 3. However, because

TABLE 6
(Continued)

Condition	Misspecified Item	ε	# Vectors Changed	Q Vector for Item in Col. 2					$\bar{\hat{g}} + \bar{\hat{s}}$
				α_1	α_2	α_3	α_4	α_5	
		0.20	10			X			0.5366
		0.00	1	X	X	X	X	X	0.3933
		0.01	1	X	X	X	X	X	0.3933
7	21	0.05	1	X	X	X			0.3927
		0.10	1	X	X	X			0.3927
		0.20	9			X			0.5361
		0.00	1	X	X	X	X	X	0.3933
		0.01	1	X	X	X	X	X	0.3933
8	21	0.05	1	X	X	X			0.3927
		0.10	1	X	X	X			0.3927
		0.20	10			X			0.5366
		0.00	1	X	X	X	X	X	0.3932
		0.01	1	X	X	X	X	X	0.3932
9	21	0.05	1	X	X	X			0.3926
		0.10	1	X	X	X			0.3926
		0.20	10			X			0.5365
		0.00	1	X	X	X	X	X	0.3933
		0.01	1	X	X	X	X	X	0.3933
10	21	0.05	1	X	X	X			0.3927
		0.10	1	X	X	X			0.3927
		0.20	10			X			0.5366
	1			X					
	11	0.00	3	X	X				0.3935
	21			X	X	X	X	X	
	1			X					
	11	0.01	3	X	X				0.3935
	21			X	X	X	X	X	
	1			X					
11	11	0.05	3	X	X				0.3938
	21			X	X	X	X		
	1			X					
	11	0.10	3	X	X				0.3929
	21			X	X	X			
	1			X					
	11	0.20	12	X					0.54649
	21					X			

other items were also affected by the stringent ε the overall quality of item parameter estimates was lower.

Two q vectors, one of which is the correct q vector, were recommended by the proposed method for conditions 5 through 10. In updating the item parameter estimates

and posterior distributions using these recommended q vectors, the correct q vectors consistently produced lower $\bar{\tilde{g}} + \bar{\tilde{s}}$ compared to the alternative q vectors. Thus, the δ-method ultimately replaced the altered q vectors with the correct specifications for all these conditions. For condition 11, which involved more incorrectly specified q vectors, the procedure recommended the same q vectors for items 1 and 11 when $\varepsilon < .20$, but not for item 21. However, in comparing the updated item parameter estimates, $\bar{\tilde{g}} + \bar{\tilde{s}}$ was smallest when $\varepsilon = .10$. As in the previous condition, the solution based on this criterion corresponds to the correct solution. In summary, the results above demonstrate that, for the conditions investigated in this simulation study, the sequential EM-based δ-method can correctly replace misspecified q vectors with the appropriate q vectors, while simultaneously retaining q vectors that have been correctly specified.

Application 1: Fraction-Subtraction Data

Data and Analysis

The sequential EM-based δ-method was applied to a subset of the fraction-subtraction data collected and analyzed by Tatsuoka (1990), and more recently by de la Torre and Douglas (2004) and Tatsuoka (2002). The data analyzed in this paper consisted of 2,144 middle school students responding to 15 fraction-subtraction items that measured the following five attributes: (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. Table 7 gives the items, Q matrix, and item parameter estimates for this analysis. The same Q matrix was used by Mislevy (1996) in analyzing the same data. In implementing the proposed method, the cut-off point was set from .000 to .050, with an increment of .001, and 100 EM cycles were added to obtain the final parameter estimates. Similarly, the estimation algorithms were implemented in Ox (Doornik, 2003).

Results

For the initial analysis, $\bar{\tilde{g}} + \bar{\tilde{s}} = .2461$, indicating a reasonable model-data fit. Except for five parameters, the estimates are all less than or equal to .20. The values of $\bar{\tilde{g}} + \bar{\tilde{s}}$ as a function of ε, and the proportions of correspondence between the original and proposed Q matrices are given in Figures 1 and 2, respectively. The results indicate that for cut-off points .009, .010, .011, and .012, the proposed Q matrix was identical to the original Q matrix, and the corresponding $\bar{\tilde{g}} + \bar{\tilde{s}}$ was the minimum. Consequently, the current definitions and specifications of the attributes can be retained and used.

Finally, to underscore the insufficiency of decisions solely based on a statistical index, and the importance of substantive knowledge and interpretation in determining the most appropriate Q matrix, attribute 5, converting whole number to fraction, was specified for item 1, $\frac{3}{4} - \frac{3}{8}$. Although the attribute is not relevant to the item from a substantive point of view, the resulting $\bar{\tilde{g}} + \bar{\tilde{s}} = .2379$, was lower than the optimal $\bar{\tilde{g}} + \bar{\tilde{s}}$ using the original specification.

TABLE 7
Fraction-Subtraction Item: Q-Matrix and Item Parameter Estimates

Item		Attribute					Estimate	
		1	2	3	4	5	$\hat{g}$	$\hat{s}$
1.	$\frac{3}{4} - \frac{3}{8}$	1	0	0	0	0	0.00	0.28
2.	$3\frac{1}{2} - 2\frac{3}{2}$	1	1	1	1	0	0.21	0.12
3.	$\frac{6}{7} - \frac{4}{7}$	1	0	0	0	0	0.14	0.04
4.	$3 - 2\frac{1}{5}$	1	1	1	1	1	0.12	0.13
5.	$3\frac{7}{8} - 2$	1	0	1	0	0	0.33	0.25
6.	$4\frac{4}{12} - 2\frac{7}{12}$	1	1	1	1	0	0.03	0.23
7.	$4\frac{1}{3} - 2\frac{4}{3}$	1	1	1	1	0	0.07	0.08
8.	$\frac{11}{8} - \frac{1}{8}$	1	1	0	0	0	0.16	0.05
9.	$3\frac{4}{5} - 3\frac{2}{5}$	1	0	1	0	0	0.08	0.06
10.	$2 - \frac{1}{3}$	1	0	1	1	1	0.17	0.07
11.	$4\frac{5}{7} - 1\frac{4}{7}$	1	0	1	0	0	0.10	0.10
12.	$7\frac{3}{5} - \frac{4}{5}$	1	0	1	1	0	0.03	0.13
13.	$4\frac{1}{10} - 2\frac{8}{10}$	1	1	1	1	0	0.13	0.16
14.	$4 - 1\frac{4}{3}$	1	1	1	1	1	0.02	0.20
15.	$4\frac{1}{3} - 1\frac{5}{3}$	1	1	1	1	0	0.01	0.18

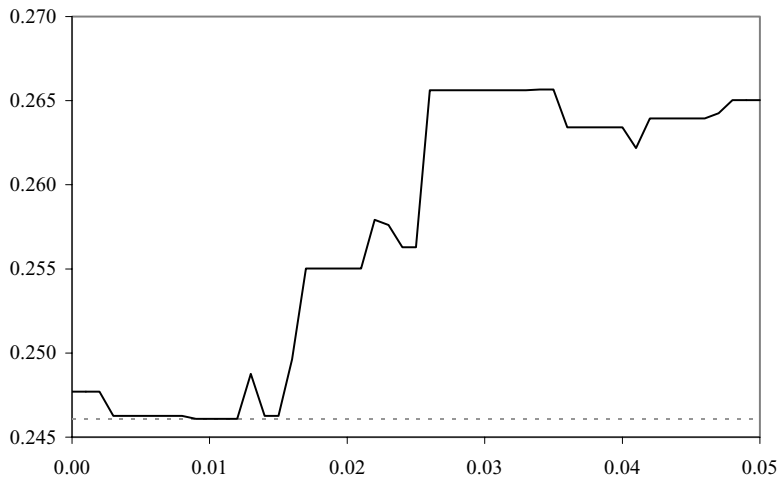


FIGURE 1. *Fraction-subtraction data: $\bar{\hat{g}} + \bar{\hat{s}}$ as a function of ε (original value represented by dotted line).*

Application 2: Analysis of 2003 NAEP Grade 8 Mathematics Data

Data and Analysis

The proposed method of empirically validating a Q matrix was also applied to the 2003 NAEP 8th grade mathematics data. The original Q matrix for this assessment

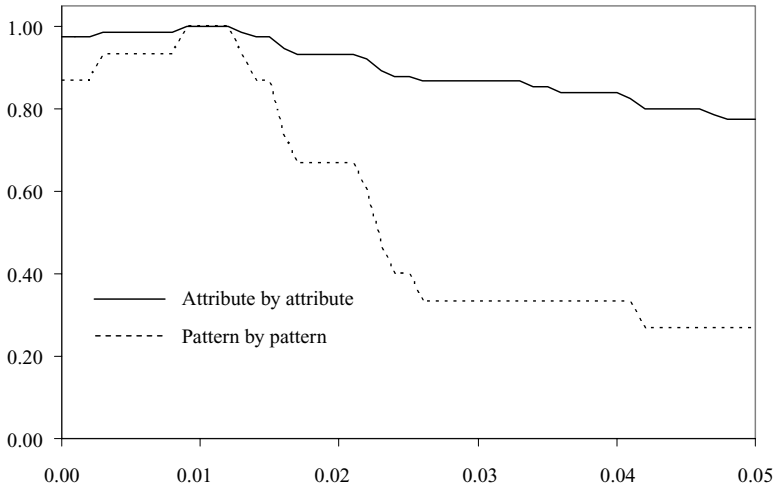


FIGURE 2. *Fraction-subtraction data: Proportion of correspondence between original and new Q matrices.*

TABLE 8
Attribute Descriptions and the Number of Times Each was Specified

Attribute	Description	Number of Times Specified
1	Calculator	15
2	Measure, Units, Conversion	17
3	Data Displays	15
4	Geometry	27
5	Fractions	13
6	Arithmetic Operations in Context	22
7	Interpolation, Extrapolation, Estimation	14
8	Spatial Perception	16
9	Higher-Order Thinking	42

was developed by L. Di Bello and colleagues (personal communication, October 6, 2005), and involves 195 multiple-choice and constructed-response items, and 17 attributes. A subset of the data, which had 90 items measuring 9 attributes, and 3,823 examinees from Texas, was used. Examinees were included if they had responses to at least 12 of the items. The minimum number of examinees responding to an item was 419, whereas the mean number was about 794. Described in Table 8 are the 9 attributes used in this analysis, and the number of times they were required. The algorithms above were modified to take into account the differential weights of the observations and missing data. The data were analyzed as binary responses. Last, for the sequential EM-based δ-method, $\varepsilon = .00, .001, \ldots, .100$, 10 EM cycles were added to update the item parameter estimates and posterior distributions. As in the previous sections, the codes were implemented in Ox (Doornik, 2003).

TABLE 9
Descriptive Statistics of Initial Item Parameter Estimates

Statistic	$\hat{g}$	$1-\hat{s}$	$\hat{\delta}$
Minimum	0.0338	.2613	−.0004
Mean	0.5022	.8099	.3077
Maximum	0.9786	1.0000	.6407

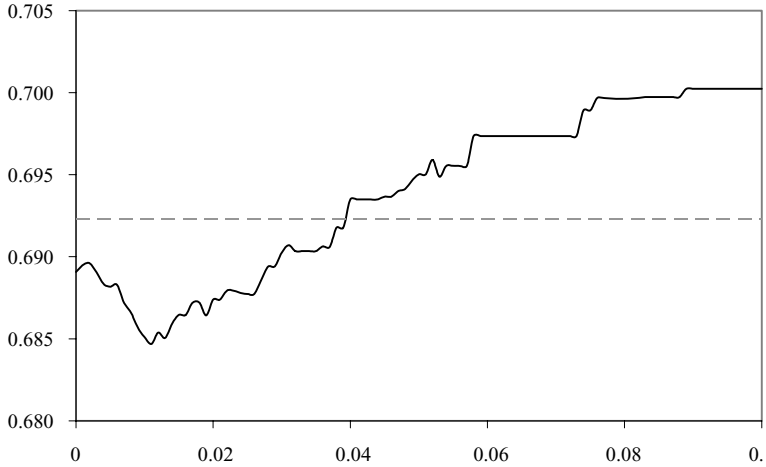


FIGURE 3. *NAEP data: $\bar{\hat{g}}+\bar{\hat{s}}$ as a function of ε (original value represented by dotted line).*

Results

Initial estimates. Table 9 gives a summary of the item parameter estimates using the original Q matrix. In contrast to the fraction-subtraction data, the values in Table 9 indicate that the misfit between the model and the data is very apparent. However, the analysis was continued for purposes of illustration. The second column shows that, based on the attributes employed in the assessment, examinees who did not have the prescribed attributes for the items can correctly answer the items about 50% of the time, on the average. In addition, one item gave students who lack the attribute for that item almost a sure chance, 98%, of getting the item right. The third column shows that the slip parameters are reasonable, with an average estimate of about .19. However, at least one item was considered extremely difficult in that only about 1 in 4 students who have the necessary attributes for that item got it right. Finally, the last column reveals that for at least one item, those who are in group $\eta = 0$ have a slightly higher probability of getting the item right compared to their counterparts in group $\eta = 1$ (i.e., δ is negative). Finally, the mean discrimination index of .31 can be considered rather low.

Test-level comparison. Figures 3 and 4 give the plots of $\bar{\hat{g}}+\bar{\hat{s}}$, and the proportions of attributes and attribute patterns in the proposed Q matrices identical to the original Q matrix as a function of ε. Figure 3 shows that $\bar{\hat{g}}+\bar{\hat{s}}$ initially generally decreases

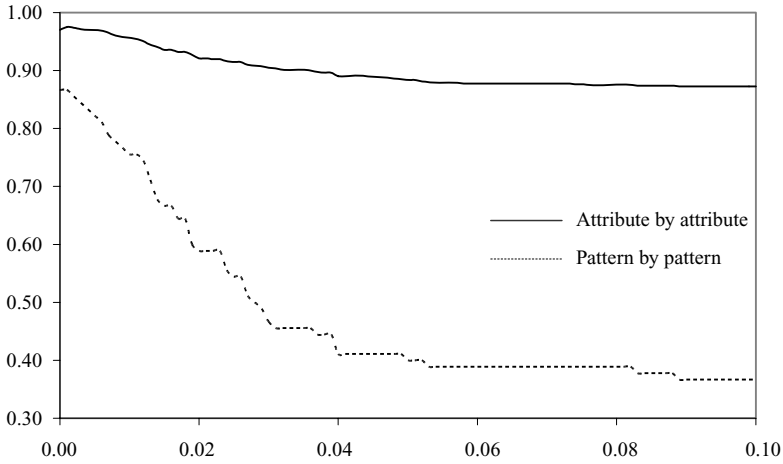


FIGURE 4. *NAEP data: Proportion of correspondence between original and new Q matrices.*

as ε increases; however, after a certain point (i.e., $\varepsilon = .011$), $\bar{\hat{g}} + \bar{\hat{s}}$ shows an upward trend as ε increases. Figure 4 reveals that the proportion of attributes in the proposed Q matrix unchanged from the original Q matrix is generally smaller with larger ε. The proportion of unchanged attributes levels off at about .88 (713 out of 810 attributes), whereas the proportion of unchanged attribute patterns levels off at about .37 (33 out of 90 attribute patterns). Based on this analysis, the optimal solution can be found when $\varepsilon = .011$, and at this cut-off point $\bar{\hat{g}} + \bar{\hat{s}} = .6847$, which is lower than the original .6923. Moreover, 773 attributes (95%) and 68 attribute patterns (76%) in the new Q matrix agree with the original Q matrix. If maximum correspondence between the proposed and original Q matrix is desired, the solution when $\varepsilon = .001$ can be used. It has 98% and 87% attributes and attribute pattern agreements, respectively, and its $\bar{\hat{g}} + \bar{\hat{s}} = .6895$ still represents an improvement over the original estimate.

Item-level comparison. For $\varepsilon = .001$ and $\varepsilon = .011$, the numbers of attributes in agreement with the original Q matrix are 78 and 68, respectively. Figure 5 gives the improvement of the item parameters estimated using the new Q matrices obtained for $\varepsilon = .001, .011$ relative to those estimated using the original Q matrix. For each cut-off point, the differences are arranged according to their magnitude. The plot indicates that about 10 items showed obvious deteriorations using the new Q matrices. However, most of the items exhibited improvements (i.e., positive changes). For $\varepsilon = .001$, most of the improvements were between .00 and .01, whereas the improvements for $\varepsilon = .011$ were distributed mostly between .00 and .02. Only a few items showed somewhat large improvements for both cut-off points. Although the gains using the proposed method were modest for this particular data set, the examples below show how the results can be instructive in validating the original Q matrix. The optimal solution (i.e., $\varepsilon = .011$) is used in discussing these examples.

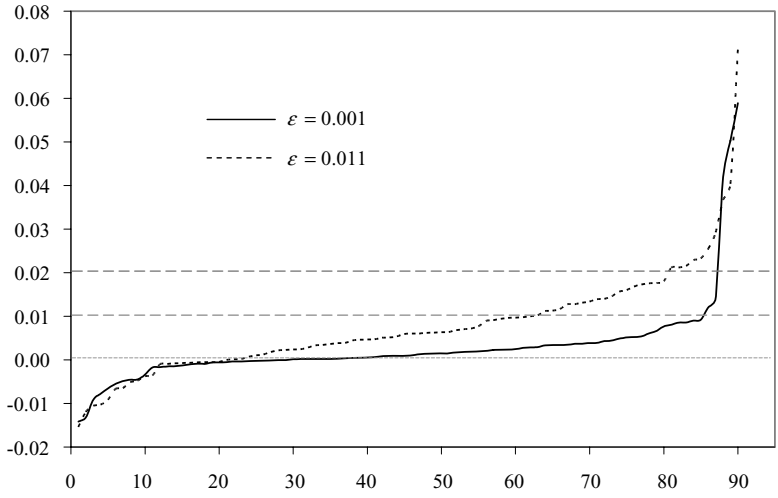
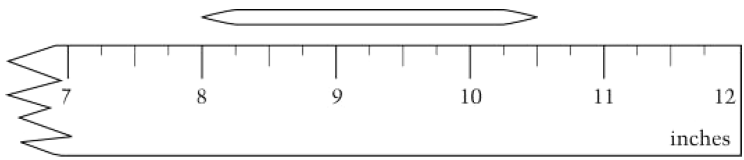


FIGURE 5. $(s^* + g^*) - (s^{(0)} + g^{(0)})$ *for* $\varepsilon = .001, .011$.

First, the proposed q vectors can provide evidence to validate the original q vectors. Item 45 (given in Figure 6) can be used to illustrate this point. The q vector based on the solution confirmed the original specification that prescribed attributes 2 (measure, units, conversion), 3 (data displays), 5 (fractions), and 7 (interpolation, extrapolation, estimation) for this item.

Second, the q vectors can call into question the validity of attribute specifications. For item 80, the original specification requires examinees to possess attributes 6, 7, and 9, and the resulting $\hat{g}$ and $\hat{s}$ are .31 and .50, respectively. However, using a q vector that required attribute 6 only, the guessing parameter of the item dropped to .24. Thus, the alternative specification questions the necessity of requiring the other two attributes for this problem.

Third, the proposed q vectors can suggest alternative attribute specifications. Item 52 in Figure 7 originally required 2 attributes: 2 (measure, units, conversion) and 7 (interpolation, extrapolation, estimation), where $\hat{g} = .72$ and $\hat{s} = .07$. The proposed q vector prescribed attribute 9 (higher-order thinking) instead of attribute 7, but provided item parameter estimates that were practically identical to the previous estimates. Therefore, it is legitimate to ask which of the two attributes better explains the examinees' performance on item 52.



What is the length of the toothpick in the figure above?

FIGURE 6. *Item 45 of the analyzed data.*

The Breakfast Barn bought 135 dozen eggs at $0.89 per dozen. What was the total cost of the eggs?

A) $116.75

B) $120.15

C) $135.89

D) $151.69

FIGURE 7. *Item 52 of the analyzed data.*

Finally, the proposed method can confirm the lack of informativeness of an item. Item 19 has $\delta \approx 0$ (i.e., $P(X = 1|\eta = 0) \approx P(X = 1|\eta = 1)$). For this particular problem, both the probabilities of a correct response were .54. Items with close to zero discrimination index are not necessarily uninformative—they could have been misspecified. However, in this instance the analysis confirmed that the item had been correctly specified, and that the initial δ estimate represents the best discrimination the item can provide. Therefore, after applying the proposed method, it is be safe to conclude that item 19 is diagnostically uninformative under the current set of attributes.

Summary and Conclusion

The appropriateness of the Q matrix, which is part of a cognitive diagnosis model specification, is often overlooked in model-data fit analysis. Consequently, model misfits due to an inappropriate Q matrix cannot be detected and dealt with. This article proposed the sequential EM-based δ-method, an empirically based procedure of validating a Q matrix, to address this concern. The purpose of the method is to improve model-data fit by selecting the optimal q vectors. Moreover, it can provide information that can be useful in re-evaluating a Q matrix.

Results from the simulation study, which used a variety and degree of q-vector misspecification, indicated that the proposed method is potentially viable. In particular, the method was able to identify and correctly replace inappropriate q vectors, while at the same time retain those which were correctly specified. In other words, the Type I and II errors of method are both zero, at least for the conditions investigated in the simulation study.

The method was applied to subsets of fraction-subtraction data and the 2003 NAEP 8th grade mathematics assessment data. The results for the fraction-subtraction data indicated that the proposed method can recognize and retain appropriately specified q vectors. Results for the NAEP data showed that the proposed method yielded parameter estimates with modest improvements. However, the information contained in the recommended q vectors can be useful in establishing or repudiating q-vector specifications, suggesting alternative specifications, and confirming noninformative items.

The sequential EM-based δ-method is a tool that provides statistical information about the Q matrix. As such, it only addresses an aspect of Q-matrix validation, and is intended to supplement, rather than supplant existing methods. A more complete process of Q-matrix validation requires utilizing both statistical information,

and substantive knowledge and domain expertise. For example, the proposed method can be used in conjunction with other methods of attribute and Q-matrix validations that use written and verbal protocols from students and experts (Leighton, 2004; Tatsuoka, Corter, & Tatsuoka, 2004; Wang & Gierl, 2007). As demonstrated by the fraction-subtraction example, decisions based purely on statistical information can be misleading (i.e., they can result in attribute specifications that run counter to substantive knowledge). Thus, for a successful implementation of Q-matrix validation, and cognitive diagnosis modeling for that matter, collaboration between experts from various fields cannot be overemphasized. Lastly, it should be reiterated that Q-matrix validation is a small but significant component of model-data fit analysis.

This research represents an initial step in understanding how a Q matrix can be validated empirically, and much work remains to be done in this area. For one, more conditions (e.g., degree of Q-matrix misspecifications, test length, sample size) need to be investigated to establish the viability of the method across a wide range of situations. In addition, more real data that cover broader domains (e.g., language testing) need to be analyzed to gain additional insights on how the method works in practice. Also, the method is based on a particular statistic (δ); other statistics may be more appropriate or useful in validating Q matrices. Moreover, most applications of cognitive diagnosis are implemented with the number of attributes K assumed to be known, and this method is developed in line with this assumption. The method can be generalized by relaxing this assumption, and including the determination of the correct number of attributes as part of the validation process. Finally, the method is developed primarily for the DINA model. It would be interesting to see how the concept developed here can be extended to other cognitive diagnosis models.

Acknowledgment

Support for this project was provided by the U.S. Department of Education, National Center for Education Statistics through a NAEP Secondary Analysis grant (R902B050007).

References

de la Torre, J. (in press). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*.

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*, 333–353.

de la Torre, J., & Douglas, J. (in press). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*.

Doignon, J. P., & Falmagne, J.-C. (1999). *Knowledge spaces*. New York: Springer-Verlag.

Doornik, J. A. (2003). *Object-oriented matrix programming using Ox (version 3.1)* [Computer software]. London: Timberlake Consultants Press.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 333–352.

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.

Leighton, J. (2004). Avoid misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, *23*(2), 6–15.

Leighton, J. P., Gierl, M. J., & Hunka, S. (2004). The attribute hierarchy model: An approach for integrating cognitive theory with assessment practice. *Journal of Educational Measurement*, *41*, 205–236.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, *2*, 99–120.

Mislevy, R.J. (1996). Test theory reconceived. *Journal of Educational Measurement*, *33*, 379–416.

Rupp, 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 and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Safto (Eds.), *Monitoring skills and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Erlbaum.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327–359). Hillsdale, NJ: Erlbaum.

Tatsuoka, K. K., Corter, J., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMMS-R across a sample of 20 countries. *American Educational Research Journal*, *41*, 901–926.

Wang, C., & Gierl, M. (2007, April). *Investigating the cognitive attributes underlying student performance on the SAT® critical reasoning subtest: An application of the attribute hierarchy method*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Author

JIMMY DE LA TORRE is an Assistant Professor of Educational Psychology at Rutgers University, 10 Seminary Place, New Brunswick, NJ, 08901; j.delatorre@rutgers.edu. His primary research interests include item response theory, cognitive diagnosis, Bayesian analysis, and the use of diagnostic assessments to support classroom instruction and learning.