Category-Level Q-matrix Refinement for Polytomous Cognitive Diagnosis: A Relative Fit Statistics Perspective

Authors: Wang Daxun, Gao Xuliang, Cai Yan, Tu Dongbo, Dongbo Tu

Date: 2019-09-16T00:00:00+00:00

Abstract

The development of polytomous cognitive diagnosis models plays an important role in the advancement of cognitive diagnosis, but research on Q-matrix validation under polytomous models remains to be explored. This study attempts to investigate Q-matrix validation for polytomous cognitive diagnosis, and focuses on the more diagnostically valuable item-category-level Q-matrix validation. Relative fit statistics are applied to Q-matrix validation in polytomous cognitive diagnosis and compared with the existing stepwise method (Ma & de la Torre, 2019). The results show that: the BIC method yields higher pattern and attribute classification accuracy rates for Q-matrix validation in polytomous cognitive diagnosis models, its Q-matrix recovery rate is also higher than that of the stepwise method, and the Q-matrix revised by the BIC method fits the data better; in complex models, the relative fit index BIC performs better than AIC and -2LL; in practice, users can select the BIC method for test Q-matrix validation; Q-matrix validation effectiveness is affected by sample size, and increasing the number of examinees can improve the accuracy of Q-matrix validation. In summary, this study provides important methodological support for Q-matrix validation in polytomous cognitive diagnosis.

Full Text

Category-Level Q-Matrix Validation for Polytomous Cognitive Diagnosis Models: A Relative Fit Statistics Perspective

WANG Daxun[1], GAO Xuliang[2], CAI Yan[1], TU Dongbo[1]
[1]School of Psychology, Jiangxi Normal University, Nanchang, 330022
[2]School of Psychology, Guizhou Normal University, Guiyang, 550000

Abstract

The development of polytomous cognitive diagnosis models plays a crucial role in advancing cognitive diagnosis assessment. However, research on Q-matrix validation under polytomous models remains limited. This study investigates Q-matrix validation for polytomous cognitive diagnosis models, focusing on the more diagnostically valuable category-level Q-matrix validation approach. Relative fit statistics were applied to polytomous cognitive diagnosis Q-matrix validation and compared with the existing stepwise method (Ma & de la Torre, 2019). The findings demonstrate that the BIC method achieves high pattern match rates and attribute match rates for Q-matrix validation in polytomous cognitive diagnosis models, with superior Q-matrix recovery compared to the stepwise method. Q-matrices corrected by the BIC method show better fit to the data. In complex models, the relative fit index BIC outperforms both AIC and -2LL, making BIC the recommended method for practical Q-matrix validation. Q-matrix validation effectiveness is influenced by sample size, with increased sample sizes improving correction accuracy. Overall, this study provides important methodological support for Q-matrix validation in polytomous cognitive diagnosis applications.

Keywords: cognitive diagnosis; Q-matrix; seq-GDINA; BIC

1 Introduction

Traditional psychological and educational assessments evaluate and rank students' abilities to measure learning outcomes or inform selection decisions, but they cannot provide detailed information about the underlying psychological processes and cognitive skills. As measurement technology has evolved, there is growing demand for assessments that provide more detailed diagnostic information to enable targeted remediation and personalized instruction. Cognitive diagnosis, which combines cognitive psychology and psychometrics, can diagnose individuals' internal psychological processing and cognitive skills, thereby providing a basis for targeted remediation and adaptive teaching (Chang, 2015; Chen, 2017; Zhang & Wang, 2016). To this end, researchers have developed numerous cognitive diagnosis models (CDMs), including DINA (Haertel, 1984), NIDA (Maris, 1999), DINO (Templin & Henson, 2006), R-RUM (Hartz & Roussos, 2008), A-CDM, and G-DINA (de la Torre, 2011), all of which apply to dichotomously scored test items.

To accommodate polytomous scoring contexts, researchers have developed polytomous cognitive diagnosis models such as the polytomous GDM (von Davier, 2008), P-DINA model (Tu, Cai, Dai, & Ding, 2010), polytomous LCDM (Hansen, 2013), and seq-GDINA (Ma & de la Torre, 2016). Unlike other polytomous models, seq-GDINA can define Q-matrices at both the item level and the score category level. Ma and de la Torre (2016) defined the item-level Q-matrix as an "Unrestricted Q-matrix" and the category-level Q-matrix as a "Restricted Q-matrix." For example, solving the expression

$63/45-\sqrt{?}$ involves three steps: step 1 calculates $45/3=15$, step 2 calculates $15-6=9$, and step 3 calculates $\sqrt{9}=3$. These three steps assess three attributes: A1 (division), A2 (subtraction), and A3 (square root operation). The item-level Q-matrix defines this item' s q-vector as $[1 \quad 1 \quad 1]$, indicating the item measures all three attributes. In contrast, the category-level Q-matrix requires Q-matrix specification for each step (or each score category), defining the item' s category-level Q-matrix as $\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}$, where step 1 measures A1, step 2 measures A2, and step 3 measures A3. Category-level Q-matrices can more deeply probe students' solution processes and thereby improve classification accuracy (Ma & de la Torre, 2016). Detailed discussions of item-level and category-level Q-matrices can be found in Ma and de la Torre (2016). Overall, category-level Q-matrices more accurately reflect the attributes required at each step. However, in practice, specifying category-level Q-matrices is more complex than item-level Q-matrices, increasing the burden on content experts. Similarly, validating category-level Q-matrices is more challenging than item-level Q-matrix validation because it requires examining each category of every item, whereas item-level validation only considers which attributes the entire item measures collectively.

Numerous Q-matrix validation methods have been proposed. Some methods only apply to simplified cognitive diagnosis models (e.g., DINA and DINO), including the γ method (Tu, Cai, & Dai, 2012), Hamming distance method (Wang, Gao, Han, & Tu, 2018), ICC-IR method (Wang, Gao, Cai, & Tu, 2018), δ method (de la Torre, 2008), and RSS method (Chiu, 2013). Additionally, researchers have proposed methods applicable to saturated cognitive diagnosis models, including the GDI method (de la Torre & Chiu, 2016), likelihood-based methods (Xu & Shang, 2018), and residual-based methods (Chen, 2017). The latter three methods have broader applicability, working for both saturated and simplified models. Among these three, the GDI method is relatively complex computationally and requires setting a cutoff value (PVAF=0.95). Moreover, pilot studies show this method is heavily influenced by sample size and performs poorly with small samples (N=500). While the residual-based method can examine whether test attributes are redundant or missing at the test level, it is not sensitive to attribute redundancy at the item level (Chen, 2017). Xu and Shang (2018) employed a TLP (truncated L1 penalty function) regularization algorithm to infer item q-vectors from estimated sparse item parameter matrices, combined with information criteria (BIC) for Q-matrix estimation or validation. Their research included both theoretical proofs and Monte Carlo experiments demonstrating good performance. Furthermore, Chen, de la Torre, and Zhang (2013) applied -2LL, AIC, and BIC indices to discriminate among different Q-matrices, finding that -2LL performed well in the DINA model, but in saturated models, -2LL tended to select Q-matrices with additional attributes, while BIC demonstrated the best performance.

These methods apply to dichotomous models. For polytomous model Q-matrix

validation, Ma and de la Torre (2019) proposed a stepwise method combining GDI and Wald tests for seq-GDINA model Q-matrix validation. This method first selects the q-vector with the maximum GDI value among single-attribute q-vectors as the baseline, then uses Wald tests to determine whether to add or delete attributes, and calculates the GDI value after the Wald test to decide termination. This approach requires multiple Wald tests and standard error calculations for each category' s q-vector, making it computationally complex. Moreover, it examines attributes from the perspective of whether they are missing or redundant, without considering the overall model fit after Q-matrix validation.

Therefore, this study applies model relative fit statistics to Q-matrix validation for polytomous cognitive diagnosis models, focusing on the more diagnostically valuable category-level Q-matrix validation. Specifically, this research employs -2LL, AIC (Akaike' s Information Criterion), and BIC (Bayesian Information Criteria) indices for Q-matrix validation in polytomous cognitive diagnosis models. Our approach shares similarities with Xu and Shang (2018) in that both require model parameter estimation and use information criteria for Q-matrix validation, and both determine q-vectors item-by-item (or category-by-category) while keeping other items' Q-matrices unchanged. The difference lies in that Xu and Shang (2018) use TLP regularization to infer q-vectors from sparse parameter matrices combined with BIC, thus avoiding estimation of all possible q-vectors. In contrast, our method selects the optimal q-vector from all possible q-vectors based on fit statistics. Additionally, Xu and Shang (2018) focused on dichotomous Q-matrix estimation or validation, while this study investigates polytomous Q-matrix validation. Monte Carlo simulation studies and empirical data analysis validate our methods and compare them with Ma and de la Torre' s (2019) stepwise method, providing methodological support for practitioners in polytomous cognitive diagnosis Q-matrix validation and specification.

2 The seq-GDINA Model

As previously mentioned, among various polytomous CDMs, the seq-GDINA model can specify Q-matrices at the score category level, enabling more detailed probing of examinees' solution processes. Additionally, this model uses the G-DINA model as the link function for each category and can transform into different polytomous models (e.g., seq-DINA and seq-RRUM) under different assumptions, making it more flexible. Therefore, this study employs Ma and de la Torre' s (2016) seq-GDINA model. The model is introduced as follows:

For an examinee with attribute mastery pattern $_c$, the probability of correctly responding to category h of item j is S_{jh} . If H_j is the total number of categories for item j, then the probability of this examinee receiving a score of h on item j is:

$$P(Y_{ij} = h|\ _c) = S_{jh} \prod_{h'=0}^{h-1} (1 - S_{jh'}\ )$$

The probability of receiving a score of 0 is:

$$P(Y_{ij} = 0|\ _c) = 1 - S_{j1}$$

For each mastery pattern, the sum of probabilities across all score categories for item j equals 1.

Examinees' score probabilities for each category of item j are influenced by the attributes measured by each category of item j. In a test with K attributes, let K_j^* be the number of attributes measured by item j, and K_{jh}^* be the number of attributes measured by category h of item j. Let $\ ^*_{ljh}$ represent the l-th reduced mastery pattern among the $2^{K_{jh}^*}$ possible patterns. For mastery pattern $\ ^*_{ljh}$, the probability function for category h of item j is expressed as:

$$S_{jh\ ^*_{ljh}} = \frac{\exp(\phi_{0jh} + \sum_{k=1}^{K_{jh}^*} \phi_{jhk}\alpha_{lk} + \sum_{k'=k+1}^{K_{jh}^*} \phi_{jhkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \phi_{jh12\cdots K_{jh}^*} \prod_{k=1}^{K_{jh}^*} \alpha_{lk})}{1 + \exp(\phi_{0jh} + \sum_{k=1}^{K_{jh}^*} \phi_{jhk}\alpha_{lk} + \sum_{k'=k+1}^{K_{jh}^*} \phi_{jhkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \phi_{jh12\cdots K_{jh}^*} \prod_{k=1}^{K_{jh}^*} \alpha_{lk})}$$

where ϕ_{0jh} is the intercept parameter, ϕ_{jhk} is the main effect of attribute α_{lk}, $\phi_{jhkk'}$ is the interaction effect between attributes α_{lk} and $\alpha_{lk'}$, and $\phi_{jh12\cdots K_{jh}^*}$ is the interaction effect of all attributes. Similar to the GDINA model (de la Torre, 2011), under different constraints, the seq-GDINA model can be transformed into seq-DINA, seq-RRUM, and other models.

3 Q-Matrix Validation for Polytomous Cognitive Diagnosis Models

In cognitive diagnosis, when defining the q-vector for item j, the q-vector that yields better relative model fit among all possible q-vectors (while keeping other items'Q-matrices unchanged) should be selected as item j's q-vector. Commonly used relative fit indices in cognitive diagnosis include -2LL, AIC, and BIC. In constrained models (e.g., DINA), misspecification of item q-vectors leads to increased guessing and slip parameters, thereby reducing model likelihood. Thus, -2LL can identify appropriate q-vectors in constrained models. However, in complex models, research (Chen et al., 2013; Chen, 2017) shows that adding attributes (overspecification) increases model likelihood due to more parameters. Consequently, -2LL tends to select q-vectors of all 1s ($[1111\cdots]$) as item j' s q-vector in complex models. Therefore, penalizing model complexity is necessary when specifying q-vectors in complex models, which AIC and BIC accomplish by penalizing parameter count.

3.1 The -2LL Method

In dichotomous model Q-matrix validation, for a test with J items and K attributes, the reduced Q-matrix rQ represents the set of all possible attribute patterns, containing $2^K - 1$ patterns when attributes are independent. Using -2LL for Q-matrix specification involves treating each pattern in rQ as item j' s attribute pattern (keeping other $J-1$ items' Q-matrices unchanged), estimating parameters with all J items, and calculating -2LL. The pattern with minimum -2LL is selected as item j' s attribute pattern:

$$\hat{q}_j = \arg \min_{l=1,\ldots,2^K-1} -2LL(\mathbf{T}_{jl})$$

where $-2LL(\mathbf{T}_{jl}) = -2\sum_{i=1}^{N} \log \sum_{c \in \mathbf{s}} P(\mathbf{Y}_i|\ _c, \mathbf{T}_{jl})\pi(\ _c)$, $P(\mathbf{Y}_i|\ _c, \mathbf{T}_{jl})$ is the likelihood of item j for examinee i with mastery pattern $_c$, $\pi(\ _c)$ is the posterior probability of mastery pattern $_c$, $\mathbf{s}$ is the set of all possible mastery patterns, and $\mathbf{T}_{jl}$ is the test Q-matrix when item j' s attribute pattern is the l-th pattern.

Unlike dichotomous models, polytomous models require sequential validation and correction of each category' s q-vector for each item. Similarly, the pattern with minimum -2LL is selected as category h of item j' s attribute pattern. Each pattern in rQ can serve as category h of item j' s attribute pattern (keeping other categories of item j and all categories of other $J-1$ items unchanged), parameters are estimated with all items, and -2LL is calculated. The pattern with minimum -2LL is selected:

$$\hat{q}_{jh} = \arg \min_{l=1,\ldots,2^K-1} -2LL(\mathbf{T}_{jhl})$$

where $-2LL(\mathbf{T}_{jhl}) = -2\sum_{i=1}^{N} \log \sum_{c \in \mathbf{s}} P(\mathbf{Y}_i|\ _c, \mathbf{T}_{jhl})\pi(\ _c)$, and $\mathbf{T}_{jhl}$ is the test Q-matrix when category h of item j' s attribute pattern is the l-th pattern q_l.

Using -2LL for Q-matrix validation requires cycling through all categories of all items sequentially, enabling Q-matrix correction.

3.2 The AIC Method

AIC (Akaike' s Information Criterion), developed by Akaike (1974), is a commonly used relative fit index in psychometrics for comparing model fit. Similar to -2LL, AIC validation for category h of item j involves sequentially using each attribute pattern in rQ as category h of item j' s attribute pattern, estimating parameters with other items, calculating AIC, and selecting the pattern with minimum AIC. The AIC formula is:

$$AIC = -2\log L + 2d$$

where L is the model' s marginal likelihood (calculated similarly to -2LL) and d is the number of parameters estimated. Compared to -2LL, AIC penalizes parameter count, with patterns having more parameters receiving greater penalty.

3.3 The BIC Method

BIC (Bayesian Information Criteria), developed by Schwarz (1978), is typically used alongside AIC for model comparison. Unlike AIC, BIC also considers sample size' s impact on model fit. The BIC formula is:

$$BIC = -2\log L + d\log(N)$$

where N is the sample size, and L and d are marginal likelihood and parameter count, respectively. BIC validation for category h of item j follows the same procedure as AIC, selecting the pattern with minimum BIC.

3.4 Exhaustive and Sequential Search Algorithms

Let $q^{(h)}$ be the h-th q-vector in rQ, $K^{(h)}$ be its attribute count, and $q^{(h')}$ be nested in $q^{(h)}$ (de la Torre, 2011). Let $\mathbf{S}_j$ denote the set of q-vectors.

The exhaustive algorithm uses each pattern in rQ as category h of item j' s attribute pattern, calculates relative fit indices, and selects the best-fitting q-vector from $\mathbf{S}_j$ as category h of item j's q-vector. This method is computationally intensive; when $K = 5$, exhaustive search requires estimating $2^5-1 = 31$ models per category.

Sequential algorithms include: (1) Forward search algorithm starts by selecting the best-fitting single-attribute q-vector from $\mathbf{S}_j$, denoted $q^{(y)}$, then compares fit between q-vectors in $\{q^{(h)} \in \mathbf{S}_j | q^{(y)} \subset q^{(h)}\}$ and $q^{(y)}$. If the best-fitting q-vector in this set outperforms $q^{(y)}$, it replaces $q^{(y)}$. This repeats until no q-vector in $\mathbf{S}_j$ outperforms $q^{(y)}$ or $q^{(y)}$ contains all attributes (i.e., $q^{(y)} = [111\cdots]$). (2) Backward search algorithm starts from the all-1 q-vector $q^{(y)} = [111\cdots]$, compares fit between q-vectors in $\{q^{(h)} \in \mathbf{S}_j | q^{(h)} \subset q^{(y)}\}$ and $q^{(y)}$, and replaces $q^{(y)}$ if a better-fitting q-vector is found. This repeats until no q-vector outperforms $q^{(y)}$ or $q^{(y)}$ measures only one attribute (i.e., $K^{(y)} = 1$). (3) Forward-then-backward search algorithm uses the expert-specified q-vector from the original Q-matrix as $q^{(y)}$, performs forward search, then backward search. (4) Backward-then-forward search algorithm also uses the expert-specified q-vector as $q^{(y)}$, but performs backward search first, then forward search.

The forward and backward algorithms do not utilize expert-specified Q-matrix information, while the latter two algorithms search based on expert-specified q-vectors. The latter two sequential algorithms' required estimations per category vary with the degree of q-vector misspecification, but they substantially reduce computations compared to exhaustive search.

3.5 Q-Matrix Validation Procedure

Let the Q-matrix requiring validation be $\mathbf{Q}^{(0)}$, typically specified by experts. For a test with J items, where each item j has H_j categories (which may vary across items), the test has $\sum_{j=1}^{J} H_j$ total categories. Let $\mathbf{S}^{(0)}$ be the set of all categories: $\{1,1;\ldots;1,H_1;\ldots;J,1;\ldots;J,H_J\}$. The procedure is:

Step 1: Select the first item j from test $\mathbf{S}^{(0)}$ and validate the q-vector for its first category, keeping Q-matrices for other $J-1$ items and other categories of item j unchanged.

Step 2: Use the sequential algorithm to select the optimal q-vector for category 1 of item j based on -2LL (or AIC, BIC if using those methods).

Step 3: Repeat Steps 1-2 to determine optimal q-vectors for other categories of item j using the same method.

Step 4: After completing Steps 1-3 for all categories of item j, repeat Steps 1-3 for remaining items until optimal q-vectors are determined for all categories.

Step 5: Among all categories, select the category whose modification yields optimal relative fit (-2LL, AIC, or BIC) and modify it, then remove this category from $\mathbf{S}^{(0)}$. The reduced set is $\mathbf{S}^{(1)}$, and the modified Q-matrix is $\mathbf{Q}^{(1)}$.

Step 6: Verify if $\mathbf{Q}^{(1)}$ matches $\mathbf{Q}^{(0)}$. If not, replace $\mathbf{Q}^{(0)}$ with $\mathbf{Q}^{(1)}$ and $\mathbf{S}^{(0)}$ with $\mathbf{S}^{(1)}$, then repeat Steps 1-5. The algorithm stops when $\mathbf{S}^{(0)} = \emptyset$.

4 Simulation Study Design

To examine different methods' performance in polytomous cognitive diagnosis Q-matrix validation, simulation studies investigated method effectiveness across different sample sizes, Q-matrix error types, and polytomous CDMs, comparing results with Ma and de la Torre' s (2019) stepwise method. Specifically, Study 1 examined methods in simplified polytomous CDMs (seq-DINA and seq-RRUM), while Study 2 examined saturated polytomous CDMs (seq-GDINA).

4.1 Study 1: Method Comparison in Simplified Models

4.1.1 Q-Matrix This study used the Q-matrix from Ma and de la Torre (2016), containing 21 items and 5 attributes, shown in Table 1 .

4.1.2 Cognitive Diagnosis Models, Examinee Parameters, and Item Parameters Study 1 used seq-DINA and seq-RRUM models. Examinee mastery patterns were generated from a multidimensional normal distribution ($\mathbf{0}$,) with attribute correlations set to 0.5, following previous research (Chen, 2017; Liu, Xin, Andersson, & Tian, 2019). Sample sizes were 500, 1000, and 2000, representing small, medium, and large samples. Item parameters were simulated such that for examinees mastering all attributes required for category h of item j, the probability of receiving score h was randomly drawn from

[0.75, 1], i.e., $S_{jh} \sim U(0.75, 1)$. For examinees mastering no attributes required for category h of item j, the probability was randomly drawn from [0, 0.25], i.e., $S_{jh} \sim U(0, 0.25)$. For seq-RRUM, probabilities for other mastery patterns were randomly drawn from $[S_{jh}^{\text{no mastery}}, S_{jh}^{\text{full mastery}}]$ following monotonicity constraints, where examinees mastering more attributes have higher probabilities of receiving score h on item j than those mastering fewer attributes: if $_{l} \geq _{l'}$, then $S_{jh\,l} \geq S_{jh\,l'}$.

4.1.3 Q-Matrix Error Simulation Following previous research (Chen et al., 2013; Liu, Tian, & Xin, 2016; Chen, 2017; Liu et al., 2019), we examined attribute redundancy, attribute omission, and combined omission-redundancy, creating six Q-matrix error types:

- **Q1:** Randomly selected 5 categories measuring one attribute and changed one "0" to "1" in each category.
- **Q2:** Randomly selected 5 categories measuring two or more attributes and changed one "1" to "0" in each category.
- **Q3:** Randomly selected 5 categories measuring two or more attributes, changed one "0" to "1" and one "1" to "0" in each category.
- **Q4:** Combined all errors from Q1, Q2, and Q3.
- **Q5 and Q6:** Simulated 10% and 20% random errors, respectively, while ensuring each category measured between 1 and 3 attributes.

Error types are summarized in Table 2 .

4.1.4 Response Simulation Based on simulated examinee and item parameters, we calculated response probabilities P_{ijh} for examinee i on each category h of item j, then generated observed responses from a categorical distribution with probabilities P_{ijh}.

4.1.5 Evaluation Metrics We calculated pattern match rate (PMR) as the proportion of correctly specified attribute patterns across all categories between corrected and true Q-matrices, and attribute match rate (AMR) as the proportion of correctly specified individual attributes. False Positive Rate (FPR) and True Positive Rate (TPR) represented proportions of incorrectly specified attributes left uncorrected and correctly specified attributes left unchanged, respectively. All experiments were replicated 200 times, and average PMR, AMR, FPR, and TPR were computed.

To compare Q-matrix quality before and after correction, we calculated absolute fit index RMSEA (Liu et al., 2016) before and after correction, averaging across 200 replications.

In complex models (seq-RRUM and seq-GDINA), BIC consistently achieved the highest Q-matrix recovery rates across all conditions, with -2LL and AIC producing worse RMSEA values than BIC. In simplified models (seq-DINA), AIC, BIC, and -2LL were equivalent. Among the four sequential algorithms,

forward-then-backward and backward-then-forward algorithms performed similarly, while forward and backward algorithms were slightly inferior in some conditions. Compared to exhaustive search, forward-then-backward algorithm did not reduce correction accuracy, with attribute match rate differences within 1%. For brevity, we report only forward-then-backward algorithm results for BIC and stepwise methods, including corrected Q-matrix RMSEA values.

4.2 Study 1 Results

Tables 3 and 4 present BIC and stepwise method results for seq-DINA and seq-RRUM models. Table 3 shows that BIC performed well for Q-matrix validation in seq-DINA, with average PMR and AMR of 83.0% and 96.9% across all conditions, compared to 78.1% and 95.7% for stepwise. Overall, BIC' s PMR and AMR were slightly higher, with most attribute match rate differences within 1%.

Across Q-matrix error types, BIC showed similar recovery rates for Q1-Q5, with AMR between 95%-98%, while Q6 results were slightly lower at 93%-95%. Stepwise showed comparable performance across error types, with AMR above 92%. Thus, error type had minimal impact on overall correction effectiveness for either method in seq-DINA.

For FPR and TPR, BIC's TPR reached approximately 95% across all conditions, indicating it rarely changed correctly specified attributes. BIC' s FPR was lower for Q2, possibly due to DINA' s characteristic that all measured attributes must be mastered for correct responses, making BIC 倾向于将缺失的属性修改过来. Stepwise' s TPR was similar to BIC, but its FPR was higher for Q2-Q4, suggesting stepwise is less sensitive to attribute omission.

Sample size affected correction effectiveness: larger samples yielded higher recovery rates. At N=500, BIC and stepwise AMR were 95.6% and 94.6%; at N=2000, they were 97.9% and 96.7%. Thus, increasing sample size improves Q-matrix validation.

Regarding fit, corrected Q-matrices from both methods had lower RMSEA values than original Q-matrices, indicating better data fit. Across all conditions, original Q-matrix average RMSEA was 0.048, while BIC and stepwise corrected Q-matrices averaged 0.007 and 0.017, respectively. BIC-corrected Q-matrices fit better than stepwise, with an average difference of 0.01. Larger sample sizes produced smaller RMSEA for BIC-corrected Q-matrices (e.g., 0.003-0.004 for Q1-Q5 at N=2000).

Table 4 shows that in seq-RRUM, BIC generally outperformed stepwise, with average PMR of 87.5% vs. 78.1% and AMR of 98% vs. 96% across all conditions. Sample size effects were similar: at N=500, AMR was 97.4% (BIC) vs. 94.8% (stepwise); at N=2000, AMR was 98.6% vs. 96.9%.

Error type had minimal impact on overall recovery rates. For Q1-Q5, both methods' AMR hovered around 96% (stepwise) and 98% (BIC). For Q6, AMR

decreased slightly but not substantially.

For FPR and TPR, both methods' TPR exceeded 95% across error types, while FPR was slightly higher for Q2-Q6 than Q1, indicating both methods are more sensitive to attribute redundancy. This suggests attribute omission in Q-matrices has greater impact.

Regarding absolute fit, Q1-corrected Q-matrices showed RMSEA nearly identical to original Q-matrices, as Q1 involves attribute redundancy, which doesn' t worsen fit in complex models. For Q2-Q6, original Q-matrix average RMSEA was 0.037, while stepwise and BIC corrected Q-matrices averaged 0.007 and 0.005, respectively, showing better fit. BIC-corrected Q-matrices fit better than stepwise, and larger samples improved corrected Q-matrix fit (RMSEA of 0.006 and 0.003 for stepwise and BIC at N=2000).

5 Study 2: Method Comparison in the seq-GDINA Model

Study 1 models have constrained link functions at each category that can be derived from seq-GDINA. As a saturated model, seq-GDINA has broader applicability. Study 2 validates and compares methods in seq-GDINA.

5.1 Study 2 Design

Study 2 design mirrored Study 1, except using seq-GDINA model. Other conditions are described in Study 1.

5.2 Study 2 Results

Table 5 presents BIC and stepwise results in seq-GDINA. Similar to Study 1, BIC generally outperformed stepwise, with average PMR of 90.5% vs. 84.5% and AMR of 98.6% vs. 97.1%. Recovery rates increased with sample size: at N=500, BIC PMR and AMR were 86% and 97.9%, stepwise were 78% and 95.9%; at N=2000, BIC were 94.8% and 99.3%, stepwise were 90.8% and 98.5%. Error type had minimal impact. Regarding absolute fit, Q1-corrected Q-matrices showed RMSEA nearly identical to original, again due to attribute redundancy. For Q2-Q6, original Q-matrix average RMSEA was 0.036, while stepwise and BIC corrected Q-matrices averaged 0.007 and 0.006, indicating better fit. Larger samples improved corrected Q-matrix fit.

6 Empirical Data Analysis

This study analyzed two TIMSS (Trends in International Mathematics and Science Study) datasets: 2011 8th-grade and 2007 4th-grade mathematics data. The TIMSS 2011 data, with Q-matrix specified by Park, Lee, and Johnson (2017) and used by Ma and de la Torre (2019) for polytomous Q-matrix validation, included 23 items, 7 attributes, and 748 students. Item 11 was polytomous; others were dichotomous. The Q-matrix is shown in Table 6 .

Before analysis, model fit indices (deviance, AIC, BIC) were compared across models, with different indices favoring different models. To avoid model selection errors, seq-GDINA was used, which reduces to GDINA for dichotomous items. Correction results show BIC adjusted 12 items with 14 attributes, while stepwise adjusted 6 items with 6 attributes, all of which were included in BIC' s adjustments. Neither method adjusted item 11' s two category q-vectors.

Table 7 shows agreement rates among Q-matrices. BIC-corrected and original Q-matrices had 0.92 agreement, stepwise-corrected and original had 0.96, and BIC- and stepwise-corrected had 0.95, indicating high agreement.

Table 8 compares relative fit indices (-2LL, AIC, BIC) and absolute fit indices (M2 test, RMSEA, SRMSR). Both corrected Q-matrices outperformed the original on relative fit. On absolute fit, original Q-matrix M2 test was $p < 0.01$, while corrected Q-matrices were $p = 0.2$ and 0.3, indicating better fit. RMSEA and SRMSR also favored corrected Q-matrices. The two corrected Q-matrices had similar fit, with stepwise better on M2 and RMSEA, BIC better on relative fit and SRMSR.

TIMSS 2007 data, with Q-matrix specified by Lee, Park, and Taylan (2011) and used by Ma and de la Torre (2016), included 11 items, 8 attributes, and 823 students. Items 3, 7, and 9 were polytomous; others were dichotomous. The original Q-matrix is shown in Table 9 .

Stepwise adjusted attribute 5 (A5) to be unmeasured by any item, so detailed results are not shown. BIC-adjusted Q-matrix is shown in Table 9. Absolute and relative fit indices were calculated. Due to low degrees of freedom, M2 test could not be performed. Original and BIC-corrected SRMSR were 0.0312 and 0.0246. On relative fit, BIC-corrected Q-matrix (AIC=11222.25; BIC=12677.42) fit better than original (AIC=11513.79; BIC=13195.01). Both empirical analyses show BIC-corrected Q-matrices fit better.

7 Discussion and Conclusions

7.1 Conclusions

This study investigated category-level Q-matrix validation for polytomous cognitive diagnosis assessment, validating and comparing stepwise and relative fit statistics methods through Monte Carlo simulation and empirical studies, providing methodological support for practical polytomous test Q-matrix validation. Findings indicate: (1) BIC method achieves high pattern and attribute match rates for polytomous CDM Q-matrix validation, with higher recovery rates than stepwise, and BIC-corrected Q-matrices fit data better. (2) In complex models, BIC outperforms AIC and -2LL, making BIC the recommended method. (3) Q-matrix validation effectiveness is affected by sample size, with larger samples improving correction accuracy.

7.2 Discussion

(1) Polytomous Cognitive Diagnosis Q-Matrix Validation Methods

Polytomous items are common in practice and provide more information than dichotomous items, making polytomous CDM development important. This study applied relative fit statistics to polytomous CDM Q-matrix validation, improving validation algorithms. BIC showed good performance, was less affected by sample size, and worked well across error types. To improve efficiency, we used sequential algorithms. Simulation showed our sequential algorithm' s attribute match rate differed from exhaustive search by less than 1%. Our forward-then-backward algorithm performed similarly to backward-then-forward, while forward and backward algorithms performed slightly worse, likely because they didn' t utilize expert-specified q-vector information. When the true model is uncertain, saturated models (seq-GDINA) can be used for Q-matrix validation without reducing accuracy.

(2) Category-Level vs. Item-Level Q-Matrix Validation in Polytomous CDMs

Category-level Q-matrices require separate Q-matrix specification for each category, enabling more accurate probing of solution processes and higher classification accuracy. However, specifying Q-matrices for each category is difficult and increases workload. Item-level Q-matrix specification is simpler but ignores step-level information, reducing classification accuracy (Ma & de la Torre, 2016). Category-level Q-matrix validation is also more challenging. This study investigated category-level Q-matrix validation in polytomous models, finding BIC effective and providing methodological support for category-level Q-matrix specification and validation.

(3) Integrating Q-Matrix Validation Results with Expert Opinion

Data-driven Q-matrix validation methods can avoid subjectivity in expert specification and reduce expert burden. However, objectively validated Q-matrices should not be adopted directly; they should be integrated with expert opinion. Data-driven validation can inform and support expert specification but cannot replace experts' crucial role in test design and Q-matrix development. While BIC-corrected Q-matrices fit data better, modifications may not always be appropriate and require expert review. When objective methods conflict with expert opinion, multiple experts should discuss or disputed items should be removed.

(4) Future Directions

This study proposed polytomous Q-matrix validation methods and found BIC applicable. However, further research is needed on: effects of different item parameter quality, performance of different methods at item-level Q-matrix validation, validation effectiveness with hierarchical attribute relationships, and more. Additional issues requiring investigation include theoretical proofs of

Q-matrix completeness and identifiability, automatic identification of incorrect attribute counts, and more real-data studies. Further research on polytomous Q-matrix validation methods is needed.

References

Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transactions on Automated Control, 19*, 716–723.

Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika, 80*, 1–20.

Chang, H.-H., & Wang, W. Y. (2016). "Internet plus" measurement and evaluation: a new way for adaptive learning. *Journal of Jiangxi Normal University (Natural Science), 40*(5), 441–455. [张华华, 汪文义. (2016). "互联网 +" 测评: 自适应学习之路. 江西师范大学学报 (自然科学版), 40(5), 441–455.]

Chen, J. S. (2017). A residual-based approach to validate Q-matrix specifications. *Applied Psychological Measurement, 41*, 277–293.

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.

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

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

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

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*, 253–273.

Haertel, E. H. (1984). An application of latent class models to assessment data. *Applied Psychological Measurement, 8*, 333–346.

Hansen, M. (2013). Hierarchical item response models for cognitive diagnosis. Unpublished doctoral dissertation, University of California at Los Angeles.

Hartz, S. M., & Roussos, L. A. (2008). The fusion model for skills diagnosis: Blending theory with practice. *Educational Testing Service, Research Report, RR-08-71.* Princeton, NJ: Educational Testing Service.

Lee, Y.-S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing, 11*, 144–177.

Liu, Y., Tian, W., & Xin, T. (2016). An application of M2 statistic to evaluate the fit of cognitive diagnostic models. *Journal of Educational and Behavioral Statistics, 41*(1), 3-26.

Liu, Y., Xin, T., Andersson, B., & Tian, W. (2019). Information matrix estimation procedures for cognitive diagnostic models. *British Journal of Mathematical and Statistical Psychology, 72*, 18-37.

Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology, 69*, 253-275.

Ma, W., & de la Torre, J. (2019). An empirical Q-matrix validation method for the sequential generalized DINA model. *British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1111/bmsp.12156.

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

Park, J. Y., Lee, Y.-S., & Johnson, M. S. (2017). An efficient standard error estimator of the DINA model parameters when analysing clustered data. *International Journal of Quantitative Research in Education, 4*, 159-190.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

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

Tu, D. B., Cai, Y., & Dai, H. Q. (2012). A new method of Q-matrix validation based on DINA model. *Acta Psychologica Sinica, 44*(4), 558-568. [涂冬波, 蔡艳, 戴海崎. (2012). 基于 DINA 模型的 Q 矩阵修正方法. 心理学报, 44(4), 558-568.]

Tu, D. B., Cai, Y., Dai, H. Q., & Ding, S. L. (2010). A polytomous cognitive diagnosis model: P-DINA model. *Acta Psychologica Sinica, 42*(10), 1011-1020. [涂冬波, 蔡艳, 戴海琦, 丁树良. (2010). 一种多级评分的认知诊断模型: P-DINA 模型的开发. 心理学报, 42(10), 1011-1020.]

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

Wang, D. X., Gao, X. L., Cai, Y., & Tu, D. B. (2018). A new Q-matrix estimation method: ICC based on ideal response. *Journal of Psychological Science, 41*(2), 466-474. [汪大勋, 高旭亮, 蔡艳, 涂冬波. (2018). 一种非参数化的 Q 矩阵估计方法: ICC-IR 方法开发. 心理科学, 41(2), 466-474.]

Wang, D. X., Gao, X. L., Han, Y. T., & Tu, D. B. (2018). A simple and effective Q-matrix estimation method: From non-parametric perspective. *Journal of Psychological Science, 41*(1), 180-188. [汪大勋, 高旭亮, 韩雨婷, 涂冬波. (2018). 一种简单有效的 Q 矩阵估计方法开发: 基于非参数化方法视角. 心理科学, 41(1), 180-188.]

Xu, G., & Shang, Z. (2018). Identifying latent structures in restricted latent class models. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2017.1340889.

Note: Figure translations are in progress. See original paper for figures.