Joint Validation of Attribute Hierarchies and Q-Matrix in Cognitive Diagnostic Models: A Practical Perspective

Authors: Wang Lingling, Wang Lingling

Date: 2024-08-06T00:00:00+00:00

Abstract

In cognitive diagnostic assessment practices, the correctness of Q-matrix construction and attribute hierarchy relationships influences the accuracy of parameter estimation in cognitive diagnostic models and the classification accuracy of examinees. Typically, both attribute hierarchy relationships and Q-matrices are primarily established through domain expert judgment. Existing research has separately examined and revised either the Q-matrix or attribute hierarchy relationships. This study proposes a method based on Bayesian network conditional independence testing for the joint validation of Q-matrices and attribute hierarchy relationships. Two simulation studies were conducted to investigate the joint revision accuracy of the proposed method and the specific factors influencing this accuracy, thereby providing guidance for practical applications. Results indicate that when Q-matrix error rates are moderate or lower, the method can effectively revise both Q-matrices and attribute hierarchy relationships, particularly when item quality is high, sample size is adequate, and test length is longer, resulting in enhanced joint revision performance. Finally, the algorithm was applied in a concrete cognitive diagnostic assessment practice to jointly examine and revise, in a data-driven manner, the attribute hierarchy relationships and Q-matrix defined by experts. Results demonstrated improved model-data fit following revision.

Full Text

Preamble

Self-Check Report for Submissions to *Acta Psychologica Sinica*

Please complete the following items and paste them on the first page of your manuscript.

1. Does this study represent a significant contribution? *Acta Psychologica Sinica* aims to publish cutting-edge psychological research that is "both scientifically excellent and of particularly broad interest and significance." Studies with only minor incremental contributions, those that do not attempt to open new areas of inquiry or propose unique and innovative perspectives, or work that purely investigates algorithms or techniques without addressing clear psychological questions, have low acceptance probability and are recommended for submission elsewhere.

Response: In cognitive diagnostic assessment practice, the accuracy of both Q-matrix specification and attribute hierarchy construction affects model parameter estimation and examinee classification accuracy. While existing research has examined Q-matrix or attribute hierarchy validation separately, these two components are interdependent. Joint validation is urgently needed in practice. This study proposes a Bayesian network-based method for jointly validating attribute hierarchies and Q-matrices. Through simulation and empirical studies, we examine the joint correction accuracy of this method and its influencing factors to provide guidance for practical applications.

2. Have you published or submitted other articles using the same data as this study? If yes, please attach them for review. (We discourage publishing multiple articles with the same variables from one dataset or splitting a series of related studies into multiple publications.)

Response: Not applicable.

3. For non-experimental, non-intervention studies in management, clinical, personality, and social psychology that rely solely on self-report (questionnaire) methods, you must check for common method bias. What methods did you use to control or demonstrate that such bias does not affect the validity of your conclusions? (See http://journal.psych.ac.cn/xlkxjz/CN/abstract/abstract894.shtml for relevant literature). Studies based on cross-sectional data with only self-reports from convenience samples are easy to conduct but typically lack innovative value and have low acceptance probability.

Response: Not applicable.

4. Did you report and analyze effect sizes (e.g., Cohen's d for t-tests, ^{2} or ^{2}p for ANOVA, standardized regression coefficients)? (Many studies mechanically report effect sizes without necessary analysis or explanation, such as whether the effect size is small, medium, or large, or its theoretical/applied significance. Convenient calculators are available through Google. For explanations, see Chinese: http://journal.psych.ac.cn/xlkxjz/CN/abstract/abstract1150.shtml; English: http://www.uccs.edu/lbecker/effect-size.html). Did you report 95% CIs for statistical analyses (e.g., 95% CI for differences, correlations/regression coefficients)? For CI calculation and graphing, see https://thenewstatistics.com/itns/esci/.

Response: Not applicable.

5. Please state your planned and actual sample sizes. If they differ, explain why. Previous psychological research has suffered from low statistical power due to insufficient sample sizes. We recommend explaining your sample size determination in the Methods section, using justified effect sizes and desired power, and reporting the software/program used. See https://osf.io/5awp4/ for guidance.

Response: This is a psychometric theory study; not applicable.

6. [Question missing in original]

Response: Not applicable.

7. For data reporting completeness, if you excluded data in statistical analyses, did you report this in the text? Why? How would results change with inclusion? How did you handle missing data? Did you delete individual items from scales? Why? How would results change with inclusion? Are there unreported measures/variables? Why? Please indicate locations in the paper.

Response: Not applicable.

8. For experimental materials, scales, or questionnaires not peer-reviewed, are they attached for review? If not, explain why. If published, are you willing to share them with other researchers?

Response: Not applicable.

9. This journal requires authors to provide raw data. Please select one option:

a) Submit data to editorial office email after submission

b) Data available at the following link

c) Raw data and programs have been shared on the Psychological Science Data Bank (https://psych.scidb.cn/)

d) If data cannot be provided, explain why or provide justification.

Response: This is a statistical measurement theory methods study that does not involve actual collected participant data. All main simulation data are presented in the paper; some results are omitted due to space constraints but available from authors upon request. This study primarily uses newly developed statistical measurement methods to collect simulation data; empirical data comes from publicly available R packages (as stated in the paper). If reviewers or readers are interested in our new methods and developed programs, they may contact the corresponding author.

10. Is your study a clinical intervention or laboratory experiment?

Yes ( ) If yes, provide pre-registration number. If no, explain why.
Note: Pre-registration is recommended for clinical interventions or lab experiments and encouraged for other experimental studies. Pre-registration requires stating all hypotheses with justification and detailed procedures. Our pre-registration site: https://os.psych.ac.cn/preregister (manual on journal website) or https://osf.io/ or https://aspredicted.org/. Pre-registration significantly increases acceptance probability. See https://osf.io/5awp4/ for importance.

Response: No (√)

11. If your study involved human or animal subjects, was it approved by your institution's ethics committee? If yes, send scanned copy to editorial office. If no, explain why.

Response: Not applicable.

12. Did you write a 400-500 word extended English abstract following the "English Abstract Writing Guidelines" on the editorial website? Has the English title and abstract been reviewed by a native English speaker or professional SCI/SSCI editing service?

Response: Yes, revised and polished.

13. If the first author is a student, the advisor must email the editorial office (xuebao@psych.ac.cn) separately confirming they have read and reviewed the paper. Has the advisor been reminded to email? (Editorial processing begins only after receiving advisor's email.)

Response: First author is not a student.

14. Please download the "Manuscript Non-Confidentiality Certificate" from the "Download Center" on the editorial website, stamp it with your institution's confidentiality office seal, and email scanned copy to xuebao@psych.ac.cn. If no confidentiality office seal is available, use the institution's official seal. Has this been emailed?

A Joint Validation Method for Attribute Hierarchies and Q-Matrices in Cognitive Diagnosis Models: A Practice-Oriented Perspective

In cognitive diagnostic assessment practice, the correct specification of both Q-matrices and attribute hierarchy relationships significantly impacts the accuracy of model parameter estimation and examinee classification. Typically, both attribute hierarchies and Q-matrices rely heavily on expert judgment. While several studies have developed methods to validate or correct Q-matrices or attribute hierarchies separately, these two components are interdependent and influence each other. Therefore, joint validation is urgently needed in practice. This study proposes a Bayesian network-based method for jointly validating attribute hierarchies and Q-matrices. Through two simulation studies, we examine the joint correction accuracy of this method and its specific influencing factors to provide guidance for practical applications. Results show that when Q-matrix error rates are medium or lower, the method can effectively correct both Q-matrices and attribute hierarchies, particularly when item quality is high, sample size is adequate, and test length is longer. Finally, we apply the algorithm to real cognitive diagnostic assessment data, jointly validating and correcting the expert-defined attribute hierarchy and Q-matrix in a data-driven manner, demonstrating improved model-data fit after correction.

Keywords: cognitive diagnosis; attribute hierarchy relationships; Q-matrix; Bayesian network

Cognitive diagnostic theory (CDT; Leighton & Gierl, 2007) represents the core of next-generation measurement theories and provides an important pathway for formative assessment and personalized learning. In cognitive diagnostic practice, fine-grained multidimensional cognitive attributes are extracted from individuals' cognitive processes, processing skills, or knowledge structures through cognitive-psychological analysis. These attributes are the most important features characterizing test items, and various psychometric models analyze examinees' response patterns to diagnose their attribute mastery profiles. The mapping relationship between test items and cognitive attributes is represented by the Q-matrix (Tatsuoka, 1983, 1985). Meanwhile, cognitive attributes in a diagnostic test typically exhibit certain psychological, logical, or hierarchical relationships, making attribute hierarchy relationships central to cognitive model construction (Ding et al., 2012). The cognitive model is precisely the core question that cognitive diagnostic tests aim to answer—what are the cogni-

tive processing procedures of examinees in the test domain (Jiang, 2020)? The correct specification of both Q-matrices and attribute hierarchy relationships affects the accuracy of model parameter estimation and examinee classification (Chiu, 2013; de la Torre, 2009; Rupp & Templin, 2008; Liu, Huggins-Manley, & Bradshaw, 2017). Typically, both attribute hierarchies and Q-matrices depend primarily on expert judgment, and different experts may produce different specifications that may also differ from examinees' actual cognitive processes. Thus, initial Q-matrices and attribute hierarchies likely contain specification errors. Given their importance for model parameter estimation, validating and correcting these expert judgments is necessary.

In attribute hierarchy validation, early researchers proposed the Hierarchy Consistency Index (HCI) within the Rule Space Model (RSM) and Attribute Hierarchy Model (AHM) frameworks (Cui & Leighton, 2009; Cui, 2007; Wang & Gierl, 2011). Additionally, Yu et al. (2011) proposed using Bayesian network structure learning to mine attribute hierarchy relationships from attribute mastery patterns obtained from examinee response data. However, this method directly analyzes attribute mastery patterns, which must be estimated from response data and Q-matrices. Without a prespecified attribute hierarchy, estimating attribute mastery patterns may not be accurate. Other researchers developed the Hierarchical Diagnostic Classification Model (HDCM) based on the parametric Log-linear Cognitive Diagnosis Model (LCDM; Henson, Templin & Willse, 2009), creating a psychometric framework that parameterizes attribute hierarchies to confirm or refute specific structures (Templin & Bradshaw, 2014b). Templin (2014b) proposed using likelihood ratio tests (LR) to validate attribute hierarchies, but this approach tests entire prespecified hierarchies as a whole (accepting or rejecting the structure) without providing specific correction information.

Numerous studies have developed methods for validating and correcting expert-specified Q-matrices. Nonparametric methods include Euclidean distance (Chiu, 2013) and Hamming distance (Wang, Gao, Han, et al., 2018), which require small sample sizes and are easy to implement but have strict assumptions limiting their extensibility and practicality (Liu & Wu, 2023). Parametric correction methods include the d method (de la Torre, 2008), g method (Tu et al., 2012), S statistic (Liu et al., 2012), iterative modified sequential search (Terzi & de la Torre, 2018), RMSEA statistic (Kang et al., 2019), weighted residual R method (Yu & Cheng, 2020), and optimal response distribution purity method (Li et al., 2022). These methods were developed under specific CDM frameworks (e.g., DINA, DINO, R-RUM). Under saturated CDM frameworks like G-DINA (generalized deterministic input noisy output "and" gate; de la Torre, 2011), parametric Q-matrix correction methods include GDI (G-DINA discrimination index; de la Torre & Chiu, 2016), residual methods (Chen, 2017), TLP (truncated L1 penalty function; Xu & Shang, 2018), and relative fit statistics (Wang et al., 2020). Ma and de la Torre (2020) proposed a stepwise method combining GDI and Wald tests for Q-matrix correction in polytomous models, also applicable to 0-1 scored G-DINA models.

The Q-matrix validation methods described above typically ignore whether attribute hierarchies are correctly specified, while existing attribute hierarchy validation methods do not consider Q-matrix accuracy. In practice, these two components are inseparable and mutually influential, and both inevitably contain specification errors due to expert subjectivity. Wang and Lu (2021) proposed two exploratory methods to learn attribute hierarchies directly from data without Q-matrix information. Ma et al. (2022) proposed a penalized likelihood method to first determine the number of attributes, then jointly estimate attribute hierarchies and Q-matrices. While statistically sophisticated, these methods are computationally complex, using only 3-4 attributes in simulations. When attribute numbers increase, the computational complexity of jointly estimating Q-matrices and attribute hierarchies from noisy data becomes prohibitive. In practice, domain experts typically provide substantial prior knowledge about attribute hierarchies and Q-matrices, making a method that can simultaneously validate and correct existing specifications more urgently needed. Therefore, this study proposes a Bayesian network-based joint validation and correction method for attribute hierarchies and Q-matrices. The following sections introduce the principles and algorithmic implementation of this Bayesian network approach, evaluate its validation and correction efficacy through simulation and empirical studies, and conclude with a discussion.

2 Principles of the Joint Validation Method Based on Bayesian Networks

Bayesian Networks (BN) are probabilistic models combining probability theory and graph theory (Zhang & Guo, 2006). A BN consists of a structural model and conditional probability parameters. The structural model is a directed acyclic graph where nodes represent random variables and directed edges represent dependency or causal relationships. For any connected pair, the originating node is the parent and the target node is the child. These dependencies are quantified by conditional probabilities of each node given its parents. Attribute hierarchy relationships and Q-matrix information in cognitive diagnosis can be represented as BN structures. For example, Figure 1 [Figure 1: see original paper] shows a cognitive diagnostic test with 3 attributes and 5 items, where the BN structure comprehensively represents both attribute hierarchy and Q-matrix information: ellipses represent cognitive attributes, rectangles represent test items, the network shows connections between items and measured attributes (Q-matrix information), and presents hierarchical relationships among attributes (e.g., Attribute A1 is a prerequisite for A2, and both are prerequisites for A3). In BN, unidirectional arrows (edges) between attributes and items represent Q-matrix elements of "1" (attributes measured by items), while edges between attributes represent correlations. BN enables significance testing of each edge to determine its validity. Testing every edge in the network structure is equivalent to testing whether each item-attribute specification in the Q-matrix is correct and simultaneously testing the significance of attribute-attribute edges (i.e., validating attribute hierarchy relationships).

Figure 1. Bayesian network structure comprehensively representing attribute hierarchy relationships and Q-matrix

In BN, the validity of edges connecting two nodes can be tested by examining conditional independence between the node variables, implemented through conditional chi-square tests (Pearson Chi-square) or conditional log-likelihood ratio tests (Log likelihood Ratio; Xue & Chen, 2012). The null hypothesis of conditional independence testing is that two nodes are conditionally independent (no edge). If the resulting significance level is very low, we reject the null hypothesis and conclude that an edge exists, using the p-value as an indicator of edge strength. If this p-value is large, the risk of error in rejecting the null hypothesis increases, suggesting the edge lacks data support—indicating either the item does not measure the attribute or no hierarchical relationship exists between the attributes. Thus, BN can simultaneously validate Q-matrix and attribute hierarchy correctness, and correct misspecified item attributes or attribute relationships through node independence tests. While conditional independence tests evaluate individual edges, network scores (AIC, Akaike, 1974; BIC, Schwaz, 1978; log-likelihood criterion, Pinheiro & Bates, 1995) assess overall model-data fit. Therefore, BN structures representing attribute hierarchies and Q-matrices can be validated through conditional independence tests or network scores. The specific implementation steps are:

Step 1: Estimate examinees' attribute mastery patterns based on the initial Q-matrix and attribute hierarchy.

Step 2: Initial Q-matrices may contain redundant or missing attributes, and initial attribute hierarchies may contain redundant or missing edges. Testing a BN structure built only from initial specifications can detect redundant edges but not missing ones. Therefore, we initially construct a saturated BN model, assuming each item measures all attributes and all attribute pairs are connected, enabling testing of every possible edge.

Step 3: Input response data and estimated attribute mastery patterns into this saturated BN model and test the strength of each edge's existence (significance level of conditional independence between attribute-item or attribute-attribute nodes). Retain significant edges to obtain a corrected Q-matrix and attribute hierarchy. Based on pilot studies, this research uses a significance level of 0.001. Notably, BN tests all edges simultaneously rather than sequentially, ensuring computational efficiency.

Step 4: Since initial attribute mastery pattern estimates contain errors, the BN structure corrected from this noisy data is not completely accurate. We further refine it using model-data fit indices. For each corrected element in the Q-matrix and attribute hierarchy, we calculate the network score (BIC in this study) and compare it with the pre-correction score. If correction reduces BIC, we retain the correction; otherwise, we keep the original specification. This optimizes the BN structure based on overall model-data fit.

Step 5: Re-estimate examinees' attribute mastery patterns using the corrected

Q-matrix and attribute hierarchy.

Step 6: Input the re-estimated attribute mastery patterns and response data into the BN model and repeat Steps 3-5. In the second iteration, more precise attribute mastery pattern estimation reduces input data noise, yielding more accurate BN structure corrections. Pilot studies show that stable BN structures (and thus stable Q-matrices and attribute hierarchies) are typically achieved after 3 iterations, with further iterations providing negligible improvement. The algorithm maintains high computational efficiency even with 3 iterations. In simulation studies, we compare the final Q-matrix with the true matrix to calculate correction accuracy rates, and similarly compare the final attribute hierarchy with the true hierarchy. Edge strength testing and network score calculation are implemented using the R package *bnlearn* (Scutari et al., 2021).

3 Analysis of Factors Influencing the Joint Validation Method

We conducted two simulation studies to examine the accuracy and influencing factors of the BN method in validating and correcting Q-matrices and attribute hierarchies. Simulation studies control extraneous variables to evaluate method performance under abstract test conditions, providing basic numerical stability and precision. Many existing Q-matrix correction methods ignore attribute hierarchy effects, while attribute hierarchy validation studies have not rigorously examined correction accuracy or considered Q-matrix specification accuracy. The two existing studies on joint Q-matrix and attribute hierarchy estimation (Wang & Lu, 2021; Ma et al., 2022) focus on estimation rather than validation. Therefore, as an initial exploration, this study uses simulation to investigate correction accuracy and influencing factors, providing valuable reference information for cognitive diagnostic assessment practitioners.

3.1.1 Research Design

The Q-matrix measured 5 attributes with test lengths of 25 and 40 items. The 25-item Q-matrix design followed Jiang (2020); the 40-item Q-matrix was created by repeating the first 15 items while ensuring balanced attribute measurement. For generality, we used the saturated G-DINA model. Sample sizes were set at two levels: 1,000 and 2,000. Item parameters followed previous research (Li et al., 2022): let $P(1)$ and $P(0)$ represent the probabilities of correctly answering item i for examinees who have mastered all measured attributes versus those who have mastered none. Item quality had two levels: high-quality items with $P(0) \sim U(0.05, 0.25)$ and $P(1) \sim U(0.75, 0.95)$, and low-quality items with $P(0) \sim U(0.05, 0.4)$ and $P(1) \sim U(0.6, 0.95)$.

We simulated Q-matrices with error rates of 0%, 10%, 20%, and 30%, randomly generating erroneous Q-matrices at each level. Attribute hierarchy relationships were set as four types: linear, convergent, divergent, and unstructured, as shown in Figure 3 [Figure 3: see original paper]. True attribute mastery patterns were

generated under each hierarchy, and response data were simulated using the specified CDM and item parameters.

After generating data, we used HDCM to estimate examinees' attribute mastery patterns (via the R package *GDINA*), specifying attribute hierarchies with missing or redundant edges based on the true hierarchies. These simulated hierarchies served as initial values for estimating attribute mastery patterns. We then applied the BN method to jointly correct attribute hierarchies and Q-matrices. For missing edges, we randomly selected attribute pairs and removed their connections. For redundant edges, in divergent and unstructured hierarchies, redundant connections existed among the 5 attributes (e.g., in divergent structure: (A2,A3), (A4,A5), (A3,A4), (A3,A5); in unstructured: multiple possible pairs). Linear structures cannot generate valid redundant edges due to complete connectivity through intermediate nodes, and convergent structures have only one valid redundant edge (A2,A3), which was excluded from redundancy simulations due to random generation requirements. Therefore, missing-edge scenarios included all four hierarchy types, while redundancy scenarios included only divergent and unstructured types.

Figure 3. Four true attribute hierarchy relationships for 5 attributes and divergent structure for 7 attributes used in simulation studies

3.1.2 Evaluation Metrics

For missing-edge scenarios, the simulation design included: 2 (test length) × 2 (sample size) × 2 (item quality) × 4 (hierarchy type) × 4 (Q-matrix error rate) = 128 conditions. For redundancy scenarios: 2 (test length) × 2 (sample size) × 2 (item quality) × 2 (hierarchy type) × 4 (Q-matrix error rate) = 64 conditions. For each condition, we randomly generated 100 datasets, estimated attribute mastery patterns using initial values with errors, and applied BN conditional independence testing combined with model-data fit indices to obtain corrected hierarchies and Q-matrices. Correction accuracy for attribute hierarchies was calculated as the percentage of 100 simulations achieving complete match with the true hierarchy. Q-matrix correction accuracy used Pattern Classification Rate (PCR) for item measurement patterns and Average Attribute Classification Rate (AACR) for individual attributes, plus True Positive Rate (TPR) and True Negative Rate (TNR). TPR is the rate of correctly correcting misspecified attributes; TNR is the rate of retaining correctly specified attributes (detailed results available from authors). All experiments were repeated 100 times, and averages were computed.

3.1.3 Results

Tables 3 and 4 show BN correction results for attribute hierarchies. When the Q-matrix is completely correct, the algorithm achieves 100% correct correction across all conditions, regardless of item quality, hierarchy type, sample size, or test length. As Q-matrix error rate increases, correction accuracy gradually

decreases. At 10% Q-matrix error, most conditions still achieve $ $90% complete correction, except for isolated cases. At 20% error, test length and sample size significantly improve correction accuracy, with most conditions achieving $ $80% complete correction, especially with low item quality. Among hierarchy types, unstructured hierarchies show greater sensitivity to item quality and Q-matrix error at low sample sizes, while BN performance is similar across other types. No substantial differences appear between missing-edge and redundancy error types.

Table 3. Number of completely correct corrections (out of 100) for 4 hierarchy types under missing-edge scenarios

Table 4. Number of completely correct corrections (out of 100) for 2 hierarchy types under redundancy scenarios

Tables 5 and 6 show Q-matrix correction accuracy (PCR and AACR) when attribute hierarchies contain missing or redundant edges. At 10% Q-matrix error, BN achieves high correction accuracy across all hierarchy types, error types, item qualities, sample sizes, and test lengths: PCR $ $70% and AACR $ $95% in most cases, even with low item quality. At 20% error, with high item quality, PCR remains $ $70% and AACR $ $92%; with low item quality, PCR $ $50% and AACR $ $90% in most cases. At 30% error, PCR $ $53% with high item quality and short tests, while AACR remains $ $90% except in isolated cases. Longer tests partially compensate for low item quality, maintaining PCR $ $50% except for unstructured hierarchies. Sample size substantially affects accuracy: increasing from 1,000 to 2,000 notably improves correction rates across conditions. Test length also significantly impacts accuracy, particularly when Q-matrix error rates are high or item quality is low, because longer tests provide more evidence for estimating initial knowledge states.

Table 5. Q-matrix correction accuracy (PCR and AACR) under missing-edge scenarios

Table 6. Q-matrix correction accuracy (PCR and AACR) under redundancy scenarios

3.2.1 Research Design

To further examine BN performance with more attributes, we set attribute number to 7, using divergent structure based on Simulation 1 results showing minimal accuracy differences across hierarchy types. Test length was fixed at 40 items and sample size at 2,000. While Simulation 1 limited hierarchy errors to one edge, this study examined more extensive errors: (1) one random missing edge, (2) two random missing edges, (3) one random redundant edge, (4) two random redundant edges, and (5) both missing and redundant edges. Other conditions (item quality, Q-matrix error rates) matched Simulation 1, using the G-DINA model for data generation.

3.2.2 Evaluation Metrics

We examined 5 (hierarchy error types) × 2 (item quality) × 4 (Q-matrix error rates) = 40 conditions, using PCR, AACR, and complete correction counts (out of 100) as in Simulation 1.

3.2.3 Results

Table 7 shows Q-matrix correction accuracy (PCR, AACR) for the 7-attribute divergent structure with sample size 2,000. Across all hierarchy error types, BN Q-matrix correction accuracy is similar. At 10% Q-matrix error, the method achieves high accuracy regardless of item quality. At 20% error, accuracy remains high with high item quality but decreases as item quality drops. At 30% error, accuracy declines with decreasing item quality. Figure 4 [Figure 4: see original paper] shows complete correction counts for attribute hierarchies across conditions. With correct Q-matrices, BN achieves 100% correction regardless of item quality. As Q-matrix error increases and item quality decreases, complete correction counts decline. Notably, even with 7 attributes, BN effectively identifies hierarchy specification errors when Q-matrix error rates are moderate ($ $20%) or item quality is high.

Table 7. Q-matrix correction accuracy for 7-attribute divergent structure

Figure 4. BN correction accuracy for 7-attribute divergent hierarchy across experimental conditions (complete correction counts out of 100)

4 Application of the Joint Validation Method to Empirical Data

We further validated the method using empirical data from the Examination for the Certificate of Proficiency in English (ECPE). Templin et al. (2014a) used ECPE data to develop HDCM and identify real attribute hierarchies. The ECPE data include 28 items measuring three attributes: lexical rules (a), integration rules (b), and grammatical rules (c). According to Templin et al. (2014a), these three attributes follow a linear relationship: a→b→c. The Q-matrix and examinee data are publicly available in the *CDM* R package.

The BN validation process was: First, construct a saturated BN with directed edges between all 3 attribute pairs (following the initial theoretical relationship: a→b, b→c, a→c) and between each item and each attribute (with direction uniformly set from attribute to item). This saturated model enables complete testing of all possible item-attribute and attribute-attribute connections, as in the simulation studies. Next, estimate examinees' mastery states for the three attributes from the empirical data, then input both response data and estimated mastery states into the network for conditional independence testing of each edge. This process essentially tests the fit between the theoretically constructed model (including both Q-matrix and attribute hierarchy) and empirical data, providing information to refine the original theoretical model. However,

in empirical applications, model modification should combine statistical results with domain expert judgment.

After jointly testing the initial attribute hierarchy and Q-matrix with BN, we found that some originally specified edges were statistically validated while others showed low existence strength, suggesting conditional independence between nodes. A few item-attribute connections not present in the original Q-matrix showed dependency after BN testing. The initial BN testing revealed attribute relationships of a→c and b→c. We then invited experts to discuss these controversial edges. Combining BN results with expert judgment, the final hierarchy remained a→b→c, while the Q-matrix required some corrections (detailed correction process available from authors). Finally, we re-estimated examinee attribute mastery patterns using the validated hierarchy and corrected Q-matrix, then compared the initial and corrected models using model-data fit indices. BIC results showed the new model (maintaining the original hierarchy with Q-matrix corrections) fit the data better after statistical testing and expert judgment. Thus, BN conditional independence testing can jointly validate attribute hierarchies and Q-matrices, providing valuable information for model optimization.

BN Model Fit Indices After Empirical Data Correction

Attribute Hierarchy	Initial Model	BN-Corrected Model
a→b→c	[Initial BIC]	[Corrected BIC]
a→c, b→c	[Initial BIC]	[Corrected BIC]

5 Conclusion and Discussion

This study proposes a Bayesian network-based method using conditional independence tests to jointly validate attribute hierarchies and Q-matrices. Two simulation studies systematically examined how sample size, Q-matrix error rate, item quality, test length, hierarchy type, hierarchy error type, and attribute number affect BN validation accuracy. Results show that when the Q-matrix is completely correct, BN perfectly corrects attribute hierarchies across all conditions. When Q-matrix error rates are medium or low ($ \$20\%$), BN effectively corrects both components across hierarchy types and error types, though accuracy declines as error rates increase and item quality decreases. Sample size significantly impacts correction accuracy: results with $N = 2{,}000$ are notably better than with $N = 1{,}000$. Test length also substantially affects accuracy, particularly when Q-matrix error rates are high or item quality is low, because longer tests provide more evidence for estimating initial knowledge states. The method performs well even with more attributes, especially when Q-matrix error rates are low and item quality is high. Different hierarchy error types have minimal impact on correction accuracy.

Based on these findings, we recommend ensuring adequate item quality and

sample size when jointly correcting attribute hierarchies and Q-matrices, and expanding test length when item quality is assured. Although we examined high Q-matrix error rates (30-40%), such rates in practice typically warrant Q-matrix redevelopment rather than correction. High error rates were included in simulations to directly examine method efficacy.

To explore practical extensibility, we applied the algorithm to empirical data. Results demonstrate that BN provides valuable reference information for correcting expert-specified hierarchies and Q-matrices, and the corrected model shows better fit when combined with expert judgment. Notably, like all Q-matrix correction methods, BN cannot solely determine final specifications in practice—even with demonstrated accuracy in simulations, controversial items or attributes still require domain expert judgment.

A key point is that BN validation requires initial estimation of attribute mastery patterns, which we implemented using traditional CDMs in simulations. In practice, appropriate diagnostic models should be selected based on empirical data, though BN can also estimate patterns directly (Wang et al., 2021). Current direct estimation requires specialized BN software (e.g., Netica); future research could integrate Netica with R for a more generalized algorithm independent of traditional CDMs and model selection issues. This study used iterative validation, but pilot studies show diminishing returns: stable corrections are typically achieved after 3 iterations, and more iterations do not significantly improve accuracy. In empirical studies, iteration number can be set based on specific circumstances—if a single validation provides sufficient information, it can be combined with expert judgment for final determination. Despite joint validation, BN maintains good computational efficiency: for test length 25, sample size 2,000, and 3 iterations, one joint validation takes 123 seconds. Li et al. (2022) reported 147 seconds for their Gini coefficient method (test length 20, sample size 300). Thus, despite jointly correcting hierarchies and Q-matrices, BN maintains comparable time complexity to single Q-matrix correction methods.

Previous research (Yu et al., 2011) used BN structure learning (K2 algorithm) for data-driven hierarchy exploration, but their method estimated attribute mastery patterns by performing "OR" operations on attribute vectors of correctly answered items—a rudimentary approach that may fail with complex models. Moreover, this estimation depends on the prespecified hierarchy used for Q-matrix development. While Templin and Bradshaw (2014) proposed LR tests for hierarchy validation, their method tests entire hierarchies through model comparison without examining correction accuracy. In contrast, our BN method provides data-driven correction for any pair of attributes simultaneously—one conditional independence test yields validation results for all edges. Future research could compare BN and LR methods for hierarchy correction when Q-matrices are correct.

Currently, no method jointly validates both Q-matrices and attribute hierarchies. Studies estimating both from noisy data without prior information (Wang

& Lu, 2021; Ma et al., 2022) have strong statistical foundations but obvious computational complexity, using only 3-4 attributes. In practice, domain experts provide substantial prior knowledge, making a method to validate and correct existing specifications more valuable. As no baseline method exists for joint validation, we compared BN Q-matrix correction accuracy against the Stepwise method (Ma & de la Torre, 2020) under attribute independence, finding BN advantages at high Q-matrix error rates (results available from authors).

References

Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transactions on Automatic Control, 19*(6), 716–723. http://dx.doi.org/10.1109/TAC.1974.1100705

Chen, J. (2017). A Residual-Based Approach to Validate Q-Matrix Specifications. *Applied Psychological Measurement, 41*, 014662161668602. doi:10.1177/0146621616686021

Chiu, C.-Y. (2013). Statistical Refinement of the Q-Matrix in Cognitive Diagnosis. *Applied Psychological Measurement, 37*(8), 598-618. doi:10.1177/0146621613488436

Cui, Y. (2007). *The hierarchy consistency index: Development and analysis*. Doctoral Dissertation. University of Alberta.

Cui, Y., & Leighton, J. P. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement, 46*(4), 429-449.

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*(4), 346−362.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*(1), 115–130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179–199.

de la Torre, J., & Chiu, C. Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*(2), [pages].

Ding, S.L., Mao M.M., Wang, W.Y., Luo, F., & Cui.Y.(2012). Evaluating the Consistency of Test Items Relative to the Cognitive Model for Educational Cognitive Diagnosis. *Acta Psychologica Sinica, 44*(11), 1535-1546.

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

Jiang Yu.(2020). *Research On The Test Method Of Attribute Hierarchy Based*

On Information Matrix. Unpublished doctoral dissertation, Beijing Normal University.

Kang, C. H., Yang, Y. K., & Zeng, P. H. (2019). Q-matrix refinement based on item fit statistic RMSEA. *Applied Psychological Measurement, 43*(7), 527–542.

Leighton, J. P., & Gierl, M.J. (Eds.). (2007). *Cognitive Diagnostic Assessment for Education: Theory and Practices*. Cambridge: Cambridge University Press.

Li, J., Mao, X., & Wei, J. (2022). A simple and effective new method of Q-matrix validation. *Acta Psychologica Sinica, 54*(8), 996–1008.

Liu, J., Xu, G., & Ying, Z. (2012). Data-Driven Learning of Q-Matrix. *Applied Psychological Measurement, 36*(7), 548-564. doi: 10.1177/0146621612456591

Liu, R., Huggins-Manley, A. C., & Bradshaw, L. (2017). The impact of Q-matrix designs on diagnostic classification accuracy in the presence of attribute hierarchies. *Educational and Psychological Measurement, 77*(2), 220–240.

Liu, Y., Wu Q. (2023). An empirical Q-matrix validation method using a complete information matrix in cognitive diagnostic models. *Acta Psychologica Sinica, 55*(1), 142-158.

Ma, W., & de la Torre, J. (2020). An empirical Q-matrix validation method for the sequential G-DINA model. *British Journal of Mathematical and Statistical Psychology, 73*, 142-163.

Ma, C., Ouyang, J., & Xu, G. (2022). Learning Latent And Hierarchical Structures In Cognitive Diagnosis Models. *Psychometrika.88*(1),175-207. doi.org/10.1007/s11336-022-09867-5

Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics,4*(1), 12–35. doi:10.1080/10618600.1995.10474663

Rupp, A. A., Templin, J. (2008).The effects of Q-matrix misspecification on parameter estimates and classification accuracy in DINA model. *Educational and Psychological Measurement. 68*(1),78-96

Scutari, M., & Denis, J. B. (2021). *Bayesian networks with examples in R*. New York: Chapman and Hall/CRC.

Schwaz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461–464. http://dx.doi.org/10.1214/aos/1176344136

Tatsuoka，K.K.(1983).Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement,20*(4)，345-354.

Tatsuoka, K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics, 12*, 55-73.

Terzi, R., & de la Torre, J. (2018). An iterative method for empirically-based Q-matrix validation. *International Journal of Assessment Tools in Education, 5*(2), 248–262.

Templin, J., & Bradshaw, L. (2014a). Hierarchical diagnostic classification models: a family of models for estimating and testing attribute hierarchies. *Psychometrika, 79*(2), 317-339.

Templin, J. , & Bradshaw, L. (2014b). The use and misuse of psychometric models. *Psychometrika, 79*(2), [pages].

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.

Wang, W., Song, L., Ding, S., Meng, Y., Cao, C., & Jie, Y. (2018). An EM-Based Method for Q- Matrix Validation. *Applied Psychological Measurement, 42*(6), 446-459. doi:10.1177/0146621617752991

Wang, C., & Gierl, M. (2011). Using the Attribute Hierarchy Method to Make Diagnostic Inferences about Examinees' Cognitive Skills in Critical Reading. *Journal of Educational Measurement, 48*(2), 165-187.

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.

Wang, D., Gao, X., Cai, Y., & Tu, D. (2020). A method of Q-matrix validation for polytomous response cognitive diagnosis model based on relative fit statistics. *Acta Psychologica Sinica, 52*(1), 93–106.

Wang, C., & Lu, J. (2021). Learning Attribute Hierarchies From Data: Two Exploratory Approaches. *Journal of Educational and Behavioral Statistics.* doi.org/10.3102/1076998620931094

Xue, W. & Chen, H. G.(2012). *Data mining based on Clementine.* China Renmin University Press.

Xu, G. & Shang, Z. (2018). Identifying latent structures in restricted latent class models. *Journal of the American Statistical Association.* doi.org/10.1080/01621459. 2017.1340889

Yu, X. F., & Cheng, Y. (2020). Data-driven Q-matrix validation using a residual-based statistic in cognitive diagnostic assessment. *British Journal of Mathematical and Statistical Psychology, 73*(Suppl 1), 145–179.

Yu X.F., Ding, S.L., Qin C.Y., & Lu., Y.N.(2011). Application of Bayesian Networks to Identify Hierarchical Relations Among Attributes in Cognitive Diagnosis, *Acta Psychologica Sinica, 43*(03), 338-346.

Zhang, L.W. & Guo, H.P.(2006). *Introduction to Bayesian networks.* Science Press, Beijing.

Abstract

Cognitive diagnostic models (CDMs) are developed to diagnostically evaluate subjects' cognitive strengths and weaknesses based on the Q-matrix mappings of their items and attributes. The traditional calibration of cognitive attributes in Q-matrix mainly relies on the subjective judgment of experts. Due to the subjective process of Q-matrix construction, there inevitably are more or less misspecifications in the Q-matrix, which, if left unchecked, may result in a serious negative impact on cognitive diagnostic assessment. From another important perspective, in the empirical applications of CDMs, cognitive attributes generally do not operate independently but rather belong to an interrelated network, and a certain psychological order, logical order, or hierarchical relationship may be present among the cognitive attributes. The correctness of both the Q-matrix and the attribute hierarchy significantly impacts the parameter estimation ability of a CDM and the accuracy of the examinee's classification result. Recently, considerable studies have developed approaches for validating Q-matrices or testing attribute hierarchies respectively. However, there is no method that can validate both a Q-matrix and an attribute hierarchy simultaneously. From the empirical application perspective, an approach that can simultaneously validate both a prespecified Q-matrix and an attribute hierarchy is more desirable.

An approach based on Bayesian networks (BN) for validating both Q-matrices and attribute hierarchies simultaneously is proposed in this research. To explore the performance of the BN method, this article conducted two simulation studies and empirical data analysis to theoretically and practically evaluate the accuracy of the Q-matrix validation and attribute hierarchy correction processes. The correctness of each element in the Q matrix and the attributes hierarchy can be checked by testing the strength of edge existence in the network structure.

When validating attribute hierarchy relationships and Q-matrix jointly in the first simulation, we explore the effects of Q-matrix error rate, item quality, test length, sample size, and the attribute hierarchy type on the correction accuracy of both the Q-matrix and the attribute hierarchy.

The results show that the BN method can effectively correct the Q-matrix and attribute hierarchy simultaneously when the error rate of the Q-matrix is at a medium or low level, especially when the item quality is high or the sample size is sufficient or the test length is long, the accuracy of the correction is generally high. As the Q-matrix error rate increases and the quality of the items decreases, the correction accuracy gradually decreases. The BN method can correct the attribute hierarchies exactly right when the Q matrix is correct. The results in the second simulation show that when the attribute number in the Q-matrix increases, the BN method is still performing well.

Different types of attribute hierarchy errors have a small impact on the correction accuracy across different conditions. The effectiveness of the BN method in the empirical dataset was demonstrated by the better model data fit index of BIC.

In conclusion, the initial specified Q-matrix and attribute hierarchy can be simultaneously validated via the BN method. Then the corrected Q-matrix and the refined attribute hierarchy obtained from the data-driven BN method can again be combined with the theoretical judgments of experts to obtain a more optimized model, finally achieving more accurate diagnostic outcomes in CDA practice.

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