

CDM Q-Matrix Discovery with LLMs: Fusing Domain Knowledge with Empirical Evidence

Susu Zhang

University of Illinois Urbana-Champaign
szhan105@illinois.edu

V.N. Vimal Rao

University of Illinois Urbana-Champaign
raovnv@illinois.edu

Abstract

We present a method for diagnostic assessment Q-matrix specification combining LLM input with model-based empirical validation. An LLM generates an initial Q-matrix from item content, and response data guide subsequent Q refinement. The approach integrates substantive rationale with empirical evidence to support scalable and measurement theory-grounded diagnostic assessment.

1 Introduction

Diagnostic assessments are designed to measure examinees' multidimensional profiles on fine-grained, discrete attributes, such as the skills, procedures, and knowledge required to solve assessment items (Rupp et al., 2010). Compared with a single continuous proficiency score, an attribute-mastery profile provides more specific information about students' strengths and weaknesses and may therefore support targeted remediation, instructional decisions, and personalized feedback. Cognitive diagnosis models (CDMs), also referred to as diagnostic classification models, provide a psychometric framework for relating these latent mastery profiles to observed item responses.

Central to diagnostic assessment is specifying the Q-matrix, a binary item-by-attribute incidence matrix indicating which attributes are required by each item. The Q-matrix specifies the relationship between attribute mastery and item responses and thereby determines the substantive interpretation of the resulting diagnostic classifications. Misspecification can affect both item parameter estimation and examinee classification (Rupp and Templin, 2008). A defensible Q-matrix should be grounded in both the cognitive demands implied by item content and empirical fit on the observed response data.

Traditional Q-matrix development relies on cognitive task analysis conducted by subject-matter specialists. These specialists map each item to a

set of attributes and, where relevant, specify relationships such as attribute hierarchies. To support consistent coding by trained raters, the attributes are accompanied by clear labels, operational definitions, and rules for assigning item-attribute links. Think-aloud evidence from examinees can be used to assess the proposed attributes and to revise them when necessary. Support for a Q-matrix may therefore combine prior literature, expert ratings or panel discussion, evidence about examinees' solution processes, and subsequent validation using response data (Gierl et al., 2008; Li and Suen, 2013; Akbay et al., 2018). Coordinating these inputs can require subject-matter and measurement expertise, specialized data collection, and considerable time, making it costly for routine classroom assessments.

Empirically driven approaches aim to learn item-attribute relations from response data, addressing both the substantial human input required for expert construction and the possibility of misspecification despite careful review. These approaches are commonly divided into Q-matrix refinement methods and direct estimation methods (Liu and Kang, 2019; Zhang et al., 2023). Refinement methods begin with an expert-specified Q-matrix and use item-level diagnostics to identify potentially misspecified entries while preserving substantive interpretability. Examples include parameter-based validation under the deterministic-input, noisy-and-gate (DINA) model (Junker and Sijtsma, 2001; de la Torre, 2008), residual-based refinement (Chiu, 2013), and extensions to more general CDMs or methods that favor parsimonious revisions through fit-parsimony criteria or regularization (de la Torre and Chiu, 2016; Nájera et al., 2021). Expert judgment is therefore involved both in constructing the starting matrix and in deciding whether proposed revisions are substantively defensible. Direct estimation methods infer an unknown Q-matrix from response data (e.g., Liu et al., 2012, 2013; Chen et al., 2018). Dependable recovery may require

more sample information than is available in small or moderately sized classroom datasets, especially early in an assessment’s deployment. The estimated binary incidence matrix also requires post-estimation interpretation to identify the cognitive processes represented by its columns and explain particular item-attribute associations: Similar to exploratory factor analysis, the columns (here, attributes) are not inherently interpretable.

Recent work has examined the emerging capabilities of general-purpose large language models (LLMs) for psychometric tasks, including automated scoring, automated item generation, and response simulation (e.g., [Gorgun and Bulut, 2024](#); [Suk et al., 2025](#)). Conceptual and review papers have also discussed their potential utility for formative assessment, including assistance with item writing and more timely, customized instructional feedback ([Chang, 2024](#); [Hao et al., 2024](#); [Ho, 2024](#)). Meanwhile, LLM outputs vary across repeated generations and prompt specifications, and their accuracy requires empirical verification when they inform measurement decisions.

Within cognitive diagnosis, recent studies have examined content-based Q-matrix construction with LLMs from item and attribute descriptions and evaluated the resulting matrices through expert comparison and empirical model fit ([Aşiret and Sünbül, 2025](#); [Xue and Appleton, 2026](#)). Two key challenges remain for classroom use. First, an LLM’s content-based proposal needs to be evaluated against examinee response patterns and revised when the proposed item-attribute relations are inconsistent with those patterns. Second, the initial specification and each empirical revision need to remain interpretable so that teachers and domain experts can evaluate whether the proposed relations reflect the skills required by the items.

The current work draws from automated statistical model discovery with language models ([Li et al., 2024](#); [Zhang et al., 2025](#)) and proposes a method for Q-matrix discovery under the DINA model that integrates LLM-generated proposals based on domain knowledge and item content with empirical evidence from examinee responses. The LLM first proposes a Q-matrix from the item content and attribute descriptions and provides a brief rationale for each item Q-vector. After the DINA model is fitted, an item-level Q-matrix validation statistic identifies alternative Q-vectors that are more consistent with the observed responses. These alternatives are returned to the LLM for evaluation against

the item content, and the LLM provides a rationale for each decision to revise or retain the flagged Q-vector. Global model fit is then used to select among the Q-matrices visited during refinement. This study makes the following contributions:

- An LLM-assisted algorithm for initializing the Q-matrix from item content and refining the initial specification using empirical response evidence under the DINA model;
- Support for substantive interpretation and expert review through brief LLM-generated rationales for the initial item Q-vectors and subsequent refinement decisions;
- Evaluation of the procedure with (1) a proof-of-concept simulation study and (2) an empirical application from a large university general-education statistics course. The empirical evaluation combines quantitative evaluation (e.g., expert agreement, model fit) with expert evaluation of the LLM-generated rationales.

2 Method

Let $\mathbf{Y} = (Y_{ij})$ be the $N \times J$ matrix of binary responses and $\mathbf{Q} = (q_{jk})$ the $J \times K$ binary Q-matrix, where $q_{jk} = 1$ indicates that item j requires attribute k , and $q_{jk} = 0$ otherwise. The procedure begins with an LLM proposal based on item content and then alternates between Bayesian DINA estimation, item-level Q-vector validation, and LLM refinement of the Q-vector on empirically flagged items.

2.1 Initial Q-Matrix Proposal

To obtain the initial LLM-proposed Q-matrix, we designed a prompt around three components recommended for LLM-based Q-matrix construction: a clearly specified role and task, the content needed to make item-attribute judgments, and explicit constraints on the requested output ([Xue and Appleton, 2026](#)). The LLM was assigned the role of a subject-matter expert for the assessment domain and target population and asked to construct an initial Q-matrix. For each item, it received the stem, response options, correct answer, and any table or figure needed to solve the item. Each attribute was represented by its label, definition, and examples in the assessment domain.

The prompt’s item-coding instructions were informed by expert Q-matrix development practices, in which an item is solved, the necessary steps are

identified, and those steps are mapped to the specified attributes (Gierl et al., 2008; Rupp et al., 2010). The LLM was therefore asked to infer a defensible solution and code only attributes required by that solution. The prompt provided brief coding guidance, distinguishing required skills from incidental/helpful skills, and imposing additional constraints (e.g., attribute coverage, number of single-attribute items, and distinguishable columns) for DINA model identifiability (Gu and Xu, 2019). The full prompt and application-specific constraints are reported in Appendix F.

The LLM returned a binary $J \times K$ Q-matrix and a brief rationale for every item’s Q-vector. Each rationale linked the coded attributes to the steps that made them necessary, and a final consistency check required the rationale and Q-vector to identify the same attributes. Although only the Q-matrix is used for subsequent model fit, the rationale provides an explanation to the to the initial proposal and allowed substantive review.

2.2 Bayesian DINA Estimation and Q-Validation

Given a proposed Q-matrix, a DINA model is fitted using the Bayesian Gibbs sampling algorithm of Culpepper (2015). The DINA model is a simple conjunctive model, where an examinee answers an item correctly with high probability only if they mastered all requisite skills of the item. For examinee i and item j , let α_i denote the binary attribute profile and define the ideal response

$$\eta_{ij}(\alpha_i, \mathbf{q}_j) = \prod_{k=1}^K \alpha_{ik}^{q_{jk}}. \quad (1)$$

The conditional correct-response probability is

$$P(Y_{ij} = 1 \mid \alpha_i, \mathbf{q}_j) = (1 - s_j)^{\eta_{ij}} g_j^{1 - \eta_{ij}}, \quad (2)$$

where s_j and g_j are the slipping and guessing parameters. Each chain was run for 10,000 iterations with the first 5,000 as burn-in.

In practice, item-level Q-matrix validation statistics identify items for expert review and suggest alternative Q-vectors that agree more closely with the response data. Experts can then examine whether the suggested attributes are supported by the item content and revise the Q-matrix accordingly. We implemented this process using a Bayesian adaptation of the residual sum of squares (RSS) statistic proposed for the DINA model by Chiu (2013) for

Q-matrix validation. For every item j and every nonzero candidate Q-vector $\mathbf{q}_m \in \{0, 1\}^K$, let $\eta_{im}^{(s)}$ denote the ideal response implied by the sampled attribute profile $\alpha_i^{(s)}$ at posterior draw s . We calculated

$$\widehat{\text{RSS}}_{jm} = \frac{1}{S} \sum_{s=1}^S \sum_{i=1}^N \left(Y_{ij} - \eta_{im}^{(s)} \right)^2. \quad (3)$$

Intuitively, $\widehat{\text{RSS}}_{jm}$ captures the posterior discrepancy between the ideal responses implied by candidate Q-vector $\mathbf{q}_m$ and the observed responses on item j , where smaller values indicate closer agreement with the observed item responses.

Relative overall fit was evaluated using the widely applicable information criterion (WAIC; Watanabe, 2010). For posterior draws $\Omega^{(1)}, \dots, \Omega^{(S)}$ and examinee response vector $\mathbf{y}_i$,

$$\text{WAIC} = -2 (\text{lppd} - p_{\text{WAIC}}), \text{ where} \quad (4)$$

$$\text{lppd} = \sum_{i=1}^N \log \left[\frac{1}{S} \sum_{s=1}^S p(\mathbf{y}_i \mid \Omega^{(s)}) \right], \quad (5)$$

$$p_{\text{WAIC}} = \sum_{i=1}^N \text{Var}_s \left[\log p(\mathbf{y}_i \mid \Omega^{(s)}) \right]. \quad (6)$$

Lower WAIC values indicate better expected predictive fit relative to the other fitted Q-matrices.

2.3 Model Critique and Refinement

After each fit, a critique was generated for the previously visited Q-matrix with the lowest WAIC. An item was flagged whenever at least one candidate Q-vector had a lower RSS than its current vector. For each flagged item, the current vector, the RSS-minimizing alternative, and their differences in RSS values were presented as an empirical response-based critique of the current Q-vector.

The refinement prompt included only the flagged items, accompanied by their item content, current Q-vectors, attribute definitions, structural constraints, and RSS-based statistical critique. The LLM was instructed to solve each flagged item, decide which attributes were required, compare that with the RSS-suggested vectors, and avoid accepting a change solely because it reduced RSS. The LLM could retain the current Q-vector, adopt the RSS-minimizing vector, or return a different content-based or partially-revised vector when only part of the RSS-suggested change was supported by the item content.

In addition to the revised Q, for each flagged item, the refinement prompt instructed the LLM to provide a short rationale that began by indicating whether the current Q-vector was changed or unchanged and then explained that decision, i.e., the item content or solution step supporting the revised attribute assignment, or why the current coding better represented the item content than the RSS-suggested alternative. The prompt also instructed the LLM to perform a self-consistency check so that the rationale was consistent with attributes required by the revised Q.

To ensure that the search advanced, each refinement call was required to alter at least one flagged row. The complete refinement prompt template is provided in Appendix G. The refined Q-matrix is subsequently fit to the response data to obtain the RSS-based critique for the next iteration.

2.4 Stopping and Final Selection

At each iteration, the critique and refinement were based on the lowest-WAIC Q-matrix visited so far. The maximum number of iterations was set as 10, and the algorithm stops earlier if no item was flagged or the same Q-matrix was produced in three consecutive iterations. The matrix with the lowest WAIC across all visited iterations was selected as the final refined Q-matrix.

2.5 LLM Implementation Settings

All LLM calls used GPT-5.4 through API call with medium reasoning effort and temperature 1. GPT-5.4 was selected because it supported both textual and image inputs and inference-time reasoning. Comparison of foundation models was outside the scope of the study. The same model settings were used for initial proposals and refinements in both applications. API responses were requested as R objects, checked for formatting, and regenerated after a parsing or output-validation failure.

3 Fraction-Subtraction Simulation

3.1 Simulation Design

We examine the proposed methods with a proof-of-concept simulation study, using the item content of the 20 fraction-subtraction items originally reported by Tatsuoka (1984), and a four-attribute ground-truth Q-matrix for the fraction items reported by Chen et al. (2018), who learned the matrix from empirical response data. Köhn et al. (2025) subsequently interpreted the four at-

tributes as finding a common denominator, borrowing from an integer, subtracting integer and fractional parts, and converting a whole number to a fraction. Among the Qs proposed for the fraction-subtraction items, we used this specification because it met the DINA identifiability conditions.

We conducted 50 simulation replications. In each replication, we generated a binary response matrix for 536 examinees and 20 items under the DINA model using the four-attribute ground-truth Q-matrix. The sample size matched the empirical fraction-subtraction dataset, item-specific slipping and guessing parameters were set to the estimates reported by Köhn et al. (2025), and true attribute profiles were sampled uniformly from the 2^4 classes. In each replication, the LLM received the original item content and answer key together with the four attribute definitions and examples to propose an initial Q-matrix. Subsequent refinement incorporated the response data generated under the ground-truth Q. The inputs supplied to the LLM are reported in Appendix A. Note that the item text was developed independently of the data-generating Q-matrix, and the item content itself does not have a known ground-truth Q-matrix. The evaluation therefore examined whether incorporating response evidence enabled the algorithm to recover the Q-matrix used to generate the simulated responses.

3.2 Evaluation Criteria

For each replication, Q-matrix recovery was evaluated for the initial LLM proposal and the final matrix selected by WAIC. Entry-wise accuracy was the proportion of all $J \times K$ entries correctly recovered. Sensitivity was the proportion of true entries equal to 1 that were recovered as 1, and specificity was the proportion of true entries equal to 0 that were recovered as 0. We report the mean and standard deviation of each measure across replications. For each item, we also calculated the proportion of replications in which it was flagged by RSS, the proportion of flags followed by an LLM revision, and the proportion of replications in which the final estimated Q-vector matched the data-generating Q-vector.

3.3 Results

As shown in Figure 1, global model fit and Q-matrix recovery generally improved over refinement iterations. Table 1 shows that mean entry-wise accuracy increased from .869 for the initial proposals to .975 for the final matrices. Mean sen-

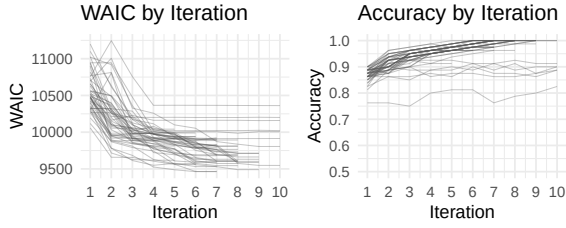


Figure 1: Fraction simulation under DINA: WAIC and Q-matrix entry-wise accuracy across refinement iterations. Each line represents a replication.

	Accuracy	Sensitivity	Specificity
Initial Q	.869	.785	.922
SD	(.026)	(.052)	(.020)
Final Q	.975	.963	.983
SD	(.040)	(.057)	(.032)

Table 1: Recovery of the data-generating Q-matrix across 50 fraction-subtraction simulation replications.

sitivity increased from .785 to .963, and specificity increased from .922 to .983. Item-level results are reported in Appendix B. Although every replication used the same item content, attribute definitions, and LLM settings, many entries of the initial Q-matrices varied across replications, producing mean entry values farther from 0 or 1. The final Q-matrices contained fewer entries with mean absolute deviations greater than .20 from the data-generating values, shown in bold in Appendix B, indicating higher item-level accuracy. Their mean entry values were also more concentrated near 0 or 1, indicating lower variation across replications after incorporating response evidence.

4 Statistics Classroom Assessment

4.1 Data Description

The empirical application analyzed the first midterm examination administered in Fall 2025 in a large general-education introductory statistics course at a Midwestern public university. The examination occurred in Week 5 of a 15-week term and assessed concepts involving foundational principles of data, measurement, and descriptive statistics. Its 25 multiple-choice items covered data structure, variable types, relative and conditional frequencies, numerical summaries and visualizations, characteristics of quantitative distributions, comparisons across groups, and associations represented in two-way tables and scatterplots.

The course instructor identified eight attributes that the examination was intended to measure and

specified two alternative organizations of those attributes. The attribute labels and definitions are detailed in Table 8. In version 1, A3 represented summaries and graphs for a single quantitative variable, whereas A4 represented categorical frequencies and bar graphs. In version 2, A3 represented numerical summaries across quantitative and categorical variables, whereas A4 represented visual displays across both variable types. The definitions of A1, A2, and A5–A8 were identical. The instructor also specified a corresponding expert Q-matrix under each version, serving as a substantive reference for comparison with the learned Q-matrices. The items were developed from a blueprint aligned with the course content and learning objectives and had been refined across semesters using item response theory analyses of student response data. Appendix C provides item summaries while protecting content that remains in operational use.

For each attribute version, the LLM received the same 25 item stems, response options, and correct answers, together with any table or plot required to solve an item and the corresponding version of the eight attribute definitions. The empirical data contained binary responses from 1,144 students. The same 1144×25 response matrix was used to fit and evaluate the candidate Q-matrices under the DINA model for both attribute versions.

4.2 Evaluation Criteria

The initial and selected final Q-matrices under each attribute version were compared with the corresponding expert Q-matrix. Treating expert-coded entries as the reference, we calculated entry-wise agreement, sensitivity, and specificity. We also fitted the expert Q-matrices using the same Bayesian DINA procedure and compared their WAIC values with those of the LLM initial and final Q-matrices.

For the best-fitting Q, two experts who are university-level statistics instructors independently rated the rationales using four binary criteria that assessed whether each rationale matched the attributes coded in the Q-vector (*attribute alignment*), was factually grounded in the item content (*item grounding*), agreed with the rater’s judgment of the attributes required or not required to solve the item (*substantive accuracy*), and was sufficiently clear and specific for a reviewer to understand the initial coding or refinement decision (*review utility*). Detailed rating instructions and examples of violations are provided in Appendix D.

The initial Q rationale evaluation included all 25

items. The refinement rationale evaluation included the last available rationale for each item flagged along the trajectory leading to the final selected Q-matrix. For each criterion, endorsement was calculated as the proportion of ratings equal to 1 across both raters; inter-rater agreement was the proportion of items receiving identical ratings for the two raters.

4.3 Results

Version 1 terminated at 10 iterations and selected iteration 6 by WAIC, where 10 items were flagged in at least one iteration, and 7 items' Q-vectors differed between the initial and final Q-matrices. Version 2 terminated after five iterations and selected iteration 2, and six items were flagged and three item Q-vectors differed. As shown in Table 2, the final refined Q-matrix had a lower WAIC than both the initial LLM Q-matrix and the expert Q-matrix for both attribute versions. The empirically refined Q under version 1 had the lowest WAIC. The final matrices also had slightly greater agreement with their corresponding expert matrices than the initial proposals. Note that whereas specificity increased, sensitivity was unchanged under version 2 and decreased under version 1.

Measure	Q-matrix	Version 1	Version 2
WAIC	Initial LLM	29167.21	29145.52
	Final refined	29141.98	29142.41
	Expert	29148.19	29145.87
Agreement	Initial LLM	.920	.915
	Final refined	.925	.925
Sensitivity	Initial LLM	.815	.741
	Final refined	.741	.741
Specificity	Initial LLM	.936	.942
	Final refined	.954	.954

Table 2: Bayesian DINA fit and agreement with the expert Q-matrix for the statistics assessment.

The LLM initial and final refinement rationales under version 1 are reported in Appendix E and were evaluated by both experts on the four criteria. Table 3 reports the average endorsement and agreement of the expert ratings. The rationales were consistently rated to be grounded in the item content. Initial-rationale endorsement ranged from .92 to 1.00 across criteria, with inter-rater agreement from .92 to 1.00. Among the 10 refinement rationales, endorsement ranged from .90 to 1.00 and agreement from .80 to 1.00.

Criterion	Initial ($J = 25$)		Refinement ($J = 10$)	
	Endorse.	Agree.	Endorse.	Agree.
Alignment	1.00	1.00	.95	.90
Grounding	1.00	1.00	1.00	1.00
Accuracy	.92	.92	.95	.90
Review utility	.98	.96	.90	.80

Table 3: Expert evaluation of version 1 LLM-generated rationales. Endorsement: proportion of binary ratings equal to 1 across two experts; agreement: agreement rate between raters across items.

5 Discussion

The current study proposes an iterative approach to Q-matrix discovery that integrates LLM judgments based on item content with empirical evidence from item responses. The LLM proposes an initial Q-matrix from the items and attribute definitions, and diagnostics from the fitted Bayesian DINA model are used to generate statistical critique and inform subsequent revisions. Additionally, to maintain interpretability of the learned Qs, the LLM provides brief, item-specific rationales for both the initial attribute assignments and later decisions to revise or retain them. These rationales allow instructors and domain experts to examine the basis for each Q-vector, assess whether the coded attributes reflect the intended item requirements, and evaluate discrepancies between the content-based proposal and empirical diagnostic evidence.

Results from the fraction-subtraction simulation indicate that empirical refinement substantially improved recovery of the data-generating Q-matrix relative to the initial LLM specifications. Each replication used the same item and attribute descriptions, yet stochastic variation in the LLM outputs produced different initial Q-matrices. Incorporating response-based diagnostics led to more consistent final specifications, suggesting that empirical evidence may stabilize content-based proposals in addition to improving average recovery.

For the statistics classroom assessment, the final refined Q-matrix had lower WAIC than the initial LLM Q-matrix and the expert Q-matrix under both attribute versions, indicating better expected predictive fit among the fitted Q-matrices. Refinement also produced a small increase in agreement with the instructor-specified Q-matrices, associated primarily with higher specificity. The expert ratings of the rationales indicated high overall alignment, grounding, substantive accuracy, and review utility for both sets of rationales. For items on which the

expert and refined Q-vectors differ, the item-level rationales and fit diagnostics provide information for evaluating whether the discrepancy reflects misspecification, an alternative solution strategy, or a difference in attribute granularity.

There are several limitations of this study. First, the procedure holds the prespecified attribute set fixed, leaving possible missing, redundant, or overly broad/narrow attributes unexamined. Future extensions could use recurring conflicts between content-based judgments and response-based diagnostics, sparse attribute coverage, or weak empirical item discrimination to prompt structured review of the attribute definitions and proposals to merge, split, remove, redefine, or add attributes, followed by expert review. Second, refinement is conducted under the conjunctive DINA model, so suggested Q-matrix revisions may also reflect misspecification of how attributes interact to affect correct response probability. Future research can extend the workflow to more general CDMs. Finally, the classroom results were based on a single assessment and examinee sample of a single course, and only one foundation model was examined (GPT-5.4). Generalizability should be evaluated across domains, item formats, sample sizes, prompts, and LLMs.

Additional extensions could incorporate attribute hierarchies, with admissible mastery profiles and Q-matrix revisions constrained by prerequisite relations among attributes. The method could also be adapted to open-ended responses, allowing response content, written work, or solution traces to supplement correctness scores when evaluating the proposed item-attribute relations. Such evidence could help experts assess whether the attributed skills correspond to the processes examinees actually used.

As LLM capabilities continue to evolve, the current study provides a practical workflow to harness LLM capabilities to support instructors in implementing of measurement-theory-based best practices in diagnostic classroom assessment.

References

Lokman Akbay, Ragip Terzi, Mehmet Kaplan, and Katibe Gizem Karaaslan. 2018. [Expert-based attribute identification and validation on fraction subtraction: A cognitively diagnostic assessment application](#). *Journal on Mathematics Education*, 9(1):103–120.

Semih Aşiret and Seçil Ömür Sünbül. 2025. Evaluat-

ing ChatGPT-4o's performance in construction of Q-matrix for a cognitive diagnostic assessment. *Research Square*.

- Hua-Hua Chang. 2024. Harnessing AI for educational measurement: Standards and emerging frontiers. *J. Educ. Behav. Stat.*, 49(5):702–708.
- Yinghan Chen, Steven Andrew Culpepper, Yuguo Chen, and Jeffrey Douglas. 2018. [Bayesian estimation of the DINA Q-matrix](#). *Psychometrika*, 83(1):89–108.
- Chia-Yi Chiu. 2013. [Statistical refinement of the Q-matrix in cognitive diagnosis](#). *Applied Psychological Measurement*, 37(8):598–618.
- Steven Andrew Culpepper. 2015. Bayesian estimation of the dina model with gibbs sampling. *Journal of Educational and Behavioral Statistics*, 40(5):454–476.
- Jimmy de la Torre. 2008. An empirically based method of Q-matrix validation for the DINA model: Development and applications. *J. Educ. Meas.*, 45(4):343–362.
- Jimmy de la Torre and Chia-Yi Chiu. 2016. [A general method of empirical Q-matrix validation](#). *Psychometrika*, 81(2):253–273.
- Mark J Gierl, Changjiang Wang, and Jiawen Zhou. 2008. Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in algebra on the sat. *Journal of Technology, Learning, and Assessment*, 6(6):n6.
- Guher Gorgun and Okan Bulut. 2024. Instruction-tuned large-language models for quality control in automatic item generation: A feasibility study. *Educ. Meas. Issu. Pr.*
- Yuqi Gu and Gongjun Xu. 2019. The sufficient and necessary condition for the identifiability and estimability of the dina model. *Psychometrika*, 84(2):468–483.
- Jiangang Hao, Alina A von Davier, Victoria Yaneva, Susan Lottridge, Matthias von Davier, and Deborah J Harris. 2024. Transforming assessment: The impacts and implications of large language models and generative AI. *Educ. Meas. Issu. Pr.*, 43(2):16–29.
- Andrew D Ho. 2024. Artificial intelligence and educational measurement: Opportunities and threats. *J. Educ. Behav. Stat.*, 49(5):715–722.
- Brian W Junker and Klaas Sijtsma. 2001. Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3):258–272.
- Hans-Friedrich Köhn, Chia-Yi Chiu, Olasumbo Oluwalana, Hyunjoon Kim, and Jiayi Wang. 2025. [A two-step Q-matrix estimation method](#). *Applied Psychological Measurement*, 49(1-2):3–28.

- Hongli Li and Hoi K Suen. 2013. [Constructing and validating a Q-matrix for cognitive diagnostic analyses of a reading test](#). *Educ. Assess.*, 18(1):1–25.
- Michael Y Li, Emily B Fox, and Noah D Goodman. 2024. Automated statistical model discovery with language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 27791–27807.
- Jingchen Liu and Hyeon-Ah Kang. 2019. [Q-matrix learning via latent variable selection and identifiability](#). In *Handbook of Diagnostic Classification Models*, pages 247–263. Springer.
- Jingchen Liu, Gongjun Xu, and Zhiliang Ying. 2012. [Data-driven learning of Q-matrix](#). *Applied Psychological Measurement*, 36(7):548–564.
- Jingchen Liu, Gongjun Xu, and Zhiliang Ying. 2013. [Theory of the self-learning Q-matrix](#). *Bernoulli*, 19(5A):1790–1817.
- Pablo Nájera, Miguel A Sorrel, Jimmy de la Torre, and Francisco José Abad. 2021. [Balancing fit and parsimony to improve Q-matrix validation](#). *Br. J. Math. Stat. Psychol.*, 74(S1):110–130.
- Andr’e A. Rupp and Jonathan Templin. 2008. [The effects of q-matrix misspecification on parameter estimates and classification accuracy in the dina model](#). *Educational and Psychological Measurement*, 68(1):78–96.
- Andre A. Rupp, Jonathan Templin, and Robert A. Henson. 2010. *Diagnostic measurement: Theory, methods, and applications*. The Guilford Press, New York, NY.
- Youmi Suk, Chenguang Pan, and Ke Yang. 2025. Using generative AI for sequential data generation in monte carlo simulation studies. *J. Educ. Behav. Stat.*, (10769986251397559).
- Kikumi K. Tatsuoaka. 1984. *Analysis of errors in fraction addition and subtraction problems*. Computer-Based Education Research Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL. Final Report #143; Final report for grant no. NIE-G-81-0002.
- Sumio Watanabe. 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(12).
- Kang Xue and James J Appleton. 2026. Evaluating general-purpose multimodal AI for q-matrix generation from math items: A cognitive diagnostic modeling exploration. *J. Educ. Meas.*, 63(1):e70028.
- Erica Zhang, Ryunosuke Goto, Naomi Sagan, Jurik Mutter, Nick Phillips, Ash Alizadeh, Kangwook Lee, Jose Blanchet, Mert Pilanci, and Robert Tibshirani. 2025. Llm-lasso: A robust framework for domain-informed feature selection and regularization. *arXiv preprint arXiv:2502.10648*.
- Susu Zhang, Jingchen Liu, and Zhiliang Ying. 2023. [Statistical applications to cognitive diagnostic testing](#). *Annual Review of Statistics and Its Application*, 10:651–675.

A Fraction-Subtraction Items and Attributes

Table 4 displays the item texts and answer keys supplied to the LLM. Table 5 shows the four attributes' definitions and examples.

Item	Expression	Accepted answer
1	$5/3 - 3/4$	$11/12$
2	$3/4 - 3/8$	$3/8$
3	$5/6 - 1/9$	$13/18$
4	$3\frac{1}{2} - 2\frac{3}{2}$	0
5	$4\frac{3}{5} - 3\frac{4}{10}$	$6/5$ or $1\frac{1}{5}$
6	$6/7 - 4/7$	$2/7$
7	$3 - 2\frac{1}{5}$	$4/5$
8	$2/3 - 2/3$	0
9	$3\frac{7}{8} - 2$	$15/8$ or $1\frac{7}{8}$
10	$4\frac{4}{12} - 2\frac{7}{12}$	$7/4$ or $1\frac{3}{4}$
11	$4\frac{1}{3} - 2\frac{4}{3}$	1
12	$1\frac{1}{8} - 1/8$	1
13	$3\frac{3}{8} - 2\frac{5}{6}$	$13/24$
14	$3\frac{4}{5} - 3\frac{2}{5}$	$2/5$
15	$2 - 1/3$	$5/3$ or $1\frac{2}{3}$
16	$4\frac{5}{7} - 1\frac{4}{7}$	$22/7$ or $3\frac{1}{7}$
17	$7\frac{3}{5} - 4/5$	$34/5$ or $6\frac{4}{5}$
18	$4\frac{1}{10} - 2\frac{8}{10}$	$13/10$ or $1\frac{3}{10}$
19	$4 - 1\frac{4}{3}$	$5/3$ or $1\frac{2}{3}$
20	$4\frac{1}{3} - 1\frac{5}{3}$	$5/3$ or $1\frac{2}{3}$

Table 4: Fraction-subtraction item information supplied to the LLM.

Attribute	Definition and example
A1: Find a common denominator	Construct a common denominator and rewrite both terms as equivalent fractions with that denominator. Example: $5/6$ and $1/9$ become $15/18$ and $2/18$.
A2: Borrow from an integer	When the fractional part of the minuend is smaller, borrow 1 from its integer part and add d/d to its fractional part. Example: $4\frac{1}{10}$ becomes $3\frac{11}{10}$.
A3: Subtract integer and fractional parts	Subtract the integer and fractional parts separately after placing them in compatible form, then combine the result. Example: $3\frac{16}{12} - 2\frac{7}{12} = 1\frac{9}{12}$.
A4: Convert a whole number to a fraction	Rewrite a whole-number operand as a fraction with the working denominator. Example: rewrite 2 as $6/3$ when computing $2 - 1/3$.

Table 5: Fraction-subtraction attribute definitions supplied to the LLM.

B Fraction-Subtraction Simulation Item-Level Results

	Initial Q				Final Q				Row-wise Q Recovery	RSS-based flag rate	Revision rate given RSS flag
	A1	A2	A3	A4	A1	A2	A3	A4			
Item1: 5/3 - 3/4	1.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000	0.120	0.000
Item2: 3/4 - 3/8	1.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000	0.080	0.000
Item3: 5/6 - 1/9	1.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000	0.120	0.000
Item4: 3 1/2 - 2 3/2	0.000	0.920	0.220	0.020	0.000	1.000	0.100	0.000	0.900	0.200	0.824
Item5: 4 3/5 - 3 4/10	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.020	0.000
Item6: 6/7 - 4/7	1.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	1.000	1.000	0.758
Item7: 3 - 2 1/5	0.000	0.140	0.280	0.900	0.000	0.100	0.120	0.940	0.880	0.400	0.554
Item8: 2/3 - 2/3	1.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	1.000	1.000	0.769
Item9: 3 7/8 - 2	0.000	0.000	1.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	-
Item10: 4 4/12 - 2 7/12	0.000	1.000	1.000	0.000	0.000	1.000	0.340	0.000	0.660	0.840	0.230
Item11: 4 1/3 - 2 4/3	0.000	0.880	0.600	0.020	0.000	0.980	0.760	0.020	0.740	0.260	0.944
Item12: 1 1/8 - 1/8	0.000	0.000	1.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	-
Item13: 3 3/8 - 2 5/6	1.000	1.000	0.980	0.000	1.000	1.000	1.000	0.840	0.840	1.000	0.248
Item14: 3 4/5 - 3 2/5	0.000	0.000	1.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	-
Item15: 2 - 1/3	0.000	0.000	0.000	1.000	0.000	0.020	0.020	0.980	0.980	0.120	0.025
Item16: 4 5/7 - 1 4/7	0.000	0.000	1.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	-
Item17: 7 3/5 - 4/5	0.000	1.000	0.460	0.000	0.000	1.000	0.720	0.000	0.720	0.260	0.933
Item18: 4 1/10 - 2 8/10	0.000	1.000	1.000	0.000	0.000	1.000	1.000	0.840	0.840	1.000	0.196
Item19: 4 - 1 4/3	0.000	0.340	0.480	0.940	0.000	0.980	0.960	0.940	0.900	0.700	0.722
Item20: 4 1/3 - 1 5/3	0.000	0.840	0.980	0.120	0.000	0.900	1.000	0.100	0.900	0.080	1.000

Table 6: Fraction simulation: Average initial Q across replications, average final Q across replications and final Q's row-wise recovery rate, proportion of replications an item is flagged at least once with RSS (i.e., RSS-based flag rate), and proportion of times Q vector is revised following RSS flag (i.e., Revision rate given RSS flag). Note: **Bolded** entries indicate an absolute deviation from ground-truth Q above .20, which could arise with either fluctuation across replications or consistent difference from ground-truth.

C Statistics Classroom Assessment Items and Attributes

Table 7 summarizes the operational items without reproducing the exact content of the item. Table 8 presents the attribute labels and definitions supplied to the LLM for Version 1 and Version 2. A1, A2, and A5–A8 were common to both versions, and A3 and A4 were version-specific.

Item	Description	Reference	$p(\text{correct})$
1	Definition of data	-	.830
2	Number of attributes in a data set about hotels	Table	.943
3	Number of observations in a data set about hotels	Table	.950
4	Graph choice for hotel floor counts	Table	.755
5	Definition of relative frequency	-	.810
6	Computation of relative frequency	Table	.751
7	Quartile interpretation from a summary table	Table	.649
8	Identification of skewness from a summary table	Table	.739
9	Interpretation of a percentile from a summary table	Table	.757
10	Interpretation of the middle 50% from a summary table	Table	.676
11	Interpretation of the middle 95% from a summary table	Table	.495
12	Graph choice for in-state percentage by college	Table	.596
13	Denominator for a conditional frequency from a two-way table	Table	.506
14	Interpretation of Cramer's V from a two-way table	Table	.686
15	Definition of a smoothed trend line	Plot	.441
16	Interpretation of association strength in a scatterplot	Plot	.778
17	Interpretation of association direction in a scatterplot	Plot	.217
18	Identification of the variable types represented in a scatterplot	-	.751
19	Graph choice for comparing a quantitative distribution between groups	Plot	.898
20	Comparison of spread between two quantitative distributions	Plot	.960
21	Interpretation of an effect size	Plot	.413
22	Comparison of spread between two quantitative distributions	Plot	.890
23	Interpretation of distribution shape	Plot	.866
24	Interpretation of the center of a distribution	Plot	.537
25	Interpretation of the middle 95% of a distribution	Plot	.698

Table 7: Item summary for the statistics classroom assessment. *Note:* Reference: additional table or plot applicable to item. $p(\text{correct})$: Observed proportion of correct response.

Version	Attribute	Label	Definition
Both	A1	Data Structure	Systematic arrangement of a data set in terms of data points (observations) and their characteristics (attributes or variables).
Both	A2	Variable Types	Types of variables and measures, including quantitative and categorical variables and common subtypes.
1	A3	Univariate Quantitative Summaries and Graphs	Univariate summary statistics and common visual displays for a single quantitative variable, including the mean, median, standard deviation, five-number summary, density plot, and box plot.
1	A4	Univariate Categorical Frequencies and Bar Graphs	Frequency tables and visual displays for a single categorical variable, including frequencies, relative frequencies, percents, and bar graphs.
2	A3	Univariate Summaries	Univariate numerical summaries for quantitative and categorical variables, including the mean, median, standard deviation, five-number summary, frequencies, relative frequencies, and percents.
2	A4	Univariate Visualizations	Univariate visual displays for quantitative and categorical variables, including density plots, box plots, and bar graphs.
Both	A5	Quantitative Distribution Characteristics	Characteristics of quantitative distributions, including shape, location, spread, skewness, central intervals, and the normal rule of thumb.
Both	A6	Two-Way Tables	Two-way tables for two categorical variables, including conditional percents, comparisons of conditional percents, and association summaries such as Cramer's V.
Both	A7	Comparing Quantitative Distributions	Comparisons of quantitative distributions across groups, including differences in location and spread and standardized comparisons such as effect size and ratios of standard deviations.
Both	A8	Scatterplots and Correlations	Joint distributions and associations for two quantitative variables, including scatterplots, correlation-type measures, and trend lines.

Table 8: Statistics classroom assessment attribute labels and definitions supplied to the LLM.

D Rationale Rating Rubric

Criterion	Definition	Example of a violation
Attribute alignment	The rationale connects each explanation to the correct named attribute and does not confuse one attribute with another.	The rationale assigns a solution step to an attribute whose definition describes a different skill.
Item grounding	The rationale remains tied to the item content and solution steps and does not introduce unsupported details.	For an item that only requires comparing fractions, the rationale introduces borrowing or mixed-number conversion even though neither operation appears in the item.
Substantive accuracy	The rationale correctly explains whether each attribute is or is not required to solve the item.	The rationale identifies an attribute as required even though the item can be solved without it, or rejects an attribute that is necessary.
Review utility	The rationale is clear and specific enough for a reviewer to understand the initial Q-vector or the decision to revise or retain it.	The rationale states only that a skill is relevant without identifying what feature or solution step makes it necessary or unnecessary.

Table 9: Instructions used to evaluate the LLM-generated rationales.

E Version 1 LLM Rationales Evaluated by Experts

Table 10 contains the LLM’s rationale for the initial Q-vector for all 25 items. Table 11 contains the last available refinement rationale for each of the 10 items flagged on the trajectory leading to the selected version 1 Q-matrix.

Item	Initial rationale
1	A1 required to recognize that data are representations of people or things.
2	A1 required to count the five attributes shown as the table’s columns.
3	A1 required to count the six observations shown as the table’s rows.
4	A2 and A3 required to recognize n_floors is quantitative and select a density plot.
5	A4 required to identify percentages of category counts as relative frequencies.
6	A1 and A4 required to use the sample total 265 and compute 80/265 as about 30%.
7	A3 and A5 required to identify Q1 = 23 and use that 75% of values are at least Q1.
8	A5 required to compare mean and median and reject a left-skew claim when they are equal.
9	A5 required to place 10 well below Q1 = 23, so it is near the bottom few percent.
10	A3 required to read the middle 50% as the interval from Q1 to Q3.
11	A3 and A5 required to use mean 30 and SD 5 and apply the about 95% rule to get 20 to 40.
12	A4 and A6 required to recognize a conditional percent for categorical data is best shown with a bar chart.
13	A6 required to identify 64% as a column percent using the in-state total as the denominator.
14	A6 required to interpret Cramer’s V = .89 as a strong relationship.
15	A8 required to identify the smoothed trend line as a line of averages.
16	A8 required to interpret Kendall’s Tau of -0.62 and the plot as a moderate relationship.
17	A8 required to evaluate claims against the scatterplot and trend, leaving none of the statements supported.
18	A2 required to know scatterplots display two quantitative attributes.
19	A3 and A7 required to recognize grouped density curves as a side-by-side density plot.
20	A7 required to compare the two group distributions and identify the one with larger spread.
21	A7 required to judge the standardized difference between groups as about 1.0.
22	A7 required to know the ratio of standard deviations compares spread across groups.
23	A5 required to identify the distribution’s shape as approximately normal.
24	A3 and A5 required to use the graph’s center and near symmetry to estimate the mean at about 105 mg.
25	A3 and A5 required to read an approximately normal distribution and use the central 95% range of about 50 to 200.

Table 10: Initial version 1 rationales evaluated by the two experts.

Item	Refinement rationale
4	changed; A3 only since the task is to choose an appropriate graph for one quantitative variable.
6	changed; compute 80/265 as a percent from the frequency table, so A4 alone is needed.
7	changed; use Q1=23 from the five-number summary to judge the 75% statement, so A3 alone is needed.
12	changed; A4 only since the goal is to depict a category percent with a bar chart, and the suggested extra step is not needed.
13	unchanged; identify 64% as a column percent in a two-way table, so A6 only; RSS-suggested A5 is not used.
15	changed; A8 identifies the smoothed trend line, and A5 supports reading it as a line summarizing typical sleep values across screen time.
17	unchanged; check each claim directly against the scatterplot and trend line, so A8 only; the all-skill RSS alternative is not supported.
19	changed; the task is only to recognize the display as a density plot for a quantitative variable, so the graph-identification skill alone is needed.
20	unchanged; compare the spreads of the two quantitative group distributions, so A7 only; RSS-suggested A1 is not used.
25	changed; A5 only since the step is to identify the middle 95% range from the quantitative distribution.

Table 11: Version 1 refinement rationales evaluated by the two experts.

F Initial Q-Matrix Prompt

The following template was used for the initial Q-matrix proposal prompt. Application-specific item, attribute, and constraint blocks were inserted at the indicated locations. References (i.e., directories) to attached figures were added for items with visual content. Operational statistics items are not reproduced.

You are a subject-matter expert in [TEST DOMAIN] for [TARGET POPULATION].

TASK: Produce an initial Q-matrix for a cognitive diagnostic model.

WORKFLOW (internal; do not print):

- 1) SOLVE: Determine a substantively defensible correct solution path for each item.
- 2) CODE: Set $q_{jk}=1$ if attribute k is required for successful solution of the item; else 0.
 - Do not weight by counts of operations.
 - Ignore non-attributes such as reading, formatting, and test-taking.
 - For polytomous items, code skills required for full credit.
 - Be parsimonious; avoid nice-to-have strategies.
- 3) SELF-CHECK: If the solution path and proposed Q disagree, fix Q internally.
- 4) Q-RATIONALE CONSISTENCY CHECK: Ensure each rationale names only attributes coded as 1. If not, revise the Q-vector before finalizing.

ITEMS (id: text with correct answer):
[ITEM BLOCK]

ATTRIBUTES (id: definition and examples):
[ATTRIBUTE BLOCK]

DOMAIN CONSTRAINTS (enforce):
[DOMAIN-CONSTRAINT BLOCK]

- Avoid all-zero or all-one columns unless strictly necessary.

[ATTACHED-FIGURE BLOCK, IF APPLICABLE]

OUTPUT (STRICT):
Return exactly TWO fenced R code blocks and nothing else.

FIRST R CODE BLOCK:

- Define ``qmatrix`` with exactly J rows and K columns using `matrix(..., byrow=TRUE)`.
- Entries must be 0/1 only.
- Do not set row or column names.

SECOND R CODE BLOCK:

- Define ``item_rationale`` as a named list with exactly J entries named `Item1`, ..., `ItemJ`.
- Each entry must be one short character string.
- Give a step-based reason for the coded skills.
- Mention only skills coded as 1 and the key step that makes them necessary.
- Verify that each rationale matches its Q-vector.
- Do not explain every omitted skill.
- Keep each rationale short, concrete, and procedural.

G Q-Matrix Refinement Prompt

The following template was used after the initial fit. Only the subset of items flagged by RSS was inserted.

You are a subject-matter expert in [TEST DOMAIN] for [TARGET POPULATION].

TASK:
Revise the Q-matrix for THIS FLAGGED-ITEM SUBSET ONLY.

GOAL:
Use the statistical review only as diagnostic evidence. It may suggest possible misspecification, but it is not the definitive standard. Final decisions must be based primarily on item content and attribute definitions.

EVIDENCE PRIORITY:

- 1) Item content and a substantively defensible correct solution path
- 2) Attribute definitions
- 3) Statistical review of flagged items
- 4) Global structural/identifiability constraints

CURRENT Q FOR FLAGGED ITEMS:
[CURRENT-Q BLOCK]

STATISTICAL REVIEW OF FLAGGED ITEMS (advisory, not definitive):
[RSS-CRITIQUE BLOCK]

WORKFLOW (internal; do not print):

- 1) SOLVE: Determine a substantively defensible correct solution path for each flagged item.
- 2) CODE: Set $q_{jk}=1$ if attribute k is required for successful solution of the item; else 0.
 - Do not weight by counts of operations.
 - Ignore non-attributes such as reading, formatting, and test-taking.
 - For polytomous items, code skills required for full credit.
 - Be parsimonious; avoid nice-to-have strategies.
- 3) COMPARE the content-based Q-vector with the current vector and the statistically suggested competing vector.
- 4) DECIDE:
 - Treat the statistical review as diagnostic evidence only.
 - Adopt a suggested change only if supported by the item and attribute definitions.
 - If only part is supported, revise only that part.
 - If the current row is substantively correct, keep it.
 - Do not copy a suggested vector merely because it has better fit.
 - Change at least one flagged item. Choose the suggested alternative most consistent with content; if none is content-consistent, randomly apply one suggested alternative.
- 5) SELF-CHECK: If the solution path and proposed Q disagree, fix Q internally.
- 6) Q-RATIONALE CONSISTENCY CHECK: Ensure each rationale names only attributes coded as 1.

REVISION CONSTRAINTS:

- Revise only the flagged items shown here.
- Return a Q-vector for every flagged item in the same order as the current subset.
- Do not add or remove rows.
- Entries must be 0/1 only.
- Ensure at least one flagged item is changed.

ITEMS:
[FLAGGED-ITEM BLOCK]

ATTRIBUTES:
[ATTRIBUTE BLOCK]

DOMAIN CONSTRAINTS:
[DOMAIN-CONSTRAINT BLOCK]

[ATTACHED-FIGURE BLOCK, IF APPLICABLE]

OUTPUT (STRICT):
Return exactly TWO fenced R code blocks and nothing else.

FIRST R CODE BLOCK:

- Define ``qmatrix`` for the flagged subset using `matrix(..., byrow=TRUE)`.
- Entries must be 0/1 only.
- Do not set row or column names.

SECOND R CODE BLOCK:

- Define ``item_rationale`` as a named list with one entry per flagged item.
- Begin each rationale with ``changed;`` or ``unchanged;``.
- Give a brief, step-based reason tied to the solution path.
- If changed, state the step supporting the revised skills.
- If unchanged, state why the RSS suggestion was not supported by the solution path.
- Verify that every rationale matches the corresponding final Q-vector.
- Keep each rationale short, concrete, and procedural.