Leveraging LLMs for Cognitive Skill Mapping in TIMSS Mathematics Assessment

Ruchi J. Sachdeva, Ph.D., Pearson
Jung Yeon Park, Ph.D, George Mason University

Abstract

This study evaluates GPT-4 for generating and validating Q-matrices for TIMSS mathematics items. Using expert mappings as benchmarks, we examined prompt design, temporal stability, and error detection. The model showed strong accuracy, substantial reliability, and effective recovery of altered skills, supporting its role as a scalable aid to cognitive diagnosis.

1 Background

Identifying the cognitive skills required to solve specific test items is a foundational task in educational assessment. This function is not only critical for ensuring the validity of test interpretations but is especially central to the development and validation of Q-matrices in Cognitive Diagnosis Models (CDMs; Rupp et al., 2010).

A carefully constructed Q-matrix aligns each assessment item with the exact constellation of cognitive skills or knowledge components required for its solution. When the identification of these skills is flawed or incomplete, the resulting model can misrepresent learners' proficiencies, undermining one of the principal strengths of cognitive diagnostic models, their capacity to deliver precise, actionable feedback. Consequently, the delineation of content and associated skills emerges as a critical, yet cognitively demanding, dimension of assessment design. Against this backdrop, large language models (LLMs), such as GPT, present a promising avenue for augmenting or supporting expert analysis, offering new opportunities to enhance the rigor and efficiency of Q-matrix development.

2 Purpose

This study evaluates GPT-4's capacity to identify skills and validate Q-matrices against a content expert–designed gold standard across Number, Algebra, Geometry, and Data and Chance. Q-matrix design requires more than simple skill matching, demanding analysis of interactions, hierarchies, cognitive load, and item context. We examine whether GPT-4 can meet these demands as a scalable, cost-effective aid to expert assessment design. This study addresses the following research questions:

RQ1: What prompt strategies enable GPT-4 to accurately map cognitive skills to test items?

RQ2: How stable are its Q-matrices across repeated prompts?

RQ3: How does GPT-4o's performance vary across different error types (skill addition and skill swapping)?

3 Sample

This study uses the TIMSS 2007 Grade 7 Mathematics Released Items with an expert-defined Q-matrix for 89 publicly available items (Johnson et al., 2013) approved by NCES to strengthen the generalizability and relevance of its findings in large-scale educational research.

4 Methodology

A three-part framework was employed to refine methods for generating and validating Q-matrix skill mappings for TIMSS mathematics items. First, two prompt templates were designed: one to construct Q-matrix entries for all 89 items and another to evaluate existing matrices for errors, each targeting a distinct cognitive mapping task.

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 223–228

Second, we assessed agreement on the number of skills per item by comparing GPT estimates with counts from the content expert gold standard. We also tested four prompt configurations to gauge their effect on Q-matrix accuracy, contrasting a full version containing all optimization elements with simplified versions that excluded skill descriptions, expert-role instruction, or procedural constraints.

Third, we evaluated the reliability of a single fixed prompt by generating 20 independent item-by-skill mappings and measuring consistency across runs. This framework offers a rigorous basis for understanding how prompt design and refinement influence the precision and diagnostic quality of automated Q-matrix construction in educational assessment research.

4.1 Prompt Design

We refined prompt structures with a subset of 10 TIMSS items, then applied the finalized versions to all 89 items to generate Q-matrix entries. Because submitting all items at once exceeded the token limit, each was provided individually with its prompt and image.

Two prompts were developed: one to generate an independent Q-matrix and another for validation. Both returned JSON-formatted outputs containing the item identifier, the corresponding Q-matrix entry, the number of skills identified, and a brief rationale for each skill selection.

For validation, a Q-matrix entry with an intentional error was presented, and GPT was asked to judge its accuracy and produce a corrected mapping for the item. The validation output also included a one-sentence explanation of the chosen skills, and a one-sentence justification for agreeing or disagreeing with the provided Q entry for the item.

4.2 Testing Prompt Strategies

To evaluate the impact of various prompt optimization strategies, we compared GPT-generated Q-matrices with and without these enhancements, holding all other conditions constant to isolate the effects of prompt structure. Each prompt variant produced a distinct Q-matrix, which was benchmarked against a content expert-Q matrix.

Four prompt configurations were evaluated for their effect on Q-matrix accuracy. (See Table 1) The full prompt included all optimization elements for constructing an 89-item, nine-skill matrix. Other versions removed detailed skill descriptions, omitted the instruction assigning an expert role, or removed all the procedural constraints, which provided guidance on selecting a primary skill, adding additional skills, and interpreting graphs and shapes, leaving only the core task specification.

Method	**Status**	**Description**
Full Prompt (P1)	Kept All	Used all prompt techniques to produce the final version for an 89-item, 9-skill Q-matrix.
Skill Details (P2)	Removed	Detailed descriptions of each skill (Taken from the TIMSS technical documentation).
Providing a Role (P3)	Removed	Omitted the instruction: *"You are a content expert in 7th-grade Mathematics assessment in the United States."*
Procedural Constraints (P4)	Removed	Eliminated procedural rules about task limits, content-domain mapping, leaving only the high-level task description.
Evaluation Metrics	**F1-score comparisons against expert-designed Q-matrix at the skill level.**	

Table 1: Framework for Prompt Engineering Strategies in Q-Matrix Generation.

This evaluation framework allowed systematic identification of prompt configurations and their impact on accuracy and informing best practices. Performance was assessed using F1 scores, with precision and recall calculated at the overall and skill level against the content expert-designed Q-matrix.

4.3 Number of Required Skills

A Q-matrix maps assessment items to the cognitive skills needed for their solution. After defining skills from learning objectives, each item is reviewed to determine how many are required. Using the same expert-defined skills, GPT estimated this count, and we compared its results with expert judgments using intraclass

correlation (ICC) to assess agreement on item complexity. Treating each item–skill decision (i.e., whether an item is linked to a given skill; coded 0/1)

4.4 Repeated Prompt Reliability

LLMs can map assessment items to cognitive skills, yet their outputs may fluctuate across identical runs, raising concerns about the reliability of Q-matrices in cognitive diagnostic models. To evaluate this, we tested temporal stability by administering the same prompt to GPT twenty times and calculating Krippendorff's alpha (Krippendorff, 2018). This statistic measures the degree of agreement among multiple coders or iterations beyond chance and accommodates various data types, including nominal scales. It was appropriate here because the entire binary skill vector (for example, 100100000) was treated as a single nominal category. This approach provided an estimate of consistency across repeated mappings and the robustness of GPT-generated Q-matrices.

4.5 Q-Matrix Error Detection Using Add and Swap Conditions

To assess GPT's ability to detect and correct errors in skill mappings, we used two procedures called Add and Swap (Table 2). Both began with the expert Q-matrix for each TIMSS item; 74 of 89 items involved a single skill, offering a clear test case. GPT received the skill definitions, learning objectives, detailed prompt, and a PNG of the item.

In the Add condition, a randomly selected extra skill was appended to the correct mapping to create an altered entry. GPT evaluated agreement or disagreement with the provided Q-matrix entry, identified the appropriate skill pattern for the item, and offered a rationale when its judgment differed. In the Swap condition, one correct skill was replaced with an unrelated alternative, and GPT assessed the substitution, stated agreement or disagreement, and proposed the correct skill or set of skills.

GPT's recommendations were compared with the content-expert mappings, and accuracy was assessed with F1 scores. Illustrative examples and procedural details for both conditions are presented in Table 2.

Condition	Purpose	Procedure	Illustrative Example
Add	Evaluate GPT's ability to identify and discard unnecessary skills in Q-matrix mappings.	GPT reviewed an augmented skill set containing one unnecessary element and determined whether to retain or remove it before producing the final mapping.	Example: For a geometry item, the expert selected only Skill 6. The Q-matrix listed Skills 6 and 1; GPT removed Skill 1, retaining Skill 6 in line with the expert mapping.
Swap	Examine GPT's capacity to identify and correct an entirely different (incorrect) skill.	For each (mostly single-skill) item, the correct skill was swapped with an unrelated one; GPT reviewed the materials and proposed the appropriate skill(s).	Example: In a number item, the correct Skill 1 was swapped for Skill 5; GPT removed Skill 5 and reinstated Skill 1.

Table 2: Procedures for Error Detection (Add and Swap)

5 Results

This section presents findings on the accuracy, reliability, and interpretability of Q-matrices generated by GPT. We report performance across prompt strategies, agreement on the number of skills per item, detection of Q-Matrix errors, and stability across repeated runs, highlighting how prompt design influenced outputs.

5.1 Prompt Optimization

Each prompt produced a distinct Q-matrix, which was evaluated against a content expert–designed

Q-matrix using F1, precision, and recall. Table 1 summarizes the prompt-engineering strategies, and Table 3 reports their effects on accuracy and partial credit. The Full Prompt (P1) yielded the highest performance, with an overall F1 of 0.83 and a partial-credit mean of 0.82. P2 (No Skill Details) achieved an F1 of 0.78 and a partial-credit mean of 0.76, while P3 (No Role Assignment) showed an F1 of 0.78 and a partial-credit mean of 0.81. P4 (No Constraints) produced the largest decline, with an F1 of 0.77 and a partial-credit mean of 0.80. F1 measured exact agreement, whereas partial-credit scores reflected overlap; the two metrics were similar because most items targeted a single skill. The relatively high partial-credit scores for P3 and P4 suggest that, although omitting these prompting techniques preserved the identification of the primary skill, it reduced the detection of secondary or supporting skills when key prompt elements were excluded.

Skill	P1	P2	P3	P4
Skill 1: Whole Numbers & Integers	0.85	0.71	0.67	0.86
Skill 2: Fractions, Decimals & Percents	0.91	0.91	0.92	0.95
Skill 3: Ratios & Proportions	0.81	0.80	0.80	0.67
Skill 4: Patterns	0.67	1.00	0.86	0.50
Skill 5: Expressions, Equations & Functions	0.96	1.00	0.96	0.88
Skill 6: Lines, Angles and Shapes	0.77	0.74	0.78	0.88
Skill 7: Measurement	0.88	0.50	0.67	0.80
Skill 8: Location & Movement	0.68	0.55	0.55	0.60
Skill 9: Data and Chance	0.93	0.80	0.81	0.81
Full Sample F1	**0.83**	**0.78**	**0.78**	**0.77**
Mean Partial-Credit Score	**0.82**	**0.76**	**0.81**	**0.80**

Table 3: F1 Score & Partial Credit Score by Skill.

Across individual skills, the strongest accuracy was observed for Expressions, Equations, and Functions (F1 = 0.88–1.00) and Fractions, Decimals, and Percents (F1 > 0.90 across all prompts). Data and Chance also showed consistently high performance (F1 = 0.80–0.93). By contrast, Location and Movement had the lowest scores (0.55–0.68), and Patterns displayed substantial variation across prompts (0.50–1.00). Measurement showed moderate sensitivity to prompt design, ranging from 0.50 to 0.88. An F1 value of 1 indicates perfect agreement between the prompt-generated Q-matrix and the content expert–designed Q-matrix. These findings suggest that skills such as Expressions, Equations, and Functions and Fractions, Decimals, and Percents are relatively stable across prompts, whereas skills like Location and Movement and Patterns, which often involve graphs and diagrams, are more susceptible to changes in prompt constraints.

5.2 Number of Skills Agreement

A Q-matrix links assessment items to the cognitive skills required for their solution. After defining skills from the learning objectives, each item was reviewed to determine the number of unique skills involved—a task that is both challenging and essential for accurate measurement. GPT analyzed the items and estimated the total number of skills required, and this estimate was compared with the corresponding counts from the content expert–designed Q-matrix. Agreement between GPT and the content expert–designed Q-matrix was evaluated using intraclass correlation (ICC) to assess item complexity. Treating each item–skill decision (0 = not linked, 1 = linked) as a subject and the two raters (content expert–designed Q-matrix and GPT; N = 2,047) as judges, single-rater ICCs (ICC1/2/3) were 0.72 (95% CI [0.70, 0.74]). When ratings were averaged, ICCs (ICC1k/2k/3k) increased to 0.84 (95% CI [0.82, 0.85]). Since many of the items involved only a single skill, future work should examine Q-matrices containing a larger proportion of items that require multiple skills.

5.3 Temporal Stability of Q-Matrix Skill Mappings

Large language models (LLMs) can map assessment items to cognitive skills, but their outputs may vary across identical runs, making reliability a critical concern for Q-matrices in cognitive diagnostic models (CDMs). To examine temporal stability, we administered the same prompt to GPT 20 times and computed

Krippendorff's alpha (Krippendorff, 2018). GPT demonstrated strong consistency, with mean alphas of 0.86 (exact match), 0.94 (Hamming distance), and 0.93 (Jaccard similarity). Sixty-five of 89 items (73%) were identical across runs, most remaining items exceeded $\alpha = 0.80$, and only a few showed lower agreement (exact-match α as low as 0.11). Items with the poorest reliability were multipart questions in which extensive information was presented on a single page, suggesting that reading complex PNGs with many components may impair repeatability. Item-level results can be provided upon request.

5.4 Detecting Q-Matrix Error (Add and Swap Conditions)

GPT showed strong performance in detecting and correcting altered skill assignments in both the Add and Swap conditions (Table 4). Accuracy was highest for Expressions, Equations and Functions and Data and Chance. Moderate scores, with F1 values ranging from 0.75 to 0.90, were observed for Whole Numbers and Integers (Skill 1), Fractions, Decimals and Percents (Skill 2), Ratios and Proportions (Skill 3), Lines, Angles and Shapes (Skill 6), and Measurement (Skill 7).

Skill	Description	Add	Swap
1	Whole Numbers & Integers	0.85	0.86
2	Fractions, Decimals, & Percents	0.90	0.88
3	Ratios & Proportions	0.81	0.82
4	Patterns	0.61	0.67
5	Expressions, Equations, & Functions	0.96	0.97
6	Lines, Angles, & Shapes	0.77	0.78
7	Measurement	0.75	0.82
8	Location & Movement	0.55	0.56
9	Data and Chance	0.87	0.88
	Overall F1	**0.78**	**0.80**

Table 4: Performance in Detecting and Correcting Altered Skill Assignments.

Lower accuracy emerged for Patterns (Skill 4) and Location and Movement (Skill 8), where F1 scores were consistently below 0.70 across both conditions. For Lines, Angles and Shapes, performance was also slightly reduced, which may reflect current challenges in interpreting graphs, diagrams, and geometric figures. These areas may benefit from enhanced visual-processing capabilities or additional expert review to ensure reliable skill detection. Overall, the findings indicate that GPT can accurately identify and correct altered skill assignments, particularly in numerical and algebraic contexts, while tasks involving geometry and spatial reasoning may require refined prompts or closer collaboration with human experts.

6 Conclusion

This study demonstrates that GPT-4 can meaningfully support the cognitively demanding task of Q-matrix construction and validation. When provided with explicit skill definitions, structured prompts, and item images, GPT achieved high agreement with content expert mappings (F1 = 0.83) and substantial reliability across repeated runs ($\alpha \approx 0.86$). It also detected and corrected injected errors in the add and swap conditions with moderate to strong accuracy, particularly in number and algebra content domains. Performance declined for geometry and spatial-reasoning items, suggesting that visual interpretation remains a limiting factor. Items with heavy reading loads, multi-part content presented on a single page, or complex graphical information in PNG format also showed weaker repeatability, indicating that such features may challenge the model's consistency. Overall, these findings suggest that large language models, when carefully prompted, can offer scalable and replicable assistance in skill identification, complementing rather than replacing expert judgment. Future research should focus on refining methods for items with extensive text, multi-part layouts, or intricate visual elements to improve performance in these areas.

References

Chiu, C. Y., & Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification*, 30(2), 225-250.

Johnson, M.S., Lee, Y.-S., Sachdeva, R., Zhang, Z., & Park, J.Y. (2013, April). Examination of

Gender Differences using the Multiple Groups DINA Model [Paper presentation]. *National Council on Measurement in Education*, San Francisco CA.

Krippendorff, K. (2018). *Content Analysis: An Introduction to Its Methodology* (4th ed.). Thousand Oaks, CA: Sage.

Martin, M. O., Mullis, I. V. S., & Foy, P. (2008). TIMSS 2007 *International Science Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades.* TIMSS & PIRLS International Study Center, Boston College.

Ravand, H., & Robitzsch, A. (2015). Cognitive diagnostic model of attribute profiles: G-DINA model framework. *Educational Psychology*, 35(8), 1088-1110.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. Guilford Press.

van Rijsbergen, C. J. (1979). *Information Retrieval* (2nd ed.). London: Butterworths.

Appendix A: Skills, Content Domains

Skills	**Content Domain**
1 – Whole Numbers & Integers;	Numbers
2 – Fractions, Decimals, & Percents	Numbers
3 – Ratios & Proportions	Numbers
4 – Patterns	Algebra
5 – Expressions, Equations, & Functions	Algebra
6 – Lines, Angles, & Shapes	Geometry
7 – Measurement	Geometry
8 – Location & Movement	Geometry
9 – Data Analysis & Probability	Data and Chance