A Multi-Agent Large Language Model Framework for Automated Q-Matrix Generation

AERA 2026 Annual Meeting

Jihong Zhang, Ph.D. Xinya Liang, Ph.D.

Department of Counseling, Leadership, and Research Methods
University of Arkansas

American Educational Research Association, 2026

Q-Matrix Construction

- Q-matrix specifies **item-attribute relationships** in IRT and DCMs (Tatsuoka, 1983)
- Traditional construction (Buck & Tatsuoka, 1998): define attributes → code Q-matrix → estimate parameters → refine iteratively
- Misspecification → poor model fit and incorrect classification

Fraction Subtraction (Tatsuoka, 1990)
A1: Basic, A2: Simplify, A3: Separate whole, A4: Borrow

Item	Example	A1	A2	A3	A4
1	$\frac{5}{3} - \frac{1}{3}$	1	0	0	0
4	$3\frac{1}{3} - \frac{2}{3}$	1	0	1	0
7	$4\frac{1}{2} - \frac{3}{4}$	1	1	0	0
10	$5\frac{1}{3} - 2\frac{2}{3}$	1	0	1	1

Motivation

- Q-matrix construction and validation is **expert-intensive**, **time-consuming**, and **subjective**
- Requires multiple evidence sources: student think-aloud protocols, expert ratings, psychometric considerations (Li & Suen, 2013)
- Real human data can be **scarce** or **low-quality** in early test development
- Experts may have **limited availability** or **domain knowledge gaps**; prone to **subjectivity**
- Traditional procedures are susceptible to **human bias or error**

Motivation

AI-driven automation can improve **efficiency**, **consistency**, and **scalability**, while keeping humans in the loop for monitoring.

LLMs for Q-Matrix Construction

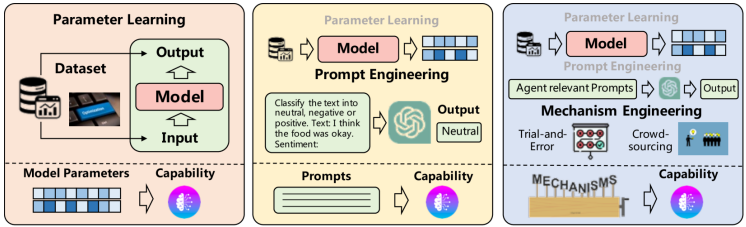


Figure 1: Three Eras of AI (Wang et al., 2023, arXiv:2308.11432)

- Related work uses **neural networks** (Tao et al., 2024) — **Gap**: no LLM + data-driven validation pipeline
- **Single-agent**: one LLM, one task at a time — generation and validation applied **separately** (Asiret et al., 2025; Xue & Appleton, 2026)
- **Multi-agent** (this study): domain expert + psychometrician + researcher agents in a **single integrated pipeline**

Research Questions

RQ1: Stability

How consistent are AI-generated Q-matrices across repeated runs of the multi-agent framework?

RQ2: Overlap

How well do multi-agent-produced Q-matrices match the reference Q-matrix?

RQ3: Finalization

How do different finalization strategies affect the overlap rate of the final Q-matrix?

Multi-Agent Framework

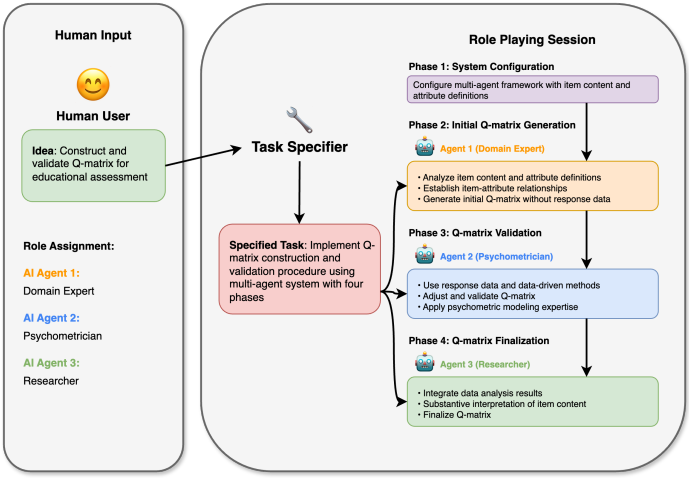


Figure 2: Q-matrix Construction Role-Playing Framework

Technical Stack and Prompt Design

- **LLM**: Llama-3.1 (70b-instruct)
- **Framework**: NVIDIA NeMo Agent Toolkit
- **Languages**: Python, R
- **Psychometric Package**: GDINA, Qval (R)
- **Validation**: PVAF

Prompt Design

1. **Agent 1** (Domain Expert — Generation): Assign J items to K attributes based solely on item content and definitions.
2. **Agent 2** (Psychometrician — Validation): Compare initial Q_0 against data-driven alternative. Tools: 'GDINA()' $\rightarrow$ fit DCM; 'Qval()' $\rightarrow$ PVAF validation.
3. **Agent 3** (Researcher — Finalization): Integrate Q_0, Q_1, item content, and parameter estimates. Rules: (1) screen agreement; (2) evaluate cross-loading; (3) apply structural constraints.

Q-matrix Finalization: Four Strategies

Plan	Phase 2	Phase 3	Phase 4	Binarization
A	1 Q_0	1 Q_1	1 Q_{final}	AI agent reviews Q_0 and Q_1
B	100 Q_0	100 Q_1	1 Q_{final}	AI agent reviews $\hat{P}_{Q_0}$ and $\hat{P}_{Q_1}$
C	100 Q_0	100 Q_1	1 Q_{final}	Researcher cutoff on $\hat{P}_{Q_1}$
D	100 Q_0	100 Q_1	100 Q_{final}	AI agent reviews $\hat{P}_{Q_{final}}$

- Plans B–D use 100 replications to quantify **stochastic variability**
- Plan A minimizes computational cost but does not capture LLM generation variability
- $\hat{P}_Q$: probability matrix from 100 replications — entry (i, k) = proportion of runs assigning item i to attribute k

Study 1: Social Anxiety Disorder Scale

- **SAD Scale**: 13-item survey measuring *social phobia* (Iza et al., 2014)
- 7-point Likert scale; 3 attributes: **A1** Public Performance, **A2** Close Scrutiny, **A3** Interaction
- **Simulated data** ($N = 787$) from `CDM` R package
- **Evaluation metric**: OR = proportion of matched q-entries, computed at overall, attribute, and item levels
- **Reference Q-matrix constructed via two methods**: (1) content-based expert review (Q_1^*); (2) data-driven PVAF validation (Q_2^*)

Q_1^* (content-based; simple structure)

Item	A1	A2	A3
1–4	1	0	0
5–8	0	1	0
9–13	0	0	1

Q_2^* (data-validated; overlapping structure)

Item	A1	A2	A3
1	1	0	0
2–3	1	1	0
4	1	1	1
5–6	0	1	0
7–8	0	1	1
9–11	0	0	1
12–13	0	1	1

Study 1: Stability of Initial Q-Matrices

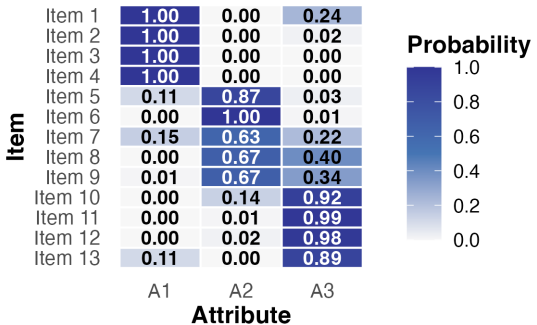


Figure 3: Initial Q-matrix Variability

Stability Findings

1. Items 1–4 (A1), 5–6 (A2), and 10–13 (A3): high certainty ($p \approx 1.0$)
2. Items 7–9 (A2): greater uncertainty, mixed patterns

Study 1: Stability of Validated Q-Matrices

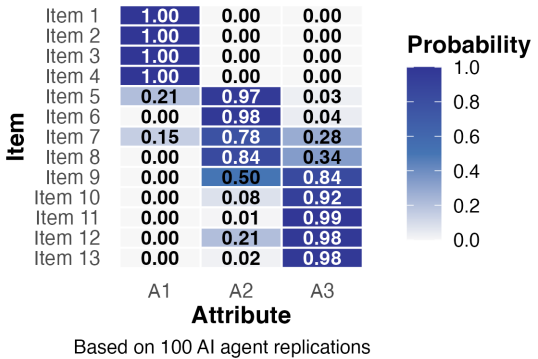


Figure 4: Validated Q-matrix Variability

Stability Findings

1. Items 1–4 (A1), Items 5–6 (A2), and 10–13 (A3): remained stable with high certainty
2. Items 7–8 (A2): stability improved with increased certainty
3. Items 9 (A2): increased uncertainty after validation

Study 1: Overlap Rates

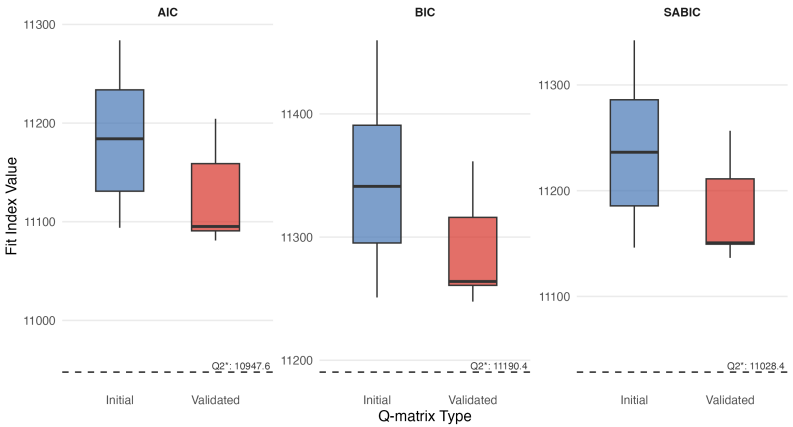
Initial Q-matrix vs. Q_1^*: Overall OR = **0.90** (95% CI [0.89, 0.92])

- A1: 0.97 | A2: 0.87 | A3: 0.86

Validated Q-matrix vs. Q_2^*: Overall OR = **0.77** (95% CI [0.76, 0.78])

- A1: 0.97 | A2: 0.55 | A3: 0.79
- A2 (close scrutiny) showed the lowest overlap — consistent with its lowest factor loading (0.62) in prior research

Study 1: Model Fit Comparison



Model: LCDM (Henson et al., 2009)

- Lower = better fit
- Dashed line = Q_2^* reference
- Validated Q-matrices consistently show **better model fit** and **lower variability** than initial Q-matrices
- Data-driven validation improves empirical fit beyond both the initial and reference Q-matrices

Study 1: Final Q-Matrix Results

Reference: Q_2^* (data-validated). Plans B, C (OR = 0.77) > Plan D (0.67) > Plan A (0.62).

	Overall	A1	A2	A3
Plan A	0.62	0.85	0.31	0.69
Plan B	0.77	1.00	0.54	0.77
Plan C	0.77	1.00	0.54	0.77
Plan D	0.67	0.92	0.31	0.77

- **A1** (public performance): consistently highest overlap across all plans
- **A2** (close scrutiny): consistently lowest — ambiguous construct boundaries and low factor loading
- **Plan A** (single run) performed worst — does not capture LLM variability

Plan A: single-run AI review; Plan B: 100-rep AI review of probability matrices; Plan C: 100-rep researcher cutoff; Plan D: 100-rep AI review of final probability matrix

Study 2: Fraction Subtraction Test

- **Fraction subtraction** dataset (Tatsuoka, 1990; de la Torre, 2008)
- 12 items, $N = 536$ middle school students
- 4 attributes: **A1** Basic fraction subtraction, **A2** Simplifying/reducing, **A3** Separating whole number, **A4** Borrowing from whole number
- **True Q-matrix** (Q_{true}) available for direct validation

Q_{true} (CDM data.fraction3)

Item	A1	A2	A3	A4
1	1	0	0	0
2	1	1	1	1
3	1	0	0	0
4	1	0	1	0
5	1	1	1	1
6	1	1	1	1
7	1	1	0	0
8	1	0	1	0
9	1	0	1	0
10	1	0	1	1
11	1	1	1	1
12	1	1	1	1

Study 2: Stability of Q-Matrices

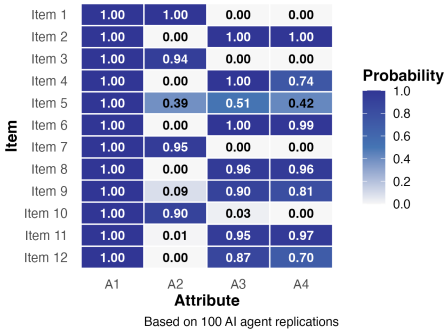


Figure 5: Initial Q-matrix Variability

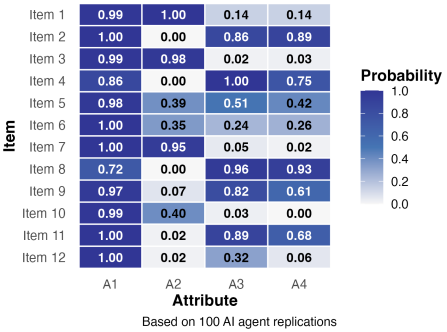


Figure 6: Validated Q-matrix Variability

- A1 assigned with **perfect certainty** ($p = 1.0$) across all items before and after validation
- Most items (items 6, 9, 10, 11, 12) increase uncertainty in the Validated Q-matrix

Study 2: Overlap Rates

Initial Q-matrix vs. Q_{true}: Overall OR = **0.71** (95% CI [0.71, 0.72])

- A1: **1.00** | A2: 0.37 | A3: 0.85 | A4: 0.63

Validated Q-matrix vs. Q_{true}: Overall OR = **0.65** (95% CI [0.64, 0.66])

- A1: 0.96 | A2: 0.44 | A3: 0.70 | A4: 0.49
- A2 improved slightly (0.37 $\rightarrow$ 0.44), but A3 and A4 **declined** after validation

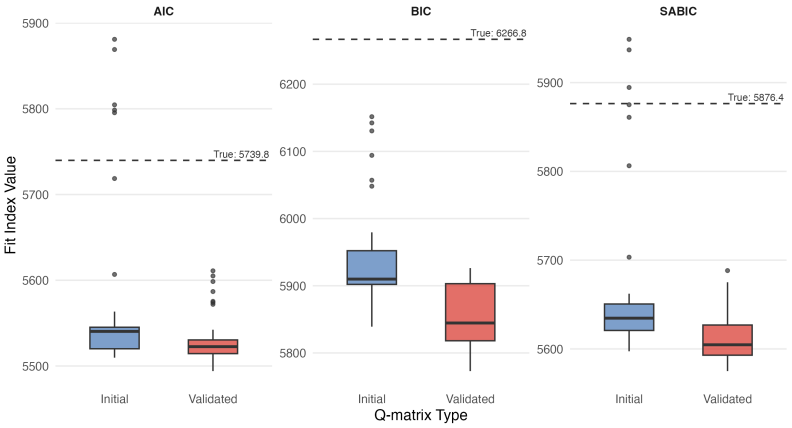
Study 2: Final Q-Matrix Results

Reference: Q_{true}. Plan A (OR = 0.77) outperformed Plans B–D.

	Overall	A1	A2	A3	A4
Plan A	**0.77**	1.00	0.50	0.92	0.67
Plan B	0.73	1.00	0.42	0.92	0.58
Plan C	0.65	1.00	0.42	0.75	0.42
Plan D	0.69	1.00	0.42	0.83	0.50

- **A1** (basic subtraction): perfect overlap (1.00) across all plans
- **A2** (simplifying): most challenging attribute to classify
- **Item 10** had lowest item-level OR (0.25) across all plans
- **Opposite pattern from Study 1** — optimal strategy depends on construct nature

Study 2: Model Fit Comparison



Model: GDINA (de la Torre, 2011)

- Lower = better fit
- Dashed line = true Q_{true}
- Validated Q-matrices consistently show **best model fit** among the three
- Data-driven validation improves empirical fit beyond both the initial and true Q-matrices

Discussion: Summary of Findings

	Study 1 (SAD)	Study 2 (Fraction)
Initial OR	**0.90**	**0.71**
Validated OR	0.77	0.65
Initial Stability	0.918	**0.942**
Validated Stability	**0.934**	0.882
Initial Model Fit (BIC)	11181.91	5550.79
Validated Model Fit (BIC)	**11115.21**	**5525.81**
Best Plan	B/C/D (0.77)	A (0.77)
Challenging Attribute	A2 (close scrutiny)	A2 (simplifying)

- Initial Q-matrices had higher overlap with content-based references than validated Q-matrices
- Validation **improved** stability in Study 1 but **reduced** it in Study 2
- Validated Q-matrices consistently achieved better model fit in both studies
- No single finalization plan is universally optimal

Discussion: Initial vs. Validated Q-matrix in LLM

Pattern across both studies

- Validated OR $<$ Initial OR in both Study 1 (0.77 vs. 0.90) and Study 2 (0.65 vs. 0.71)
- Yet validated Q-matrices consistently achieved **better model fit** (lower AIC/BIC/SABIC) than initial ones

Interpretation

- Validation pushes Q-matrices toward **empirical fit** — optimizing response data patterns
- This can **deviate from item content/context**, producing entries inconsistent with substantive attribute definitions
- LLMs anchor on item wording $\rightarrow$ initial Q-matrices align better with content-based references (Q_1^*, Q_{true})

Implication

Overlap rate and model fit capture **different** aspects of Q-matrix quality — content validity vs. empirical fit. Both matter.

Discussion: Take-Home Messages

Survey vs. Achievement Test

- Non-cognitive constructs (SAD): cleaner attribute boundaries → higher initial OR
- Achievement tests (fraction): compound skills → lower initial, validation recovery and less stability

Single vs. Multi-Agent Systems

- **Single-agent LLM** suffices for content-based straightforward, well-defined constructs
- **Multi-agent LLM** preferable when integrating response data, data-driven validation, and researcher judgment

Finalization Plan Selection

- **Plan A**: single run, highest efficiency — best for simple, well-defined constructs, but unstable and dependent on survey content
- **Plans B/C**: 100 replications with probabilistic consensus — preferred when construct boundaries are ambiguous
- **Plan D**: 100 replications with final-matrix aggregation — most thorough; suited for complex, overlapping constructs, but computationally intensive and may not always improve performance

Thank You

Jihong Zhang, Ph.D.

jzhang@uark.edu

University of Arkansas

Questions and Comments?