

Validating LLM-Rated Item Features for Explanatory Item Response Models

Ahmed H. Bediwy

Pearson

ahmed.bediwy@pearson.com

Mubarak O. Mojoyinola

Ascend Learning

mubarak.mojoyinola@ascendlearning.com

Abstract

Explanatory item response models let test developers anticipate item difficulty from item design, but they depend on subject-matter experts rating every item on every hypothesized feature—a step that is slow, costly, and the practical bottleneck limiting how many items can be modeled. We ask whether large language models can supply those ratings. Four trained experts and six LLMs independently rated 15 grade-5 mathematics items on six psychometric features. We evaluate the LLM ratings twice: against the human consensus using quadratic-weighted κ and mixed-effects models, and against 1,452 student responses by using each source's ratings as the design matrix of a linear logistic test model (LLTM) benchmarked against a Rasch baseline. Claude Opus 4.7 performed best in recovering the item difficulty ($r = 0.89$), however, the human consensus against which agreement is measured performs worst at recovering Rasch difficulty ($r = 0.33$) compared to the rest of the LLMs. The reason is visible in the expert panel itself: inter-rater Fleiss' κ is at or below chance on three of six features, so the consensus is a noisy reference rather than ground truth. We also identify two failure modes (zero-variance and perfect collinearity) that agreement statistics cannot detect but make a feature unusable as an LLTM covariate. We argue that agreement should be reported alongside criterion-referenced recovery of item parameters, not in place of it.

1 Introduction

Modern assessment programs need far more items than expert committees can write and review. Automatic item generation addresses the writing bottleneck (Gierl and Lai, 2012; Morrison and Embretson, 2018), but it displaces the problem onto calibration: newly generated items have no response data, so their psychometric properties are unknown until they are field tested.

Explanatory item response theory offers a way out. Item-explanatory models such as the linear logistic test model (LLTM; Fischer, 1973) decompose Rasch item difficulty into the weighted contributions of theoretically motivated item features, so that difficulty can be anticipated from item design rather than recovered solely from responses (De Boeck and Wilson, 2004; Embretson, 1999). The approach has been applied to automatic item generation (Gierl and Lai, 2012), construct validity (Baghaei, 2008), item-position effects (Hohensinn et al., 2008; Demirkol and Kelecioğlu, 2022), and difficulty modeling (Hauenstein and Embretson, 2020; Toyama, 2021).

Every one of these applications requires the same expensive input. Before an LLTM can be estimated, subject-matter experts must rate every item on every hypothesized feature. This is the step that limits how many items can be examined and how quickly.

Large language models are an obvious candidate for that step, having matched or exceeded the performance of crowd and trained annotators on a range of text-labeling tasks (Gilardi et al., 2023). Work on predicting item difficulty directly from item text is also advancing (Benedetto et al., 2023; AlKhuyaey et al., 2024). What has not been examined is the intermediate case that matters for explanatory modeling: whether LLM-generated features yield explanatory IRT models that recover item parameters as well as human-rated features do.

Answering that question requires deciding what counts as validation. The default move is to treat expert ratings as ground truth and report agreement. We show that this default is misleading here. Our contributions are:

1. A dual evaluation of six LLMs as item-feature raters: agreement against a four-expert consensus, and criterion-referenced recovery of Rasch difficulty through an LLTM fitted to

1,452 student responses.

2. Evidence that the two criteria agree on how to rank *LLMs* but not on the standing of the *human reference*, which performs worst at recovering difficulty among all sources examined.
3. A diagnosis of why: the expert panel’s own inter-rater reliability is at or below chance on three of six features, so agreement is measured against a noisy standard.
4. Two failure modes specific to LLM raters—zero-variance and perfect collinearity—that are invisible to agreement statistics but make the affected features unusable as LLTM covariates.

2 Method

2.1 Items and responses

The analysis uses 15 mathematics items for grade 5, administered as part of a formative assessment. The items measure three Common Core standards: 5.NBT.B.5, 5.NF.B.7, and 5.MD.B.2. Item difficulty is estimated from 1,452 student responses, which provide the empirical basis for the Rasch and LLTM models.

2.2 Features and rubric

Six features were rated on a three-point ordinal scale: Cognitive Load (CL), Visual Complexity (VC), Cognitive Domain (CD), Cognitive Coherence (CC), Response Variability (RV), and Standard Specificity (SS). Appendix A gives the operational definitions and anchors.

All six are potential candidate predictors of item difficulty in the LLTM models except for cases where a feature has zero-variance or perfect collinearity with another feature.

2.3 Rating panels

Four experienced subject-matter experts rated every item independently. The human consensus is the unique plurality of the four ratings; 2–2 splits were resolved by expert adjudication. Six LLMs—Claude Haiku 4.5 (Anthropic, 2025), Claude Opus 4.7 (Anthropic, 2026a), Claude Sonnet 4.6 (Anthropic, 2026b), GPT-5.2 (OpenAI, 2025), Gemini 3.1 Pro (Google DeepMind, 2026a), and Gemini 3.5 Flash (Google DeepMind, 2026b)—rated the same items under a common rubric and zero-shot prompting.

2.4 Aim 1: characterizing rating differences

Because the rubric scales are ordered, agreement between each LLM and the human consensus is quantified with quadratic-weighted Cohen’s κ (Cohen, 1968), which penalizes a two-category disagreement more heavily than an adjacent one. Exact agreement, adjacent agreement, and ICC are also reported because weighted κ becomes unstable when one margin concentrates in a single category. Within-panel consistency is summarized by Fleiss’ κ across the four experts and across the six LLMs separately.

Systematic differences are modeled with three linear mixed models (Bates et al., 2015), each with a random intercept for item, since ratings of the same item are dependent across raters:

$$\text{rating} \sim \text{panel} + (1 \mid \text{item}),$$

The human panel is the reference throughout, so a positive contrast means the LLM rates higher. Individual experts, not the derived consensus was used in these models, so the human side carries its own sampling variability.

2.5 Aim 2: recovery of item difficulty

The baseline is a Rasch model estimated by conditional maximum likelihood (Rasch, 1960; Mair and Hatzinger, 2007). Against it we fit an LLTM in which difficulty is constrained to a linear combination of the six rated features,

$$b_j = - \sum_{k=1}^K w_{jk} \eta_k, \quad (1)$$

where w_{jk} is the (standardized) rating of item j on feature k and η_k is the estimated feature weight. A separate LLTM is fitted per rating source—the human consensus, the AI consensus, and each LLM individually—so that the fits differ only in $\mathbf{W}$.

The quantity of interest is Pearson’s correlation r between LLTM-reconstructed and freely estimated Rasch difficulties. We also report the likelihood-ratio test of the LLTM in comparison to Rasch baseline and RMSE.

Two problems arise with only 15 items and three rating categories, and both are properties of the rater rather than coding artifacts. A rating source that assigns every item the same category on a feature contributes a constant column (zero-variance); a rating source that rates two features identically contributes perfectly collinear columns.

Such columns are dropped during LLTM estimation, so the number of usable features K varies by rating source.

3 Results

3.1 LLM ratings agree only fairly with experts

Table 1 shows LLM agreement with the human consensus across the features. Mean quadratic-weighted κ separates the models clearly, from 0.50 for Claude Opus 4.7—moderate agreement—down to 0.10 for Claude Haiku 4.5. Only Opus exceeds the conventional threshold for moderate agreement; the remaining five sit in the fair range or below.

Agreement is far from uniform across features. Visual Complexity is rated consistently with experts by every model ($\kappa = 0.43$ to 0.78), whereas Response Variability is at or below chance for four of six models, reaching -0.49 for Gemini 3.1 Pro. Exact agreement (0.48 – 0.64) understates the picture somewhat, since adjacent agreement is high throughout (0.84 – 0.94): disagreements are usually one category, not two. With 15 items per model–feature cell these estimates are imprecise, and we read the pattern across features rather than any single coefficient.

3.2 Divergence is concentrated in two features

Pooled across features, the LLM panel rates only slightly higher than the expert panel ($\Delta = 0.12$, $SE = 0.05$, $p = .031$), and the item random intercept absorbs little variance ($ICC = 0.05$). After Bonferroni correction, no individual model differs significantly from the expert panel.

The interaction model locates the difference precisely (Table 2, Figure 1a). Four features show no detectable panel difference. Two show large ones: LLMs rate Cognitive Coherence 0.53 higher and Standard Specificity 0.54 higher than experts do (both $p < .001$). On a three-point scale, these are half-category shifts, and they push the LLM panel against the ceiling.

3.3 The expert panel is an unreliable reference

Table 3 and Figure 1b report within-panel consistency, and they reframe everything above. Fleiss’ κ among the four experts reaches 0.15 on no feature at all, and is *negative* on three of the six: Cognitive Coherence (-0.08), Response Variability (-0.17), and Standard Specificity (-0.09). On

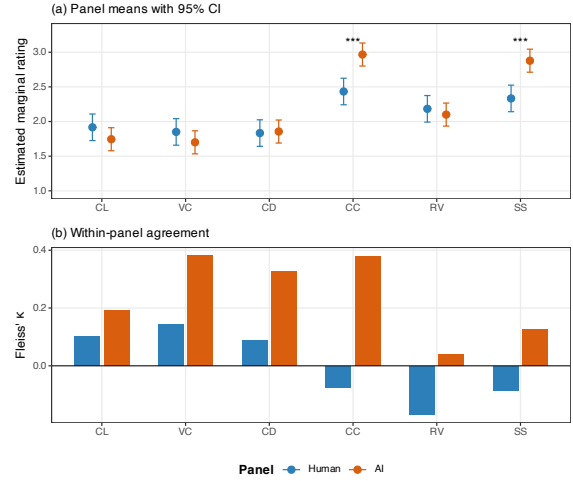


Figure 1: (a) Estimated marginal ratings by panel with 95% CIs; *** $p < .001$. (b) Within-panel agreement (Fleiss’ κ). The LLM panel is more internally consistent than the expert panel on every feature.

those three features, the experts disagree with each other slightly more than chance would predict.

The LLM panel is more consistent than the expert panel on every feature, in several cases by a wide margin (VC 0.38 versus 0.15 ; CC 0.38 versus -0.08). This does not make the LLMs correct—high agreement among models that share training data and prompt format is weak evidence—but it does mean that a low human–AI κ cannot be read as LLM error. Much of the disagreement quantified in Section 3.1 originates on the human side.

3.4 LLM features recover item difficulty better than expert features

The Rasch baseline fits well ($\log L = -8945.1$, 14 item parameters) and its difficulties correlate perfectly in rank with proportion correct ($\rho = -1.00$), confirming the calibration.

Table 4 reports the LLTM results. The ranking of the six LLMs based on how well they recover the rasch difficulty is exactly the same as the agreement ranking produced in Section 3.1: Opus, GPT-5.2, Sonnet, Flash, Gemini 3.1 Pro, and Haiku, in that order for both criteria. Agreement with experts, for all its limitations, does rank these models the same way their downstream psychometric model parameter recovery does.

The human consensus is where the two criteria differ. It is the *worst* of all eight sources used to estimate LLTM: its six-feature LLTM recovers Rasch difficulty at $r = 0.33$ ($R^2 = 0.11$). Five of six LLMs do better with the same or fewer covariates.

Model	CL	VC	CD	CC	RV	SS	Mean
Claude Opus 4.7	0.43	0.78	0.52	—	-0.18	0.94	0.50
GPT-5.2	0.09	0.76	0.08	0.36	0.06	0.55	0.32
Claude Sonnet 4.6	0.45	0.58	0.28	0.36	-0.12	0.23	0.30
Gemini 3.5 Flash	0.30	0.64	0.16	—	-0.08	—	0.26
Gemini 3.1 Pro	0.04	0.64	0.40	0.36	-0.49	0.23	0.20
Claude Haiku 4.5	0.29	0.43	-0.40	—	0.08	—	0.10

Table 1: Quadratic-weighted κ between each LLM and the human consensus, by feature, ordered by the mean. Dashes mark cells where κ is undefined because the model assigned a single category to all 15 items. Feature codes are defined in Section 2.2.

Feature	Human	AI	Δ	SE	t	p
CL	1.917	1.744	-0.172	0.110	-1.567	0.117
VC	1.850	1.700	-0.150	0.110	-1.365	0.173
CD	1.833	1.856	0.022	0.110	0.202	0.840
CC	2.433	2.967	0.533	0.110	4.853	0.000 ***
RV	2.183	2.100	-0.083	0.110	-0.758	0.448
SS	2.333	2.878	0.544	0.110	4.954	0.000 ***

Table 2: Estimated marginal ratings by panel and the AI–Human contrast from the panel $\times$ feature mixed model (Bonferroni-adjusted). *** $p < .001$.

Feature	κ_F Human	κ_F AI	ICC Human	ICC AI
CL	0.101	0.193	0.407	0.337
VC	0.145	0.382	0.632	0.880
CD	0.089	0.328	0.381	0.803
CC	-0.076	0.379	0.255	0.800
RV	-0.169	0.040	-0.159	0.627
SS	-0.085	0.128	0.391	0.596

Table 3: Within-panel inter-rater agreement: Fleiss’ κ and average-measure ICC for the four experts and the six LLMs.

Claude Opus 4.7 reaches $r = 0.89$ ($R^2 = 0.79$) using five features; GPT-5.2 reaches $r = 0.85$ and Claude Sonnet 4.6 $r = 0.81$, both with six. The gap is not a parameter-count artifact: the human consensus has the *most* covariates available and still performs worst.

Figure 2 makes the difference concrete. Under the human consensus, the reconstructed difficulties are compressed into a narrow band around zero—the features carry almost no signal about which items are hard. Under Claude Opus 4.7, they track the Rasch estimates closely across the full range.

The likelihood-ratio tests all reject the LLTMs ($p < .001$ throughout), as expected at this sample size.

3.5 Zero-variance and perfect collinearity

Two failure modes emerge that agreement statistics cannot detect.

Zero-variance. The AI consensus assigns category 3 to all 15 items on both Cognitive Coherence and Standard Specificity. The variance is exactly zero, so neither feature can be used in the LLTM: the AI consensus is left with four usable covariates rather than six. The same occurs for Claude Haiku 4.5 and Gemini 3.5 Flash, and for Cognitive Coherence alone in Claude Opus 4.7. This is the Aim 1 result seen from another angle—the two features on which LLMs significantly overrate are the two they push to the ceiling.

Perfect collinearity. Gemini 3.1 Pro assigns identical ratings to Cognitive Coherence and Standard Specificity on every item ($r = 1.00$), making the columns perfectly collinear. Therefore, one of them has to be dropped during model estimation.

Table 5 shows the consequence for the estimated weights. Where the human weights are uniformly small ($|\eta| \leq 0.18$), the AI consensus places large weights on Visual Complexity (0.52) and Cognitive Domain (-0.62) which conforms to difficulty theory, whereas the noisy human consensus does not.

Figure 3 shows that the sources disagree about which features drive difficulty. The largest weight belongs to a different feature for almost every source: Cognitive Load for Claude Opus 4.7 and Gemini 3.5 Flash, Visual Complexity for GPT-5.2, Cognitive Domain for the human and AI consensus, Cognitive Coherence for Claude Sonnet 4.6 and Gemini 3.1 Pro, and Response Variability for Claude Haiku 4.5. Recovering difficulty well and recovering it *for the same reasons* are different achievements, and only the first is established here.

4 Discussion

The two validation criteria agree about the LLMs and disagree about the humans, and that asymmetry is the finding. Ranking the six models by agreement with experts reproduces their ranking by dif-

Rating source	K	$\log L$	BIC	LR	df	r	R^2	RMSE
Rasch (baseline)	—	-8945.11	17992.14	—	—	1.00	1.00	0.00
Claude Opus 4.7	5	-9279.83	18596.06	669.44	9	0.89	0.79	0.41
GPT-5.2	6	-9391.41	18826.51	892.61	8	0.85	0.72	0.47
Claude Sonnet 4.6	6	-9488.60	19020.89	1086.99	8	0.81	0.66	0.52
Gemini 3.5 Flash	4	-9703.44	19436.00	1516.67	10	0.70	0.49	0.64
Gemini 3.1 Pro	5	-9849.80	19736.01	1809.40	9	0.65	0.43	0.68
AI consensus	4	-10069.71	20168.54	2249.20	10	0.52	0.27	0.77
Claude Haiku 4.5	4	-14039.58	28139.09	2558.73	10	0.34	0.12	0.84
Human consensus	6	-10282.84	20609.36	2675.46	8	0.33	0.11	0.85

Table 4: LLTM fit and recovery of Rasch difficulty by rating source. K is the number of usable feature covariates after dropping constant and aliased columns; r is the correlation between LLTM-reconstructed and freely estimated Rasch difficulties. Sources are ordered by r .

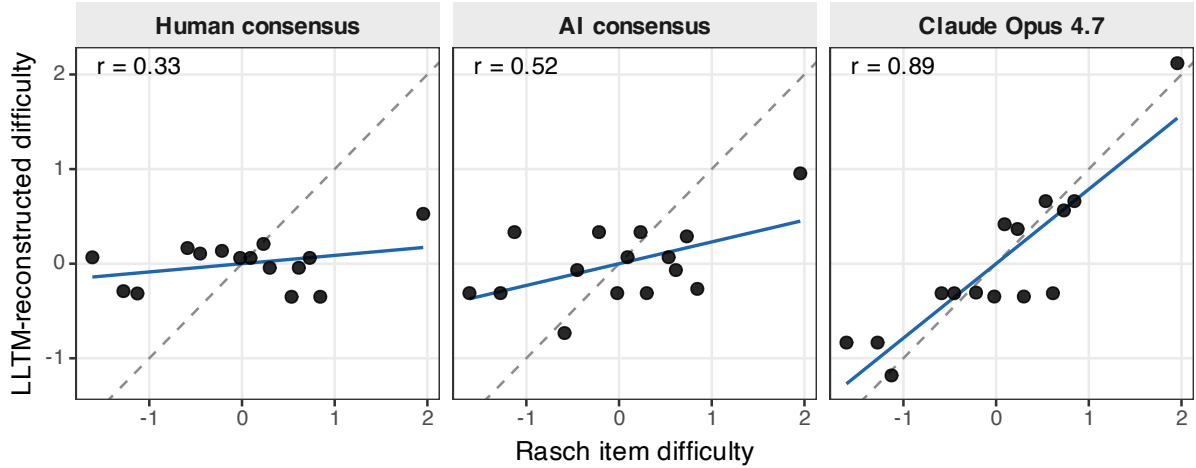


Figure 2: LLTM-reconstructed versus freely estimated Rasch difficulty for three rating sources. The dashed line is the identity line; the solid line is the least-squares fit.

Feature	Human consensus	AI consensus
CL	-0.067 (0.019)	-0.164 (0.028)
VC	0.176 (0.024)	0.518 (0.027)
CD	-0.184 (0.019)	-0.617 (0.035)
CC	-0.127 (0.023)	constant
RV	0.089 (0.018)	-0.197 (0.028)
SS	0.123 (0.025)	constant

Table 5: LLTM feature weights η (standard errors) on the difficulty scale, for standardized covariates. “constant” marks features the source rated identically on all 15 items.

difficulty recovery exactly. Agreement is therefore doing useful work: it is a cheap, response-free proxy that orders candidate raters correctly. What it cannot do is tell you where the reference standard itself sits on that scale—and here the expert consensus lands at the bottom, recovering difficulty at $r = 0.33$ against 0.89 for the best model.

The resolution is that the expert consensus is not ground truth. With Fleiss’ κ at or below zero on three features, aggregating four experts produces a

reference standard that is substantially noisy. An agreement statistic computed against it is bounded above by its reliability, so “only fair agreement with experts” and “better than experts at explaining difficulty” are not in tension once the reference is recognized as noisy.

This has a practical consequence for anyone considering LLMs for item review. Agreement with a small expert panel is cheap and is the field’s default, and our results support using it to *rank* candidate raters. They do not support using it to decide whether a rater is good enough, because that judgment is only as good as the panel. Where response data exist, criterion-referenced recovery of item parameters answers the question agreement cannot, and the two should be reported together.

The zero-variance and perfect collinearity problems identified in some LLM ratings sharpen the recommendation. Both are silent in agreement statistics—a model that rates every item 3 can still post a respectable exact-agreement rate when the

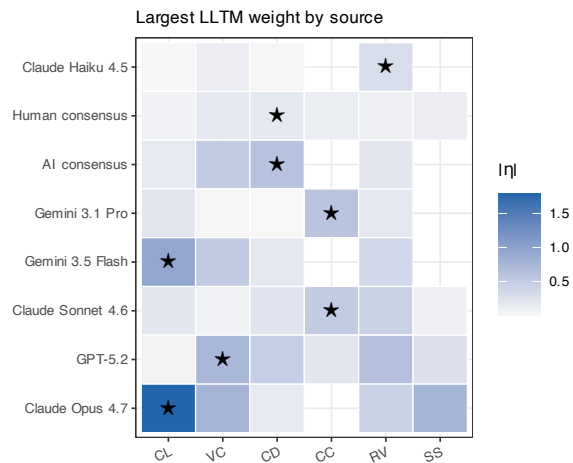


Figure 3: Absolute LLTM feature weights by rating source.

human mode is also 3, but both are fatal for explanatory modeling, since a constant or collinear covariate carries no information. Screening candidate raters for category variance and inter-feature collinearity costs nothing and should precede any agreement analysis.

These results support a narrower claim than “LLMs can replace expert raters.” Three features concentrate most of the trouble—Cognitive Coherence, Response Variability, and Standard Specificity. They are where the experts disagree with each other, where the LLMs saturate, and where one model collapses two rubric entries onto a single judgment. Yet several models place their largest difficulty weights on exactly these features. A feature can be unreliable, saturated, and still carry signal, which points at the rubric rather than the raters: these three anchors need sharper operational definitions before either panel can be held to them, and that is where we would direct effort next.

Limitations

The item sample is the binding constraint. Fifteen items with response data support a Rasch calibration comfortably but leave the LLTM comparison badly underpowered. The ordering of sources in Table 4 is consistent and large, but this study cannot establish its magnitude, and we do not claim statistical support for the difference. Extending the calibration sample is the single most valuable next step.

The human panel consists of four raters. Their low inter-rater reliability is central to our interpretation, yet it is itself estimated imprecisely from 15

items, and we cannot separate genuine construct ambiguity from insufficient rater training. A larger and better-calibrated expert panel might close the gap we report.

Each LLM rated each item once, using default settings and a single prompt. We therefore cannot distinguish stable model behavior from prompt sensitivity or sampling variability, and the within-LLM agreement we report is inflated relative to what independent replications would show. The models also share architectural lineage and training data, so their mutual agreement is not evidence of validity.

Recovery of Rasch difficulty is a proxy for usefulness, not a direct measure of it. A design matrix could recover in-sample difficulty well and still fail to predict parameters for genuinely new items, which is the use case that motivates explanatory modeling. We report in-sample recovery only; out-of-sample prediction for generated items is left to future work.

Finally, the findings are specific to grade-5 mathematics, three Common Core standards, and this six-feature rubric. Whether the dissociation between agreement and recovery generalizes to other domains, grade levels, or feature sets is untested.

References

- Samah AlKhuzayy, Floriana Grasso, Terry R. Payne, and Valentina Tamma. 2024. Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, 34(3):862–914.
- Anthropic. 2025. [Claude haiku 4.5](#). Large language model.
- Anthropic. 2026a. [Claude opus 4.7](#). Large language model. Model identifier: claude-opus-4-7.
- Anthropic. 2026b. [Claude sonnet 4.6](#). Large language model. Model identifier: claude-sonnet-4-6.
- Purya Baghaei. 2008. The Rasch model as a construct validation tool. *Rasch Measurement Transactions*, 22(1):1145–1146.
- Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.
- Luca Benedetto, Paolo Cremonesi, Andrew Caines, Paula Buttery, Andrea Cappelli, Andrea Giussani, and Roberto Turrin. 2023. A survey on recent approaches to question difficulty estimation from text. *ACM Computing Surveys*, 55(9):1–37.

- Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.
- Paul De Boeck and Mark Wilson, editors. 2004. *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. Springer, New York.
- Sinem Demirkol and Hülya Kelecioğlu. 2022. Analyzing the interaction of item position effect and student characteristics within explanatory IRT models. *Journal of Measurement and Evaluation in Education and Psychology*, 13(4):282–304.
- Susan E. Embretson. 1999. Generating items during testing: Psychometric issues and models. *Psychometrika*, 64(4):407–433.
- Gerhard H. Fischer. 1973. The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37(6):359–374.
- Mark J. Gierl and Hollis Lai. 2012. The role of item models in automatic item generation. *International Journal of Testing*, 12(3):273–298.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.
- Google DeepMind. 2026a. [Gemini 3.1 pro](#). Large language model.
- Google DeepMind. 2026b. [Gemini 3.5 flash](#). Large language model.
- Clifford E. Hauenstein and Susan E. Embretson. 2020. Modeling item difficulty in a dynamic test. *Journal of Cognitive Education and Psychology*, 19(2):93–106.
- Christine Hohensinn, Klaus D. Kubinger, Manuel Reif, Stefana Holocher-Ertl, Lale Khorramdel, and Martina Frebort. 2008. Examining item-position effects in large-scale assessment using the linear logistic test model. *Psychology Science Quarterly*, 50(3):391–402.
- Patrick Mair and Reinhold Hatzinger. 2007. Extended Rasch Modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, 20(9):1–20.
- Kristin M. Morrison and Susan E. Embretson. 2018. Item generation. In Paul Irwing, Tom Booth, and David J. Hughes, editors, *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development*, volume 1, pages 75–94. Wiley.
- OpenAI. 2025. [GPT-5.2](#). Large language model.
- Georg Rasch. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research, Copenhagen.
- Yukie Toyama. 2021. What makes reading difficult? an investigation of the contributions of passage, task, and reader characteristics on comprehension performance. *Reading Research Quarterly*, 56(4):633–642.

A Rating Rubric

All raters, human and LLM, received the same six-feature rubric. Each feature is rated 1–3 with the anchors below; the full prompt is released with the analysis code.

Cognitive Load Number of mental steps required. 1 = single-step retrieval or one operation; 2 = two connected operations, or one operation plus context interpretation; 3 = three or more operations or nested reasoning.

Visual Complexity Density of visual information. 1 = text only or a simple table; 2 = a clear diagram or basic graph that supports understanding; 3 = a complex or cluttered visual requiring interpretation.

Cognitive Domain Predominant type of mathematical thinking. 1 = procedural (applying an algorithm); 2 = conceptual (understanding relationships or why a procedure works); 3 = application (solving a contextual problem strategically).

Cognitive Coherence Clarity and singularity of the solution path. 1 = multiple valid pathways or an ambiguous statement; 2 = one clear intended path with possible alternatives; 3 = a single obvious strategy and an unambiguous statement.

Response Variability Quality of the distractors. 1 = arbitrary distractors inviting random guessing; 2 = distractors representing some common errors; 3 = distractors systematically representing documented misconceptions.

Standard Specificity Precision of alignment to one standard. 1 = draws on several standards or grade levels; 2 = one main standard plus supporting knowledge from another; 3 = tightly focused on a single standard.