Psychometric Analysis of MRBench V2

Anonymous Author(s) ANONYMOUS@DOMAIN.COM *Anonymous Institution*

Abstract

Benchmarks for evaluating the pedagogical ability of LLM tutors are increasingly central to educational AI research, yet their psychometric properties are rarely examined formally. We apply a comprehensive measurement validation pipeline to MRBench V2, a benchmark of 200 mathematical tutoring dialogues annotated across eight pedagogical dimensions. Using exploratory and confirmatory factor analyses, graded response modelling, item-level validity diagnostics, measurement invariance testing, and generalizability theory, we find that six of the eight dimensions form a coherent unidimensional scale with excellent structural fit (CFI = 0.998, RMSEA = 0.058) and strong generalizability ($G_{\text{rel}} = 0.974$). Two dimensions show psychometric properties inconsistent with their intended role. We further detect measurement non-equivalence across model sizes and highlight open questions regarding gaps in construct validity and predictive validity. Our findings demonstrate the value of psychometric analysis as a standard practice in educational AI evaluation.

Keywords: psychometric validation, AI benchmarks, large language models, item response theory, AI tutors, AI evaluation

1. Introduction

Large language models are increasingly deployed in educational applications, including AI-powered tutoring systems and automated feedback tools that provide personalized, adaptive instruction through natural language interaction (Latif et al., 2026). This has generated substantial interest in understanding the capabilities and limitations of LLMs as pedagogical agents. Research in educational NLP suggests that LLMs are often misaligned with pedagogical objectives: despite strong performance on surface-level qualities such as fluency and human-likeness, LLMs consistently underperform on deeper dimensions. Research has shown that they tend to reveal answers rather than scaffold student reasoning (Macina et al., 2023a), struggle to engage students in genuine problem-solving (Tack and Piech, 2022), and lack the expert decision-making processes that distinguish effective remediation from superficial acknowledgement of student errors (Wang et al., 2024). A key reason is that modern foundation models are alignment-tuned to be broadly helpful assistants, a design choice that is often at odds with effective pedagogy (Jurenka et al., 2024). These findings highlight the need for robust, standardized frameworks that evaluate pedagogical behaviors.

1.1. Pedagogical Benchmarking for LLM Tutors

Evaluating LLMs as tutors has attracted growing attention, though the field has historically lacked a standardized evaluation framework. Early work by Tack and Piech (2022) assessed tutor responses along three dimensions, speaking like a teacher, understanding the student, and helping the student, while Macina et al. (2023a) focused on coherence, correctness, and equitable tutoring. Wang et al. (2024) evaluated usefulness, care, and human-soundingness. The fragmentation across these frameworks makes cross-study comparison

difficult. Maurya et al. (2025) address this directly by proposing the first unified evaluation taxonomy with eight pedagogical dimensions, releasing MRBench as a benchmark of annotated mathematical tutoring dialogues. This taxonomy was subsequently operationalized in the BEA-2025 Shared Task on Pedagogical Ability Assessment (Kochmar et al., 2025). Despite this progress, benchmark development in educational AI has relied primarily on expert-grounded item construction and inter-annotator agreement as the main validity evidence, leaving structural validity, construct coherence, and measurement equivalence across model groups largely unexamined. Educational testing is the historical foundation of modern psychometrics (Embretson and Reise, 2000), and the measurement tools it produced, including item response theory, generalizability theory, and confirmatory factor analysis, were developed precisely to address these gaps. These methods are now being applied to AI benchmarks (Truong et al., 2025) more broadly, yet formal psychometric examination of pedagogical benchmarks specifically remains absent from the literature, despite their direct connection to the educational testing tradition that gave rise to these methods.

1.2. Our Work

We use MRBench V2 as a case study to examine the psychometric properties of pedagogical AI benchmarks more broadly, with the goal of strengthening measurement practice in the field rather than critiquing any particular resource. Using exploratory and confirmatory factor analyses, graded response modelling, item-level validity diagnostics, measurement invariance testing, and generalizability theory, we ask whether the eight-dimension taxonomy holds under formal measurement scrutiny. We find that a six-item reduced scale forms a coherent unidimensional construct with excellent reliability, while also identifying measurement non-equivalence across model scale tiers with implications for cross-tier comparisons. We further raise open questions about what the recovered construct actually measures, and whether current pedagogical benchmarking practice adequately captures pedagogical ability as understood in the learning sciences.

2. Methods

2.1. Data

We use MRBench V2 (Maurya et al., 2025), a publicly available dataset of 200 mathematical tutoring dialogues drawn from MathDial (Macina et al., 2023b) and Bridge (Wang et al., 2024). Each dialogue is annotated across eight pedagogical dimensions ("items" in measurement science) for responses from seven LLM tutors (GPT-4, Gemini, Claude Sonnet, Mistral, Llama 3.1 8B, Llama 3.1 405B, and Phi-3) and two human tutors (Expert and Novice). The eight dimension assess whether the tutor: (1) *identifies* the student's mistake; (2) *locates* it precisely in the student's response; (3) *avoids revealing* the final answer; (4) *provides guidance* such as hints, explanations, or supporting questions; (5) produces an *actionable* response that makes clear what the student should do next; (6) responds *coherently* with the student's prior turns; (7) uses an *encouraging tone* rather than neutral or offensive; and (8) sounds *human-like* rather than robotic. Six dimensions use a Yes/To some extent/No ordinal scale; Revealing Answer uses a correctness-aware scheme (Yes - correct / Yes - incorrect / No); and Tutor Tone uses an affect scale (Encouraging / Neutral /

Offensive). All scales are encoded as 0, 1, 2 for analysis, where higher values reflect the desired pedagogical behavior for all dimensions except Revealing Answer, where 0 (No) is the desirable outcome. Annotations were assigned by four trained human annotators following a structured guideline, with an average inter-annotator agreement of $\kappa = 0.71$ on a double-annotated subset (Maurya et al., 2025).

Five conversation identifiers contained duplicate entries in the raw JSON; all duplicates were resolved by averaging scores and rounding to the nearest integer. Human tutors (Expert and Novice) were excluded from the analysis, as our focus is on the psychometric properties of MRBench as an instrument for LLM evaluation; including them would introduce a heterogeneous response distribution across fundamentally different kinds of agents, potentially biasing item parameter estimates for the target population (Brennan, 1992). The final dataset comprises $N = 1{,}365$ observations (195 conversations $\times$ 7 LLM tutors). For the differential item functioning and measurement invariance analyses, tutors are additionally classified by model size: *large* models (GPT-4, Gemini, Sonnet, Mistral, Llama 3.1 405B; $n = 975$) and *small* models (Llama 3.1 8B, Phi-3; $n = 390$).

2.2. Analysis

All analyses were implemented in R v4.5.1 (R Core Team, 2021). We apply a sequential psychometric validation pipeline to the full 8-item response matrix unless otherwise noted.

We begin with descriptive statistics (means, standard deviations, skewness, kurtosis, score distributions) and estimate inter-dimension associations using polychoric correlations (Olsson, 1979), appropriate for ordinal data. Item-level validity is assessed via corrected item-total correlations (CITC) and Cronbach's α with each dimension deleted (Cronbach, 1951); dimensions with CITC < 0.10 and $\alpha_{-i} > \alpha$ are flagged as validity threats. To examine the factor structure of MRBench, we conduct parallel analysis (Horn, 1965) on the polychoric matrix to determine the empirically supported number of factors, followed by maximum likelihood EFA for 1-4 factors with oblimin rotation. We then fit two CFA models: M1 (unidimensional, all 8 items) and M2 (unidimensional, 6 items excluding Tutor Tone and Revealing Answer). Their fits are evaluated against accepted thresholds: CFI, TLI > 0.95, RMSEA < 0.06, and SRMR < 0.08 (Hu and Bentler, 1999).

Next, we fit a unidimensional Graded Response Model (Samejima, 1997) to both the full 8-item and 6-item matrices using the `mirt` package (Chalmers, 2012), extracting discrimination parameters (a), difficulty thresholds (b_1, b_2), and item fit statistics (S-χ^2). Latent ability scores ($\hat{\theta}$) are estimated via Expected A Posteriori scoring. We additionally test configural, metric, and scalar measurement invariance of the 6-item scale across model size groups using the WLSMV estimator with Satorra-Bentler scaling (Satorra and Bentler, 1994); metric invariance failure (ΔCFI < -0.010) would indicate that factor loadings differ across groups, precluding direct $\hat{\theta}$ comparisons.

Finally, we apply generalizability theory by fitting a fully crossed Conversation $\times$ Tutor $\times$ Dimension random effects model to the 6-item matrix via REML (Corbeil and Searle, 1976), decomposing total score variance into seven sources. Relative (G_{rel}) generalizability coefficient is computed with tutor as the object of measurement.

Table 1: Descriptive statistics and item-level validity diagnostics for MRBench V2 dimensions (N = 1,365; LLM tutors only). CITC = corrected item-total correlation; α_{-i} = Cronbach's α if dimension deleted (full-scale α = 0.750). Score distributions show percentage of responses at each ordinal level (0/1/2). Dimensions marked † show CITC < 0.10 and $\alpha_{-i} > \alpha$, indicating active degradation of scale coherence.

Dimension	**Mean**	**SD**	**Skew**	**Kurt.**	**%0**	**%1**	**%2**	**CITC**	$\boldsymbol{\alpha_{-i}}$
Mistake Identification	1.65	0.73	−1.73	1.10	15.4	3.8	80.8	0.753	0.661
Mistake Location	1.38	0.88	−0.82	−1.20	26.7	8.3	65.0	0.777	0.644
Providing Guidance	1.40	0.76	−0.82	−0.81	16.8	26.2	57.1	0.667	0.678
Actionability	1.20	0.92	−0.40	−1.69	34.1	12.2	53.8	0.342	0.752
Humanlikeness	1.84	0.50	−3.03	7.88	5.5	5.1	89.4	0.473	0.725
Coherence	1.70	0.66	−1.94	2.09	11.3	7.4	81.3	0.709	0.677
Revealing Answer†	0.33	0.73	1.78	1.27	82.1	2.6	15.4	0.021	0.797
Tutor Tone†	1.34	0.47	0.70	−1.52	0.0	66.4	33.6	−0.109	0.791

3. Results

3.1. Dimension-Level Diagnostics

Descriptive statistics reveal skewed distributions for two dimensions (Table 1). Revealing Answer exhibits a severe floor effect (82.1% scoring 0) with the intermediate category nearly empty (2.6%). Tutor Tone shows near-zero variance with 66.4% of responses at the neutral category and no offensive responses, consistent with Maurya et al. (2025). As a preliminary item-level screen prior to factor analysis, CITC diagnostics confirm both dimensions as validity threats: Revealing Answer (CITC = 0.021) is effectively orthogonal to the remaining dimensions, and Tutor Tone (CITC = −0.109) measures something in opposition to the construct. Deleting either dimension increases α from 0.750 to 0.797 and 0.791 respectively, which is a defining signature of a validity-degrading item (DeVellis, 2016). The remaining six dimensions show acceptable CITC values of 0.342-0.777.

3.2. Structural Validity

The polychoric correlation matrix (Figure 1) reveals a coherent positive manifold among the six clean dimensions (correlations 0.47–0.94), with Mistake Identification and Mistake Location showing the strongest association (r = 0.944). Both flagged dimensions disrupt this structure: Tutor Tone correlates negatively with all other dimensions (r = −0.234 to −0.010), and Revealing Answer correlates negatively with Actionability (r = −0.58), consistent with the theoretical interdependency acknowledged by Maurya et al. (2025), whereby tutors who reveal answers are less likely to produce actionable responses.

Parallel analysis on the polychoric matrix suggests four factors, but inspection of the scree plot reveals a dominant first eigenvalue (λ_1 = 4.52) that substantially exceeds the remaining eigenvalues, with only two others exceeding 1.0 (λ_2 = 1.56, λ_3 = 1.02), a pattern consistent with essential unidimensionality rather than genuine multidimensionality. The suggested four-factor solution is an artefact of the two degenerate dimensions generating

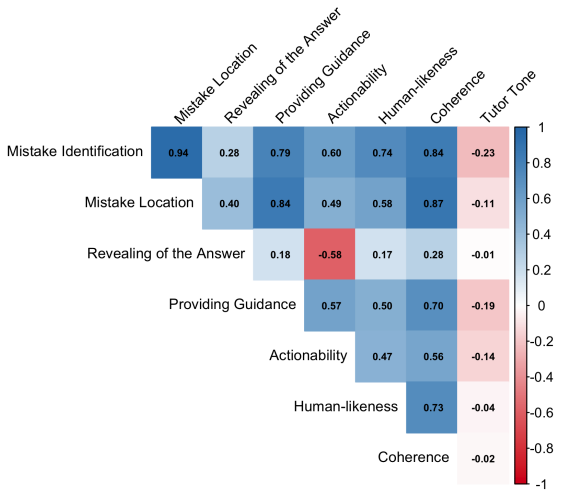

Figure 1: Polychoric correlation matrix for MRBench V2 dimensions (LLM tutors only, $N = 1{,}365$). Note the near-zero and negative correlations of Tutor Tone with all other dimensions, and the negative correlation between Revealing Answer and Actionability ($r = -0.583$).

Table 2: CFA model fit indices (WLSMV estimator, $N = 1{,}365$).

Model	**CFI**	**TLI**	**RMSEA**	**[90% CI]**	**SRMR**
M1: Unidimensional (all 8)	0.937	0.911	0.235	[0.225, 0.245]	0.246
M2: Unidimensional (6 items)	**0.998**	**0.997**	**0.058**	[0.043, 0.074]	**0.051**

artificial orthogonal variance. The one-factor EFA solution accounts for 53.6% of variance, with six dimensions loading strongly ($\lambda = 0.57$-0.97); Tutor Tone falls below the 0.30 threshold entirely and Revealing Answer loads at 0.321. CFA results confirm this picture (Table 2). The full 8-item unidimensional model (M1) fits poorly. In contrast, the 6-item unidimensional model (M2) achieves excellent fit, establishing the reduced scale as a psychometrically coherent instrument.

3.3. Item Response Analysis

GRM results corroborate the CFA findings at the item level (Table 3). The six clean dimensions show discrimination parameters ranging from $a = 1.51$ (Actionability) to $a = 10.50$ (Mistake Location). In stark contrast, Revealing Answer has $a = 0.105$ (essentially non-discriminating) with difficulty thresholds of $b_1 = 14.5$ and $b_2 = 16.3$, implying the item could only be passed by a tutor of implausibly high latent ability. Tutor Tone has a negative discrimination ($a = -0.203$), meaning higher latent pedagogical quality is associated with *lower* scores on this dimension; only one threshold was estimable ($b_1 = -3.397$, $b_2 = \text{NA}$) due to near-zero response variance.

Table 3: GRM factor loadings (F_1), communalities (h^2), and graded response model item parameters (a = discrimination; b_1, b_2 = difficulty thresholds) for all eight dimensions (LLM tutors only). Clean 6-item parameters shown; full 8-item parameters for flagged dimensions shown in italics.

Dimension	F_1	h^2	a	b_1	b_2
Mistake Identification	0.966	0.934	6.437	−1.025	−0.869
Mistake Location	0.987	0.974	10.503	−0.611	−0.389
Providing Guidance	0.861	0.742	2.893	−1.139	−0.236
Actionability	0.654	0.427	1.505	−0.634	−0.165
Humanlikeness	0.726	0.527	1.820	−2.239	−1.706
Coherence	0.886	0.785	3.300	−1.354	−0.984
Revealing Answer†	0.062	0.004	*0.105*	*14.500*	*16.264*
Tutor Tone†	−0.119	0.014	*−0.203*	*−3.397*	—

Table 4: Tutor latent ability estimates ($\hat{\theta}$) from the graded response model, ranked by clean 6-item scale score. $\Delta\hat{\theta}$ = difference between full 8-item and clean 6-item estimates. Rankings are identical across both models. Note: direct comparison between large and small models should be interpreted cautiously given metric non-invariance

Rank	**Tutor**	**Size**	$\hat{\theta}$ **(8-item)**	$\hat{\theta}$ **(6-item)**	$\Delta\hat{\theta}$
1	Llama 3.1 405B	Large	0.427	0.420	−0.007
2	GPT-4	Large	0.292	0.287	−0.005
3	Mistral	Large	0.260	0.253	−0.007
4	Sonnet	Large	0.209	0.229	+0.020
5	Gemini	Large	0.094	0.102	+0.008
6	Llama 3.1 8B	Small	−0.139	−0.142	−0.003
7	Phi-3	Small	−1.160	−1.150	+0.010

Tutor ability estimates from the clean 6-item GRM rank Llama 3.1 405B highest ($\hat{\theta}$ = 0.420), followed by GPT-4 (0.287), Mistral (0.253), Sonnet (0.229), Gemini (0.102), Llama 3.1 8B (−0.142), and Phi-3 (−1.150). Rankings are *identical* between the full 8-item and clean 6-item models (maximum $|\Delta\hat{\theta}| = 0.020$), indicating that while the degenerate dimensions corrupt construct interpretation, they do not distort the resulting leaderboard (Table 4).

3.4. Measurement Invariance

Measurement invariance testing of the 6-item scale across large and small LLMs reveals a metric invariance failure (Table 5). The metric model shows ΔCFI = −0.014, exceeding the −0.010 threshold, indicating that factor loadings differ significantly between large and small models ($\chi^2_{\text{diff}} = 131.86$, $p < .001$). This means the six dimensions do not relate to the underlying pedagogical quality construct in the same way across model scale tiers: the instrument is functioning differently for frontier models than for smaller models. As a consequence, direct $\hat{\theta}$ comparisons between large and small LLMs (such as ranking

Table 5: Measurement invariance of the 6-item unidimensional scale across large and small LLMs (WLSMV estimator). ΔCFI threshold for invariance: > -0.010. Metric invariance failure indicates factor loadings differ across model scale groups

Model	**CFI**	**RMSEA**	**ΔCFI**	**Supported**
Configural	0.999	0.039	—	Baseline
Metric	0.985	0.125	−0.014	**No**
Scalar	0.997	0.053	+0.012	Yes

Table 6: Generalizability theory variance decomposition for the 6-item MRBench scale.

Source	σ^2	%
Residual (C×T×D)	0.229	35.4
Tutor (T)	0.149	23.1
C × T	0.131	20.2
Dimension (D)	0.055	8.5
C × D	0.050	7.8
T × D	0.018	2.8
Conversation (C)	0.014	2.2
Total	0.646	100

Llama 3.1 405B against Phi-3 on a common scale) are not fully justified under the current instrument. We note that the small model group comprises only two tutors ($n = 390$), and the negative observed variable variance warnings from *lavaan* suggest estimation instability in this group; replication with a larger sample of small models is needed to confirm.

3.5. Generalizability

The fully crossed G-theory model decomposes score variance into seven sources (Table 6). Tutor variance accounts for 23.1% of total variance, substantially exceeding conversation variance (2.2%), which implies that MRBench scores primarily differentiate tutors rather than reflecting dialogue difficulty. The C × T interaction, 20.2% is nearly as large as the tutor main effect, indicating that tutoring quality is partly context-dependent: tutors differ not only in overall ability but in how they respond to specific dialogue types. The relative generalizability coefficient is $G_{\text{rel}} = 0.974$, well above the 0.90 threshold (Brennan, 1992).

4. Discussion

This study performs a psychometric validation of MRBench V2 (Maurya et al., 2025), asking whether its eight-dimension taxonomy holds under formal measurement scrutiny. The primary constructive finding is that six of the eight dimensions form a coherent, reliable and generalisable instrument. The reduced six-item scale achieves excellent structural fit (CFI = 0.998, RMSEA = 0.058) and strong generalizability ($G_{\text{rel}} = 0.974$), with tutor variance substantially exceeding conversation variance. This confirms that MRBench, in its reduced form, is doing its intended job: scores primarily reflect differences in tutoring quality

rather than dialogue difficulty. The large conversation-by-tutor interaction (20.2%) further suggests that pedagogical quality is partly context-dependent, a substantively interesting finding that motivates future work on context-stratified evaluation designs that account for dialogue type when comparing tutors. Two dimensions, Tutor Tone and Revealing Answer, show psychometric properties inconsistent with their intended role. The near-zero variance of Tutor Tone reflects an empirical reality noted by Maurya et al. (2025) themselves: LLM tutors produce virtually no offensive responses, leaving the dimension unable to discriminate. Revealing Answer functions more as an indicator of a rare adverse event than a continuous quality dimension, as evidenced by its severe floor effect and negative correlation with Actionability. We note that MRBench V3 and the BEA-2025 Shared Task (Kochmar et al., 2025) independently reduced the taxonomy, suggesting that practitioners have arrived at similar conclusions through experience; and our analysis corroborates with that. Measurement invariance testing reveals that the six-item scale functions differently across large and small LLMs, with metric invariance failing across model scale tiers. This suggests some caution is warranted when making direct ability comparisons between frontier and smaller models on MRBench V2. We stress that this finding is preliminary, as the small-model group comprises only two tutors, and should be revisited with more diverse model families.

A deeper question concerns what the recovered six-item construct actually represents. The factor analysis reveals a single dominant latent dimension underlying mistake identification, mistake location, providing guidance, actionability, coherence, and humanlikeness. While this unidimensional structure is psychometrically desirable, it raises an interpretive question: is this factor genuinely measuring pedagogical ability, or something more general? The six dimensions share a common thread: they all reflect whether a tutor produces a well-formed, contextually appropriate, and informationally complete response to a student mistake, behaviors that a capable question-answering system might also exhibit without necessarily engaging in genuine tutoring. Whether this constitutes pedagogical ability in the richer sense understood by learning scientists is an open empirical question rather than a settled claim. Constructs such as adaptive scaffolding, metacognitive support, and sustained dialogic inquiry may or may not be captured by the recovered factor, and this cannot be determined from internal structure alone. Answering it requires construct validity evidence beyond factor analysis: convergent validity studies examining whether MRBench scores correlate with related constructs such as conversational uptake (Tack and Piech, 2022), expert decision-making quality (Wang et al., 2024), or independent pedagogical ratings, would help clarify what the latent dimension represents. Equally important is predictive validity: a benchmark claiming to measure pedagogical ability should, in principle, predict outcomes that matter educationally, such as student learning gains, engagement, or satisfaction. Establishing such evidence represents an important direction for future work. More broadly, our results highlight the value of psychometric validation as a complement to the expert-judgment and inter-annotator agreement approaches that currently dominate benchmark development. We hope these methods, including item-level diagnostics, factor analysis, IRT, and generalizability theory, become routine tools in the benchmark development cycle, not as a critique of existing work but as a way to strengthen the measurement foundations of the field.

References

Robert L Brennan. Generalizability theory. *Educational Measurement: Issues and Practice*, 11(4):27–34, 1992.

R Philip Chalmers. Mirt: A multidimensional item response theory package for theREnvironment. *J. Stat. Softw.*, 48(6), 2012. ISSN 1548-7660. doi: 10.18637/jss.v048.i06. URL http://dx.doi.org/10.18637/jss.v048.i06.

R. R. Corbeil and S. R. Searle. Restricted maximum likelihood (reml) estimation of variance components in the mixed model. *Technometrics*, 18(1):31–38, 1976. ISSN 00401706. URL http://www.jstor.org/stable/1267913.

Lee J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, 1951. doi: 10.1007/BF02310555. URL https://doi.org/10.1007/BF02310555.

Robert F. DeVellis. *Scale Development: Theory and Applications*. SAGE Publications, 4 edition, 2016.

Susan E. Embretson and Steven P. Reise. *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, Mahwah, NJ, 2000.

John L. Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185, 1965. doi: 10.1007/BF02289447. URL https://doi.org/10.1007/BF02289447.

Li-tze Hu and Peter M. Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1):1–55, 1999. doi: 10.1080/10705519909540118. URL https://doi.org/10.1080/10705519909540118.

I. Jurenka, M. Kunesch, K. R. McKee, D. Gillick, S. Zhu, S. Wiltberger, S. M. Phal, K. Hermann, D. Kasenberg, A. Bhoopchand, J. Gottweis, V. Mikulik, F. Fagan, A. Novikov, A. Kumar, B. Piot, J. Terzi, C. Wang, C. Elster, others, and V. V. Ramasesh. Towards responsible development of generative ai for education: An evaluation-driven approach, 2024. URL https://arxiv.org/abs/2407.12687.

Ekaterina Kochmar, Kaushal Maurya, Kseniia Petukhova, KV Aditya Srivatsa, Anaïs Tack, and Justin Vasselli. Findings of the BEA 2025 shared task on pedagogical ability assessment of AI-powered tutors. In Ekaterina Kochmar, Bashar Alhafni, Marie Bexte, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Anaïs Tack, Victoria Yaneva, and Zheng Yuan, editors, *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1011–1033, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-270-1. doi: 10.18653/v1/2025.bea-1.77. URL https://aclanthology.org/2025.bea-1.77/.

Ehsan Latif, Vincent Liu, and Xiaoming Zhai. A systematic review of intelligent and robot tutoring systems: evolution, pedagogical design, and ai-driven classification. *Smart Learning Environments*, 13(1):1, 2026. ISSN 2196-7091. doi: 10.1186/s40561-025-00427-9. URL https://doi.org/10.1186/s40561-025-00427-9.

J. Macina, N. Daheim, L. Wang, T. Sinha, M. Kapur, I. Gurevych, and M. Sachan. Opportunities and challenges in neural dialog tutoring. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2357–2372. Association for Computational Linguistics, 2023a. doi: 10.18653/v1/2023.eacl-main.173.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.372. URL https://aclanthology.org/2023.findings-emnlp.372/.

K. K. Maurya, K. A. Srivatsa, K. Petukhova, and E. Kochmar. Unifying ai tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of llm-powered ai tutors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1234–1251. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.57.

Ulf Olsson. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44(4):443–460, 1979. doi: 10.1007/BF02296207.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

Fumiko Samejima. *Graded Response Model*, pages 85–100. Springer New York, New York, NY, 1997. ISBN 978-1-4757-2691-6. doi: 10.1007/978-1-4757-2691-6_5. URL https://doi.org/10.1007/978-1-4757-2691-6_5.

Albert Satorra and Peter M Bentler. Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye and C. C. Clogg, editors, *Latent variables analysis: Applications for developmental research*, pages 399–419. Sage, Thousand Oaks, CA, 1994.

A. Tack and C. Piech. The ai teacher test: Measuring the pedagogical ability of blender and gpt-3 in educational dialogues. In *Proceedings of the 15th International Conference on Educational Data Mining*, page 522. International Educational Data Mining Society, 2022.

Sang Truong, Yuheng Tu, Michael Hardy, Anka Reuel, Zeyu Tang, Jirayu Burapacheep, Jonathan Perera, Chibuike Uwakwe, Ben Domingue, Nick Haber, and Sanmi Koyejo. Fantastic bugs and where to find them in ai benchmarks, 2025. URL https://arxiv.org/abs/2511.16842.

Rose Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2174–2199, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.120. URL https://aclanthology.org/2024.naacl-long.120/.