It's Hard to Know How Hard It Is: Mapping the Design Space of LLM Item Difficulty Estimation

Anonymous

Anonymous Institution

Abstract. Can LLMs estimate how hard a test question is? Published results disagree wildly—correlations range from near-zero to $r = 0.87$—but each study typically tests one prompt on one model on one dataset. We map the design space systematically: 15 prompts grounded in learning science, 6 models, 3 datasets, and multiple temperature settings (~250 conditions, ~$80 in API calls). Prompts that direct the model to enumerate structural item features—prerequisites, cognitive load, error paths—achieve $\rho = 0.65$–0.69 on open-ended items. However, this advantage attenuates on MCQs: on two independent datasets (DBE-KT22, BEA 2024), prompts achieve $\rho \approx 0.45$–0.58, with simple teacher judgment nearly matching structured prompts. Prompt design is the primary lever; model size, deliberation, and temperature each contribute minimally ($\eta^2 < 0.05$). Methodologically, we find that correlations from small item sets ($n < 50$) are unreliable, and that unbalanced hyper-parameter sweeps produce Simpson's paradox artifacts—we recommend two-stage sequential DOE for future work. LLM difficulty estimates can triage items into difficulty bands at ~$0.05 per item, but the signal is strongest for items whose difficulty varies across knowledge domains.

Keywords: item difficulty estimation · LLM evaluation · design of experiments · psychometrics · prompt engineering

1 Introduction

A well-designed test gives each student questions at the right level of challenge—hard enough to reveal what they know, easy enough to avoid frustration. Getting this right depends on knowing how difficult each question is before it reaches a student. In practice, this means field-testing items with hundreds or thousands of real students and computing how many answer correctly. In classical test theory, this proportion is called p-correct; item response theory refines it with a latent difficulty parameter b that accounts for examinee ability [8]. Either way, calibrating each item requires collecting responses from 200 to 1,000+ students depending on the IRT model [5], making difficulty estimation one of the most resource-intensive steps in assessment development.

That cost is becoming a bottleneck. LLMs can now generate assessment items at scale [15], but generated items arrive without calibrated difficulty estimates.

Without such estimates, items cannot be assembled into well-targeted tests, used in adaptive systems, or compared to existing item banks. The question is whether LLMs can also estimate difficulty—replacing or supplementing the expensive field-testing step.

The task is harder than it appears. LLMs are trained on expert-written text and tend to find most items trivial; Li, Chen et al. [13] find that high model performance paradoxically impedes difficulty estimation—models that solve items easily cannot perceive what makes them hard for humans. This "curse of knowledge" creates a systematic bias: models overestimate student ability and compress difficulty estimates toward the easy end of the scale. Expert teacher judgment, by comparison, achieves median correlations of $r = 0.66$ [11] to mean $r = 0.63$ [19] with actual student performance—a useful reference point, though these studies measure judgment of individual students rather than aggregate item difficulty.

Recent work reports correlations ranging from $r \approx 0$ for direct estimation [1] to $r = 0.87$ when LLM-extracted features are combined with gradient boosting [17]. A systematic review of 37 papers on text-based difficulty prediction found rapid growth but no consensus on methods, and noted that fine-tuned small language models (BERT, RoBERTa) outperform LLMs on structured prediction tasks [16,14]. The first shared task on automated difficulty prediction (BEA 2024) [22] found that best results only marginally beat baselines on 667 medical items. Li, Chen et al. [13] evaluated 20+ models across multiple assessment domains and found that models converge on a shared "machine consensus" that systematically diverges from human difficulty perception.

A key limitation is that each study typically tests one model with one prompt on one dataset. When results disagree, it is unclear whether the discrepancy reflects model choice, prompt design, item properties, or student population. We address this by mapping the design space rather than testing any single configuration.

Our optimization target is **rank-order agreement** between LLM-predicted difficulty and empirical p-correct, measured by Spearman's ρ. We choose rank correlation because practical applications—item sequencing, adaptive test assembly, item bank stratification—depend on difficulty *ordering* rather than exact calibration. Each experimental condition prompts an LLM to estimate the proportion of students who would answer each item correctly; we then correlate these estimates against observed p-correct from real student responses.

2 Related Work

We organize prior work by approach and extract testable claims from each.

Direct LLM Estimation. Razavi & Powers [17] report $r = 0.83$ (math) and $r = 0.81$ (reading) on K–5 items using GPT-4o. However, Acquaye et al. [1] find $r \approx 0$ for direct estimation on NAEP items, and Li, Chen et al. [13] find average $\rho = 0.28$ across 20+ models on four assessment domains. *Testable claim: direct* *estimation produces meaningful correlations ($\rho > 0.30$) on curriculum-aligned items.*

Feature Extraction + ML. Razavi & Powers [17] achieve their strongest results ($r = 0.87$) by extracting features from LLMs and training gradient-boosted models. *Testable claim: LLM-extracted features combined with ML outperform direct estimation.*

Simulated Classrooms. Acquaye et al. [1] simulate classrooms of LLM students at multiple ability levels, achieving $r = 0.75$–0.82 on NAEP items. Counter-intuitively, weaker math models predicted difficulty better than stronger ones. However, Li, Chen et al. [13] find significant misalignment between AI and human difficulty perception. *Testable claim: simulation outperforms direct estimation.*

Model Uncertainty. Zotos et al. [23] use LLM uncertainty features (first-token probability, choice-order sensitivity) to predict difficulty. *Testable claim: items the model finds uncertain are items students find difficult.*

Reasoning Augmentation. Feng et al. [9] report up to 28% MSE reduction when generating reasoning before predicting difficulty. However, Li, Jiao et al. [14] find minimal improvements from chain-of-thought prompting for fine-tuned models. *Testable claim: structured deliberation improves difficulty estimation.*

Cognitive Item Models. A separate tradition predicts difficulty from structural item features. Embretson's [7] cognitive design system counts processing steps and working memory demands. Gorin [10] manipulates item features to generate items at target difficulties. The KLI framework [12] posits that the number and nature of knowledge components drive learning difficulty. Recent work has used LLMs for KC tagging [15], but not to operationalize cognitive item models as difficulty estimation prompts. *Testable claim: prompts grounded in cognitive item modeling outperform atheoretical prompts.*

3 Hypotheses

We derive hypotheses from three sources: replication of prior claims, predictions from learning science theory, and exploratory questions about the design space. Table 1 lists all 13 hypotheses with their sources and predictions.

4 Method

4.1 Datasets

SmartPaper (India, Open-ended)—Primary. 140 open-ended questions from Indian state assessments across four subjects (English, Mathematics, Science, Social Science), Grades 6–8, with 728,000+ responses. Ground truth: classical p-correct (proportion scoring full marks). Difficulty range: 0.04–0.83, mean 0.29.

Table 1: Hypotheses tested in this study. Verdicts: ✓ = supported, × = rejected, ∼ = partial/mixed.

#	Hypothesis	Source/Theory	Prediction	Verdict
From Prior Work (Replication)				
H1	Direct estimation yields $\rho > 0.30$	Razavi & Powers	$\rho > 0.30$	✓ .56
H2	Simulation beats direct	Acquaye et al.	$\rho_{\text{sim}} > \rho_{\text{direct}}$	× .19<.56
H3	Larger models do better	Scaling expectation	Larger → higher ρ	∼ mixed
From Learning Science Theory				
H4	Item analysis beats direct	Embretson (1998)	$\rho_{\text{analysis}} > \rho_{\text{direct}}$	✓ .69>.56
H5	Prerequisite counting works	KLI (Koedinger)	More prereqs → harder	✓ .69
H6	Cognitive load counting works	CLT (Sweller)	More elements → harder	✓ .67
H7	Buggy rules analysis works	Brown & Burton	More error paths → harder	✓ .66
Exploratory (Design Space)				
H8	Temperature enhances predictions	Stochastic diversity	Higher temp → higher ρ	× η^2=.001
H9	Multi-sample averaging helps	Wisdom of crowds	$\rho_{\text{3-rep}} > \rho_{\text{1-rep}}$	∼ +.03
H10	Prompt is primary lever	Design space	Prompt > model + temp	∼ dataset-dep.
H11	Analysis needs capable models	Capacity limits	Analysis × capability	✓ interaction
H12	Signal transfers across datasets	Generalizability	Signal on D1 → D2	✓ attenuates

Table 2: Experimental stages.

Stage	Purpose	Dataset	Design
Screening	Map prompt space	SmartPaper (140)	15 prompts × 2–5 temps × 3 reps
Model survey	Model generalization	SmartPaper (140)	6 models × 3 prompts × 3 reps
Confirmation	Cross-dataset transfer	DBE-KT22 (168)	3 prompts × 2 temps × 3 reps
Validation	Published benchmark	BEA 2024 (595)	3 prompts × 1 temp × 3 reps

DBE-KT22 (South Africa, MCQ)—Confirmation. 168 undergraduate computer science MCQs spanning 27 knowledge components, administered to 1,300+ students. Ground truth: classical p-correct. Differs from SmartPaper in format, domain, and population.

BEA 2024 Shared Task (USA, USMLE MCQ)—Validation. 595 text-only items from the BEA 2024 automated difficulty prediction shared task [22], comprising USMLE Steps 1–3 questions (667 total items minus 72 requiring images). Ground truth: IRT difficulty parameters from operational administrations. This benchmark allows direct comparison with other difficulty estimation approaches.

4.2 Design Space

4.3 Prompts

All 15 prompts share a common output: the LLM predicts what proportion of students would answer each item correctly. They differ along two binary dimen-

Table 3: Prompt taxonomy with factor coding. Item = analyzes item structure; Pop. = models student population. All 15 prompts except synthetic_students were tested in screening.

Prompt	Item	Pop.	Theory/Prior Work
teacher			PCK (Shulman, 1986)
verbalized_sampling			Wisdom of crowds
familiarity_gradient	✓		Transfer distance
contrastive	✓		Comparative judgment
prerequisite_chain	✓		KC theory (Koedinger et al.)
cognitive_load	✓		CLT (Sweller, 1988)
cognitive_profile	✓		CLT + profiling
devil_advocate		✓	Debiasing heuristics
teacher_decomposed		✓	IRT stratification
classroom_sim		✓	Acquaye et al. (2025)
imagine_classroom		✓	Imagery-based reasoning
synthetic_students*		✓	Student simulation
error_analysis	✓	✓	Error analysis tradition
error_affordance	✓	✓	Brown & Burton (1978)
buggy_rules	✓	✓	Brown & Burton (1978)
misconception_holistic	✓	✓	Chi et al. (1994)

*Tested in model survey only ($\rho = 0.19$, worst performer).

sions: whether the prompt directs the model to *analyze structural item features* (prerequisites, steps, cognitive demands) before estimating, and whether it directs the model to *model the student population* (reason about proficiency levels, struggles). Table 3 lists all 15 prompts. All are available in the project repository.

4.4 Models

Six models spanning 8B to frontier-scale: Gemini 3 Flash (screening model), GPT-4o, Llama-3.3-70B, Gemma-3-27B, Llama-4-Scout (17B active/109B MoE), and Llama-3.1-8B. Qwen3-32B was tested but excluded due to low parse rates (<50% valid responses).

4.5 Statistical Approach

Our primary metric is Spearman's ρ between LLM-predicted and observed p-correct. We report 95% bootstrap confidence intervals (10,000 resamples) for primary results and MAE as a calibration measure. Each condition uses 3 replications; reported ρ is the averaged-prediction correlation (mean prediction across reps correlated with ground truth).

Table 4: Prompt screening results (SmartPaper, $n = 140$), ranked by best ρ. CIs shown for prompts advanced to model survey.

Prompt	Item	Pop.	Best ρ	Temp	95% CI	MAE
prerequisite_chain	✓		**0.686**	0.5	[.576,.775]	.155
cognitive_load	✓		**0.673**	2.0	[.550,.766]	.190
buggy_rules	✓	✓	**0.655**	1.0	[.532,.752]	.117
misconception_holistic	✓	✓	0.636	2.0	—	.204
error_analysis	✓	✓	0.596	2.0	—	.121
devil_advocate		✓	0.596	1.0	—	.098
cognitive_profile	✓		0.586	1.0	[.456,.693]	.214
contrastive	✓		0.584	1.0	—	.123
classroom_sim		✓	0.562	2.0	—	.240
teacher			0.555	1.0	[.422,.664]	.439

5 Results

We present results in five parts: the headline finding on item analysis prompts, the mechanism behind it, factors that matter less than expected, generalization across three datasets, and practical implications.

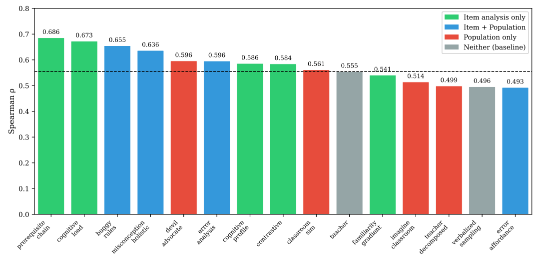
5.1 Item Analysis Prompts Achieve $\rho = 0.65$–0.69

Fifteen prompts were screened on SmartPaper (140 open-ended items) using Gemini 3 Flash at 2–5 temperature settings with 3 replications per condition. The top three prompts—all directing the model to analyze structural item features before estimating difficulty—achieved $\rho = 0.65$–0.69 (Table 4, Figure 1a): **prerequisite_chain** ($\rho = 0.686$) enumerates knowledge prerequisites before estimating; **cognitive_load** ($\rho = 0.673$) counts interacting memory elements; and **buggy_rules** ($\rho = 0.655$) identifies procedural error paths. For context, meta-analyses report that experienced teachers can predict individual student performance with $r = 0.63$–0.66 [11,19]. The constructs differ substantially—teachers judge individual students while we predict aggregate item difficulty—so direct comparison is inappropriate, but the magnitude suggests LLM estimates capture meaningful difficulty information.

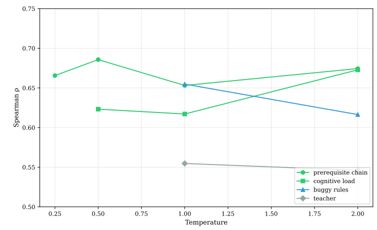
5.2 The Mechanism: Item Analysis, Not Population Modeling

Grouping prompts by their 2×2 factor membership (Table 3) reveals a clear pattern (Table **??**): item analysis is the primary driver. Prompts that model student populations without analyzing item structure perform *worse* than baseline.

Grouping by 2×2 factor membership: item analysis only achieves mean $\rho = 0.61$ (4 prompts), item + population $\rho = 0.59$ (5 prompts), neither (baseline) $\rho = 0.52$ (3 prompts), and population only $\rho = 0.47$ (4 prompts). This parallels Embretson's [7] cognitive design system, where item difficulty is predicted from



(a) Prompt screening. Color = factor membership (green = item only, blue = item + pop, red = pop only, grey = neither). Dashed = teacher baseline.



(b) Temperature effects. Most prompts show minimal sensitivity ($\Delta\rho < 0.05$).

Fig. 1: (a) Prompt screening results on SmartPaper ($n = 140$). (b) Temperature sensitivity for selected prompts.

counts of processing steps and working memory demands. The LLM operationalizes similar constructs when directed to enumerate prerequisites or interacting elements.

Population modeling alone is unreliable. Explicit student simulation (synthetic_students: $\rho = 0.19$) performs worst, consistent with Li, Chen et al.'s [13] finding that proficiency simulation does not reliably improve difficulty estimation.

5.3 What Doesn't Matter

Three factors had minimal effects (each unique $\eta^2 < 0.05$):

Temperature. Most prompts show $\Delta\rho < 0.05$ across temperatures 0.5–2.0 (Figure 1b). The exception: prerequisite_chain peaks at $t = 0.5$ while cognitive_load peaks at $t = 2.0$.

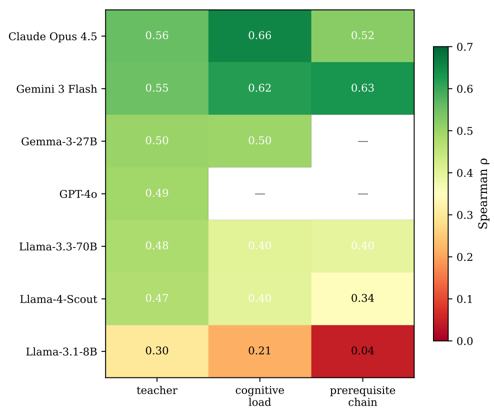
Model size. GPT-4o ($\rho = 0.49$) underperforms Gemini 3 Flash ($\rho = 0.55$) and matches Gemma-3-27B. Llama-3.1-8B struggles ($\rho = 0.30$), but scaling alone does not predict success (Table 5).

Deliberation. Undirected chain-of-thought reasoning degrades performance. The structured prompts that work use *directed* analysis—they specify exactly what to enumerate (prerequisites, interacting elements, error paths). Undirected "think step by step" produces unfocused reasoning that drifts toward solving the item.

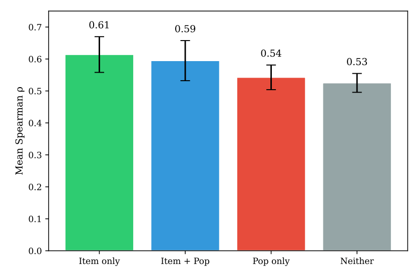
Surface features. Text length correlates $\rho = -0.44$ with difficulty on SmartPaper (shorter questions are harder). Controlling for text length, partial $\rho = 0.59$–0.66 for item analysis prompts—the correlation barely changes ($\Delta\rho < 0.04$), confirming that LLM predictions are not merely proxies for question length.

Table 5: Model survey (ρ, SmartPaper, $t = 1.0$). "—" indicates condition not run or <50 valid responses.

Model	Params	teacher	cog_load	prereq
Gemini 3 Flash	—	0.55	0.62	**0.63**
Gemma-3-27B	27B	0.50	0.50	—
GPT-4o	—	0.49	—	—
Llama-3.3-70B	70B	0.48	0.40	0.40
Llama-4-Scout	17B/109B	0.47	0.40	0.35
Llama-3.1-8B	8B	0.30	0.21	0.04



(a) Model survey. Frontier models achieve $\rho > 0.50$; Llama-8B fails on structured prompts.



(b) Factor analysis. Item analysis is the primary driver; population modeling alone underperforms baseline.

Fig. 2: (a) Model $\times$ prompt heatmap on SmartPaper. (b) Mean ρ by 2×2 factor membership with 95% CIs.

One interaction matters: item analysis prompts help capable models but hurt weak ones. On Gemini, prerequisite_chain gains +0.08 over teacher; on Llama-8B, it loses −0.26.

5.4 Generalization Across Datasets

We validated on two additional datasets: DBE-KT22 (168 undergraduate CS MCQs) and the BEA 2024 shared task (595 USMLE medical MCQs with IRT difficulty). Table 6 shows results.

The signal transfers: all prompts achieve significant correlations on all three datasets. However, the item analysis advantage attenuates on MCQs. On SmartPaper, prerequisite_chain beats teacher by +0.13; on DBE-KT22, buggy_rules achieves the highest $\rho = 0.58$ but the gap over teacher is only +0.06; on BEA, all three prompts are equivalent ($\rho \approx 0.45$).

Table 6: Cross-dataset generalization (Gemini 3 Flash, best temperature per dataset).

Dataset	n	teacher	prereq	buggy	Best	95% CI
SmartPaper (open-ended)	140	0.56	**0.69**	0.66		[.58,.78]
DBE-KT22 (CS MCQ)	168	0.52	0.53	**0.58**		[.46,.68]
BEA 2024 (USMLE MCQ)	595	0.45	0.44	**0.45**		[.39,.52]

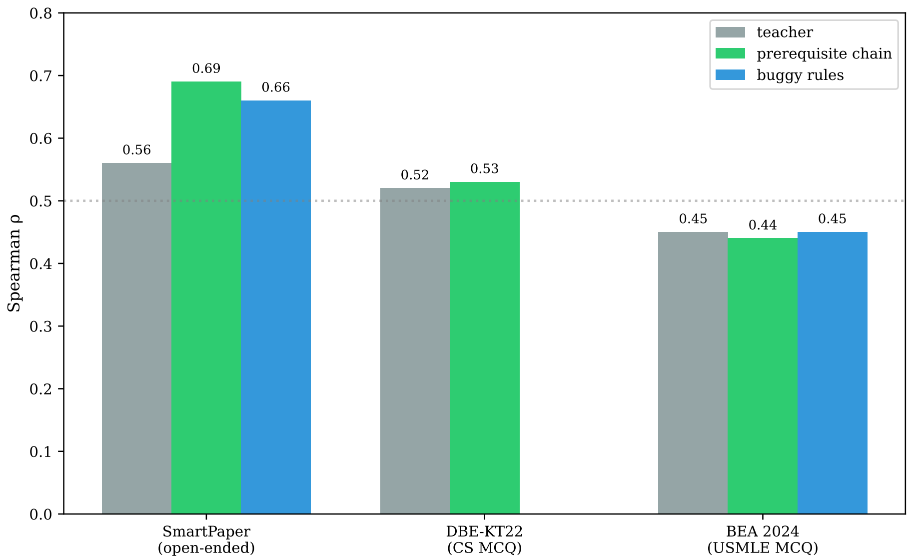


Fig. 3: Cross-dataset generalization. The item-analysis advantage (prerequisite_chain, green) is strongest on open-ended items and attenuates on MCQs, where simple teacher judgment (grey) nearly matches structured prompts.

This suggests a **structural predictability hypothesis**: LLMs estimate difficulty well when it varies across knowledge domains (open-ended items spanning topics), but the advantage of counting prompts diminishes when difficulty is driven by distractor design within a domain. Simple teacher judgment is equally effective for MCQs.

5.5 Practical Implications

At approximately \$0.05 per item (3 replications × Gemini 3 Flash pricing), LLM difficulty estimation is orders of magnitude cheaper than field testing. Rankings transfer across datasets; absolute calibration does not—models overestimate student ability on SmartPaper (MAE≈0.15–0.20) but the rank-order is preserved.

For practitioners: use a capable model (Gemini, GPT-4o), prompt it to analyze structural item features, and average 3+ replications. For standard MCQs, simple teacher judgment suffices.

6 Discussion

Directed vs. Undirected Analysis. The item analysis prompts that work use *directed* analysis—they specify exactly what to enumerate. This contrasts with *undirected* chain-of-thought ("think step by step"), which produces unfocused reasoning that drifts toward solving the item rather than analyzing its structural difficulty. Directed analysis channels implicit pedagogical content knowledge productively; undirected deliberation disrupts it.

Structural Predictability. Our three-dataset comparison suggests a structural predictability pattern: LLMs estimate difficulty well when it varies across knowledge domains (SmartPaper: $\rho = 0.69$), but the item-analysis advantage diminishes on MCQs where difficulty may be driven by distractor design (DBE-KT22: $\rho = 0.52$–0.58; BEA: $\rho = 0.45$, all prompts equivalent).

We conjecture that items whose difficulty is driven by structural features (prerequisite count, cognitive load, error paths) are predictable by LLMs, while items whose difficulty is driven by familiarity, exposure, or subtle distractor plausibility are not—because this information is absent from the item text. Testing this hypothesis would require datasets with item-level annotations of difficulty sources.

Methodological Recommendations. Our design-space exploration surfaced three lessons. (1) *Beware Simpson's paradox in unbalanced designs.* We initially swept temperature only for promising prompts; a marginal ANOVA showed temperature as dominant ($\eta^2 = 0.74$), but this was an artifact—low temperatures were confounded with high-performing prompts. When restricted to balanced conditions, temperature explained $< 1\%$ of variance. (2) *Use two-stage sequential DOE.* Screen all prompts cheaply (1 rep, single temp) to eliminate the bottom half, then run a balanced factorial on survivors. (3) *Validate on $\geq$100 items.* Our 20-item probe yielded $\rho = 0.46$; the full 140-item set achieved $\rho = 0.55$. Bootstrap analysis showed only 2.5% of 20-item subsets would produce $\rho \geq 0.50$. Many published results use <50 items—treat such samples as exploratory only.

Limitations. Three limitations warrant mention. First, all 15 prompts were screened on a single model (Gemini 3 Flash); the model survey shows effects generalize but with diminished magnitude. Second, calibration does not transfer—models overestimate ability on SmartPaper but rankings are preserved, so population-specific adjustment would be needed for absolute predictions. Third, we do not compare to fine-tuned models; fine-tuned BERT/RoBERTa can outperform zero-shot LLMs when training data is available [14], but our approach targets the zero-shot case where items arrive without calibration data.

7 Conclusion

We mapped the design space for LLM item difficulty estimation across 15 prompts, 6 models, and 3 datasets (~250 conditions, ~$80 in API calls). Item analysis prompts that enumerate structural features (prerequisites, cognitive load, error paths) achieve $\rho = 0.65$–0.69 on open-ended items, but this advantage attenuates on MCQs ($\rho \approx 0.45$–0.58), where simple teacher judgment nearly matches. Prompt design is the primary lever; model size, temperature, and deliberation each contribute $\leq 0.05\ \rho$. Rankings transfer across datasets but calibration does not—at ~$0.05 per item, LLM estimates can triage items into difficulty bands, but absolute predictions require population-specific adjustment. Methodologically, correlations from <50 items are unreliable, and unbalanced hyperparameter sweeps produce Simpson's paradox artifacts; we recommend two-stage sequential DOE.

Ethics Statement. This study uses anonymized, aggregated student response data. SmartPaper data consists of item-level statistics without individual student identifiers. DBE-KT22 and BEA 2024 are publicly available research datasets. No personally identifiable information was accessed or processed.

Data and Code Availability. Prompts and analysis code are available in the project repository. SmartPaper data are available upon request from the originating organization. DBE-KT22 is publicly available [6]. BEA 2024 data are available through the shared task organizers.

References

1. Acquaye, C., Huang, Y.T., Carpuat, M., Rudinger, R.: Take out your calculators: Estimating the real difficulty of question items with LLM student simulations. arXiv:2601.09953 (2025)
2. Benedetto, L., et al.: A survey on recent approaches to question difficulty estimation from text. ACM Computing Surveys **56**(5), 1–37 (2023)
3. Brown, J.S., Burton, R.R.: Diagnostic models for procedural bugs in basic mathematical skills. Cognitive Science **2**(2), 155–192 (1978)
4. Chi, M.T.H., Slotta, J.D., de Leeuw, N.: From things to processes: A theory of conceptual change. Learning and Instruction **4**(1), 27–43 (1994)
5. de Ayala, R.J.: The Theory and Practice of Item Response Theory. Guilford Press (2009)
6. Du Plessis, C., van Vuuren, J.: DBE-KT22: A Knowledge Tracing Dataset from South African Secondary Schools. Zenodo (2022). `https://doi.org/10.5281/zenodo.7096666`
7. Embretson, S.E.: A cognitive design system approach to generating valid tests. Psychological Methods **3**(3), 380–396 (1998)
8. Embretson, S.E., Reise, S.P.: Item Response Theory for Psychologists. Lawrence Erlbaum (2000)
9. Feng, W., Tran, P., Sireci, S.G., Lan, A.: Reasoning and sampling-augmented MCQ difficulty prediction via LLMs. arXiv:2503.08551 (2025)

10. Gorin, J.S.: Manipulating processing difficulty of reading comprehension questions. J. Educational Measurement **42**(4), 351–373 (2005)
11. Hoge, R.D., Coladarci, T.: Teacher-based judgments of academic achievement: A review. Review of Educational Research **59**(3), 297–313 (1989)
12. Koedinger, K.R., Corbett, A.T., Perfetti, C.: The Knowledge-Learning-Instruction framework. Cognitive Science **36**(5), 757–798 (2012)
13. Li, M., Chen, H., Xiao, Y., et al.: Can LLMs estimate student struggles? arXiv:2512.18880 (2025)
14. Li, M., Jiao, H., Zhou, T., et al.: Item difficulty modeling using fine-tuned small and large language models. Educ. and Psych. Measurement **85**(6), 1065–1090 (2025)
15. Moore, S., Schmucker, R., Mitchell, T., Stamper, J.: Automated generation and tagging of knowledge components from multiple-choice questions. In: L@S 2024, pp. 122–133 (2024)
16. Peters, S., et al.: Text-based approaches to item difficulty modeling: A systematic review. arXiv:2509.23486 (2025)
17. Razavi, P., Powers, S.J.: Estimating item difficulty using large language models and tree-based ML. arXiv:2504.08804 (2025)
18. Scarlatos, A., et al.: SMART: Simulated students aligned with IRT for question difficulty prediction. In: EMNLP 2025, pp. 25071–25094 (2025)
19. Südkamp, A., Kaiser, J., Möller, J.: Accuracy of teachers' judgments: A meta-analysis. J. Educational Psychology **104**(3), 743–763 (2012)
20. Sweller, J.: Cognitive load during problem solving. Cognitive Science **12**(2), 257–285 (1988)
21. Thurstone, L.L.: A law of comparative judgment. Psychological Review **34**(4), 273–286 (1927)
22. Yaneva, V., et al.: Findings from the first shared task on automated prediction of difficulty. In: BEA 2024, pp. 470–482 (2024)
23. Zotos, L., van Rijn, H., Nissim, M.: Are you doubtful? Exploring model uncertainty for difficulty estimation. In: EDM 2025, pp. 77–89 (2025)