

Does Item Response Theory Generalize Across Languages? Evidence from Cross-Lingual LLM Benchmarks

Ruchira Dhar[†] Qiwei Peng[†] Anders Søgaard[†]

[†]Department of Computer Science, University of Copenhagen

Abstract

The application of Item Response Theory (IRT) for data selection in benchmarks has mostly been limited to monolingual settings. In this work, we ask: how does IRT generalize when the same benchmark is translated across languages? We conduct experiments on Multilingual GSM and Global MMLU, covering over 100 language models and 11 languages. Fitting 1PL, 2PL, and 3PL IRT models per language, we assess cross-lingual consistency of parameters and whether IRT-filtered subsets preserve model rankings across languages: item *difficulty* emerges as the most stable parameter followed by *guessing* and then *discrimination*. A differential item analysis also shows that lack of stability can be tracked to specific items shifting under translation across languages. Our findings suggest IRT-based selection in multilingual settings must be both parameter-aware and item-aware, validating model choice and screening individual items before assuming cross-lingual generalization.

1 Introduction

Item Response Theory (IRT), a psychometric framework that characterizes each item in a test with parameters estimated from response patterns across many subjects, has its origins in clinical psychology (Harvey and Hammer, 1999) and educational testing (Yen and Fitzpatrick, 2006; Lord, 2012). In empirical NLP research, IRT has been used as a tool for selecting data points by filtering test-sets for model evaluation since not all evaluations examples turn out to be equally informative (Lalor et al., 2016; Natesan et al., 2016; Lalor et al., 2019; Rodriguez et al., 2022). More recently, it has also been shown that fitting IRT models to LLM responses, we can select data points while preserving ranking validity of the full benchmark (Hofmann et al., 2025; Zhou et al., 2026). However, the vast majority of this work has been conducted in monolingual settings, primarily English

(see Appendix A for related work). Today a key focus of NLP is cross-lingual evaluation (Hu et al., 2020; Doddapaneni et al., 2025; Chollampatt et al., 2025) where translation is a popular strategy for constructing multilingual benchmarks (Shi et al., 2023; Singh et al., 2025). This raises a natural but largely untested concern: how does IRT generalize cross-lingually when the same benchmark is translated across languages? To our knowledge, this is the first systematic study to attempt to address this.

We evaluate 102 LMs on MGSM (Shi et al., 2023) (11 languages giving 11 slices) and 103 LMs on Global MMLU (Singh et al., 2025) (7 languages, 6 domains giving 42 slices), yielding over 568,000 item-response pairs across 53 benchmark slices. We fit three IRT variants with incrementally more parameters and increasing complexity: 1PL (*difficulty* only), 2PL (adds *discrimination*), and 3PL (adds *guessing*), independently to each benchmark slice, yielding 159 fitted models. We ask three questions: 1) are IRT parameters consistent across translated versions of a benchmark; 2) do IRT-filtered subsets preserve model rankings across languages?; and 3) can cross-lingual instability in IRT parameters be traced to specific items rather than treated as a language-wide effect? Our results indicate that item *difficulty* shows highest concordance across languages for all IRT models, and *guessing* performs better than *discrimination*. This parameter hierarchy directly determines subset validity: 1PL is the most reliable cross-lingual variant. A differential item analysis also confirms this instability is item-level and not language-wide effects. We further observe that for MMLU, scientific domains are consistently more concordant across languages as compared to non-science domains.

Taken together, our results indicate that IRT does not generalise uniformly across languages: which parameters and subsets transfer depends on the IRT model, the individual item, and the domain of evaluation.

2 Methodology

2.1 LLM Evaluations

We use two multilingual translated benchmarks: Multilingual GSM (MGSM) (Shi et al., 2023) across 11 languages (bn, de, en, es, fr, ja, ru, sw, te, th, zh), yielding **250** items per language and **2,750** items per model; Global MMLU (Singh et al., 2025) across 7 languages (de, en, es, fr, ja, sw, zh) and 6 subject domains (*business, humanities, medical, other, social sciences, stem*), totaling **400** items per language and **2,800** items per model under the standard 7-language configuration. For models, we use the HuggingFace Open LLM Leaderboard to select the top instruction-tuned models under 20 billion parameters that are compatible with the lm-eval-harness framework (Gao et al., 2024), yielding 102 and 103 compatible models for MGSM and Global MMLU respectively. In total, our corpus comprises **568,900** evaluated item–response pairs, each consisting of the model’s generated response to a benchmark question and a binary correctness label (implementation details in Appendix B). These item–response pairs are used for IRT model training and for all downstream analysis¹.

2.2 IRT Model Training

The response data, consisting of binary correctness indicators $X_{ij} \in \{0, 1\}$ representing whether a specific language model i answered a benchmark item j correctly, are organised into one response matrix per language (for MGSM) and one per language–domain combination (for Global MMLU), yielding 53 matrices in total (11 MGSM $\times$ 1 + 7 MMLU languages $\times$ 6 domains). We fit three IRT model variants to each matrix independently (implementation details in Appendix B). For 1PL, we use Stochastic Variational Inference (SVI) as implemented in py-irt (Natesan et al., 2016) on top of Pyro (Bingham et al., 2019). For 2PL and 3PL, we use MAP-based implementation in PyTorch². We fit the following IRT model variants:

- **1PL (Rasch model).** The response probability is $P(X_{ij} = 1) = \sigma(\theta_i - b_j)$, where θ_i is the

ability of model i and b_j is the difficulty of item j .

- **2PL.** The response probability is $P(X_{ij} = 1) = \sigma(a_j(\theta_i - b_j))$, adding a per-item discrimination parameter $a_j > 0$ that scales how sharply the item distinguishes between high- and low-ability models.
- **3PL.** The response probability is $P(X_{ij} = 1) = c_j + (1 - c_j) \sigma(a_j(\theta_i - b_j))$, where $c_j \in [0, 1]$ is a lower asymptote (guessing parameter) that models the probability of a correct response by a minimally-capable model, capturing chance-level or knowledge-independent success on an item.

For 1PL, we use py-irt’s SVI with dropout ($p = 0.5$), vague priors, and learning rate $\eta = 0.1$, training for 2,000 epochs (selected at the point of ELBO stabilisation) and retaining parameters at the minimum-ELBO epoch. For 2PL and 3PL, we use MAP with Adam ($\eta = 0.05$) for 4,000 and 6,000 epochs respectively (selected at the point of MAP objective stabilisation, with longer schedules reflecting increased model complexity), retaining the best-objective state across training.

2.3 IRT Analysis

We conduct three sets of experiments to assess how IRT parameters generalize across language versions of these benchmarks.

Cross-lingual Consistency Analysis (§3.1) examines whether item parameters are stable across languages using four measures³: (i) Kendall’s $W \in [0, 1]$, the coefficient of concordance across all $|\mathcal{L}|$ languages simultaneously ($W = 1$: identical item rankings; $W \approx 0$: chance); (ii) Spearman Correlations $\rho \in [-1, 1]$, the monotonic rank correlation between each language pair, displayed as a $|\mathcal{L}| \times |\mathcal{L}|$ heatmap; (iii) Per-Item Rank Deviation, the standard deviation of each item’s difficulty rank across languages, where high values flag unstable items; (iv) Language Clustering (average linkage on Spearman distance $1 - \rho$) grouping languages by similarity of their IRT profiles, with alignment to accuracy-based clusters quantified via Mantel test.

³All IRT variants achieve monotonically improving within-language fit across majority of slices (3PL > 2PL > 1PL; Appendix I), suggesting that cross-lingual parameter differences in following analyses are not merely artifacts of poor within-language fit.

¹The code is publicly available at [this GitHub Repo](#).

²We retain py-irt’s SVI for 1PL, where SVI and MAP estimates are highly concordant, confirming inference method does not affect results. For 2PL and 3PL, SVI’s prior dominates gradient updates at LLM-scale and MAP avoids this. See Appendix C the full comparison and justification.

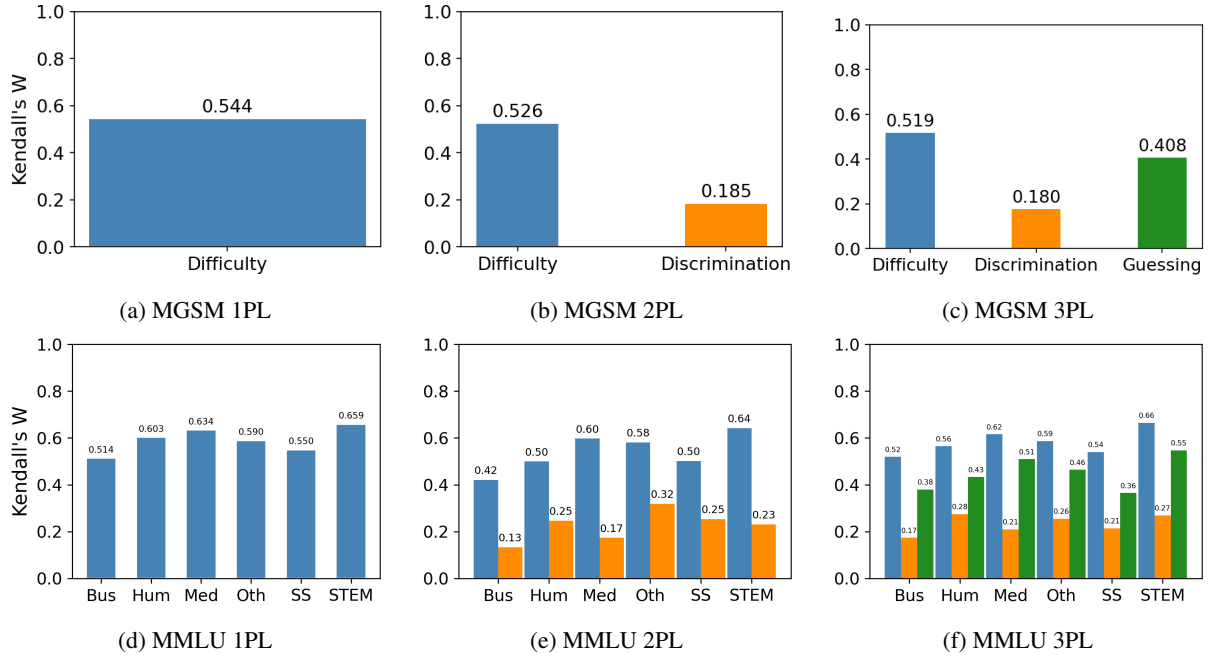


Figure 1: Kendall’s W concordance across languages for MGSM (top row) and MMLU (bottom row) under 1PL, 2PL, and 3PL models. For MGSM, each bar is one parameter type; for MMLU, each bar is one domain (Bus: Business, Hum: Humanities, Med: Medical, SS: Social Sciences, Oth: Others).

Subset Analysis (§3.2) evaluates IRT filtering along two dimensions. (i) Cross-Model IIF Concordance: items are ranked by their Item Information Function (IIF) at $\theta = 0$ under each IRT model variant, and pairwise Jaccard similarity of the top-50% subsets is measured across benchmark slices to assess whether model choice affects item selection; (ii) Subset Ranking Fidelity: Spearman ρ between model accuracy on the IRT-selected subset and on the full benchmark, evaluated within-language and cross-lingually under in-sample and held-out (80/20) model splits.

DIF Analysis (§3.3) asks whether cross-lingual instability in IRT parameters is a language-wide effect or is concentrated in specific items. Thus, for each item we ask: once we account for the fact that some languages are simply harder overall, does this specific item still get noticeably harder or easier when translated? An item where the answer is yes is exhibiting *differential item functioning* (DIF) and language-specific bias in how the item behaves, independent of overall difficulty (Holland and Wainer, 2012; Allalouf et al., 1999). To test this directly, we conduct an anchor-item DIF analysis relative to English. For each language, the bottom quartile of items by cross-lingual rank standard deviation serves as an anchor set to equate the 1PL difficulty scale between English and each

Model	Science		Non-science	
	W	$\bar{\rho}$	W	$\bar{\rho}$
1PL	0.647	0.588	0.564	0.492
2PL	0.620	0.559	0.501	0.420
3PL	0.640	0.581	0.553	0.479

Table 1: Kendall’s W and Mean Pairwise Spearman $\bar{\rho}$ for item *difficulty* across languages, averaged over science (STEM + Medical) and non-science (Business + Humanities + Other + Social Sciences) domains of MMLU, for each IRT model variant.

focal language, avoiding the equal-ability assumption that otherwise blocks a formal DIF test in our setting. For each remaining item we compute the anchor-corrected difficulty shift Δb_j , test significance via Fisher-information standard errors, and control the false discovery rate at $q = 0.05$ with Benjamini–Hochberg.

3 Results

3.1 Cross-Lingual Consistency Analysis

Kendall’s W Concordance Figure 1 shows that difficulty concordance is maintained across model complexities, while discrimination parameters are substantially less cross-lingually stable. Under 1PL, difficulty concordances are: MGSM at $W = 0.544$ and MMLU domains

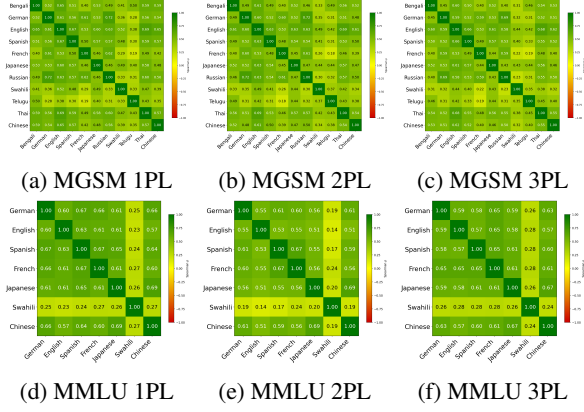


Figure 2: Pairwise Spearman rank correlation of item *difficulty* across languages for 1PL, 2PL, and 3PL IRT models on MGSM (top row) and MMLU (bottom row; each cell is the element-wise average over six domains). Colour encodes $\rho \in [-1, +1]$: red = -1 , yellow = 0 , green = $+1$.

ranging from $W \in [0.514, 0.659]$. Under 2PL, difficulty concordance is maintained: MGSM $W_{\text{diff}} = 0.526$ (barely below 1PL), MMLU $W_{\text{diff}} \in [0.421, 0.642]$. Under 3PL, difficulty concordance is similarly preserved (MGSM $W_{\text{diff}} = 0.519$; MMLU $W_{\text{diff}} \in [0.520, 0.664]$). The discrimination parameter, by contrast, shows substantially lower cross-lingual agreement across both 2PL (MGSM $W_{\text{disc}} = 0.185$; MMLU $W_{\text{disc}} \in [0.134, 0.319]$) and 3PL (MGSM $W_{\text{disc}} = 0.180$; MMLU $W_{\text{disc}} \in [0.174, 0.275]$). The guessing parameter c under 3PL retains concordance on both MGSM ($W = 0.408$) and MMLU ($W \in [0.365, 0.547]$). To rule out that this concordance is inflated by rater redundancy among typologically similar languages, we recompute W after removing the tight cluster of Indo-European languages and find that concordance does not decrease under any IRT variant (details in Appendix E).

A cross-domain analysis (Table 1) of item *difficulty* groups MMLU domains into science (STEM + Medical) and non-science (Business + Humanities + Other + Social Sciences), revealing a persistent science advantage that holds across all model variants. Scientific domains, where mathematical and factual content is more likely to be language-invariant, yield more consistent IRT parameters across languages.

Pairwise Spearman Correlations Figure 2 shows pairwise item *difficulty* and Figure 3 shows pairwise item *discrimination* and *guessing* rank correlations across languages for each IRT variant. To

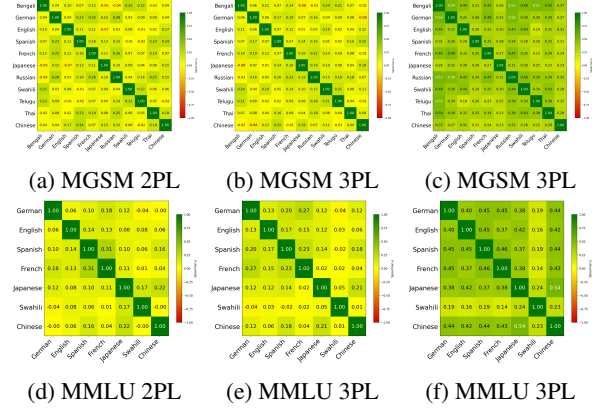


Figure 3: Pairwise Spearman rank correlation of item *discrimination* (a, columns 1–2) and *guessing* (c, column 3) across languages for MGSM (top row) and MMLU (bottom row; element-wise average over six domains). Colour encodes $\rho \in [-1, +1]$: red = -1 , yellow = 0 , green = $+1$.

ensure these correlations reflect cross-lingual differences rather than within-language estimation noise, we conduct a split-half reliability check that refits 2PL/3PL via MAP on disjoint halves of the model pool and confirms strong within-language parameter agreement (Appendix D). Under 1PL, all language pairs show positive monotonic agreement on both MGSM (mean $\rho = 0.498$) and MMLU (mean $\rho = 0.524$), with stronger correlations among Indo-European language pairs (e.g. German–Russian: $\rho = 0.718$; English–Spanish: $\rho = 0.670$) and weaker but still positive correlations for distant pairs (e.g. French–Telugu: $\rho = 0.190$; French–Swahili: $\rho = 0.290$). On MMLU, European languages form a tight cluster ($\rho \in [0.60, 0.67]$), while Swahili is a consistent outlier across both benchmarks. Under 2PL, difficulty correlations show a modest reduction on MGSM (mean $\rho = 0.479$) and MMLU (mean $\rho = 0.466$), with positive values across most language pairs, indicating that 2PL preserves the shared difficulty structure rather than absorbing it into language-specific variance. Under 3PL, difficulty correlations remain similarly positive (MGSM mean $\rho = 0.470$; MMLU mean $\rho = 0.513$). The discrimination parameter a is less cross-lingually stable across both 2PL and 3PL as seen in Figure 3 (detailed discussion in Appendix F), indicating it captures language-idiosyncratic signal. The guessing parameter c under 3PL again shows stronger positive correlations across language pairs (mean $\rho = 0.349$ on MGSM; mean $\rho = 0.359$ on MMLU), with the same Swahili isolation pattern persisting, sug-

Model	Param	MGSM	MMLU
		mean [min, max]	mean [min, max]
1PL	diff	48.5 [2.8, 91.8]	14.5 [1.6, 34.3]
2PL	diff	49.5 [4.2, 98.1]	15.5 [1.8, 40.0]
	disc	67.1 [29.4, 95.6]	19.4 [2.6, 47.1]
3PL	diff	49.9 [5.9, 99.4]	14.7 [1.4, 37.6]
	disc	67.1 [19.1, 99.3]	19.6 [2.0, 52.1]
	<i>c</i>	56.3 [6.0, 85.9]	17.0 [1.5, 47.0]

Table 2: Per-item rank standard deviation across languages for each IRT model variant and parameter. MMLU values are pooled across all six domains. Lower values indicate higher cross-lingual stability.

gesting that chance-level correctness is among the most translation-stable properties of an item. The cross-domain analysis (Table 1) shows that scientific domains in MMLU consistently show higher correlation across languages than non-science domains.

Per-Item Rank Standard Deviation While Kendall’s W and Spearman ρ capture aggregate ordering agreement, they mask item-level heterogeneity: per-item rank SD reveals which individual items preserve their difficulty rank across language versions of a benchmark, directly informing which items are most stable across languages in a multilingual compressed benchmark. For MGSM, stability is highly heterogeneous under 1PL, with some items near-fixed across languages and others shifting almost the full rank range (Table 2). Under 2PL and 3PL, difficulty rank SD is essentially unchanged from 1PL (MGSM: 49.5 and 49.9 vs. 48.5; MMLU: 15.5 and 14.7 vs. 14.5), while discrimination rank SD is substantially higher (≈ 67 on MGSM). The guessing parameter c shows intermediate rank stability (MGSM: 56.3). For MMLU, the pattern mirrors MGSM: wide spread under 1PL, with difficulty stability maintained or improved under 2PL/3PL, and discrimination consistently less stable (Figure 14 and Figure 15 in Appendix H).

Language Clustering We cluster languages hierarchically (average linkage on Spearman distance $1 - \rho$) by each IRT parameter profile and compare to accuracy-based clusterings via Mantel test and mean pairwise Spearman $\bar{\rho}$. For *difficulty* (Figure 4), MGSM dendrograms mirror the accuracy clustering: European languages cluster tightly while Telugu and Swahili are most isolated. A Mantel test quantifies alignment between each IRT difficulty clustering and the accuracy cluster-

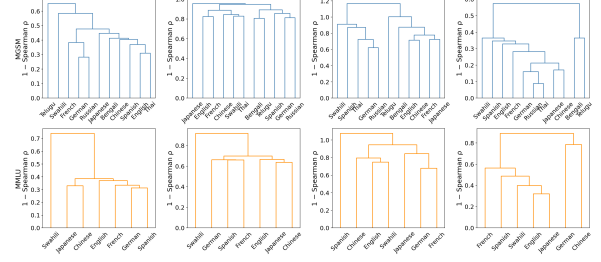


Figure 4: Language clustering by *difficulty* for MGSM (top row) and MMLU (bottom row). Columns 1–3 cluster languages by pairwise Spearman correlation of item difficulty rank profiles under 1PL, 2PL, and 3PL respectively. Column 4 clusters languages by pairwise Spearman correlation of model accuracy vectors.

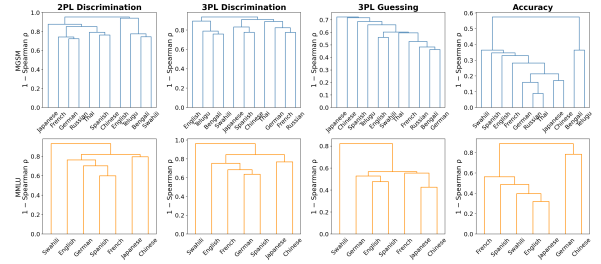


Figure 5: Language clustering by *discrimination* (for 2PL and 3PL), and *guessing* (3PL). Rows: MGSM (top), MMLU (bottom). Columns: 2PL discrimination, 3PL discrimination, 3PL guessing, accuracy (left to right).

ing (MGSM: 1PL $r = 0.63$, $p = 0.001$; 2PL $r = 0.56$, $p = 0.002$; 3PL $r = 0.53$, $p = 0.003$; all MMLU tests non-significant: 1PL $r = -0.15$, $p = 0.751$; 2PL $r = -0.23$, $p = 0.886$; 3PL $r = -0.17$, $p = 0.786$). For MMLU, the accuracy dendrogram is structurally decoupled from IRT difficulty profiles where German and Chinese cluster as nearest neighbours despite typological distance, while Swahili clusters with European languages in accuracy space, suggesting MMLU performance might be governed more by alternate factors like pretraining data distribution than item difficulty structure. For *discrimination* and *guessing* (Figure 5), the same procedure is applied with mean pairwise $\bar{\rho}$ as an additional cross-lingual consistency measure. Discrimination is the least cross-lingually consistent parameter ($\bar{\rho} \approx 0.10$ on both benchmarks), producing near-equidistant clusterings with no interpretable structure; Mantel tests are non-significant except MGSM 2PL discrimination ($r = 0.34$, $p = 0.005$), a signal that disappears under 3PL. Guessing shows substantially higher cross-lingual consistency ($\bar{\rho} \approx 0.35$ on both benchmarks), visible in lower and more differentiated

merge heights, but does not align with typological or accuracy-based groupings (all Mantel tests non-significant).

Thus, *difficulty* is the only parameter whose clustering aligns with accuracy (Mantel tests significant on MGSM, non-significant on MMLU). Neither *discrimination* nor *guessing* aligns with accuracy-based clusters (all Mantel tests non-significant; MGSM 3PL: disc $r = 0.134$, $p = 0.151$; guess $r = 0.074$, $p = 0.341$). However, guessing produces more differentiated dendrogram structure than discrimination: mean pairwise $\rho = 0.349$ vs. 0.098 on MGSM, with lower and more varied merge heights, indicating it captures some consistent cross-lingual signal, though not one that mirrors typological or accuracy-based groupings.

3.2 Subset Analysis

IIF Concordance We examine whether the three IRT variants agree on which items are most informative, a choice with direct consequences for the validity of any resulting cross-lingual subset (within-language fit comparisons are in [Appendix I](#)). Agreement is operationalised via the Item Information Function (IIF) at $\theta = 0$; a sensitivity analysis at $\theta \in \{-1, 0, +1\}$ confirms that the qualitative orderings are robust to this choice ([Appendix J](#)):

$$I_j^{(1PL)} = P_j (1 - P_j), \quad (1)$$

$$I_j^{(2PL)} = a_j^2 P_j (1 - P_j), \quad (2)$$

$$I_j^{(3PL)} = \frac{a_j^2 (P_j - c_j)^2 (1 - P_j)}{(1 - c_j)^2 P_j}, \quad (3)$$

where b_j , a_j , c_j are difficulty, discrimination, and guessing parameters; $P_j^{(1PL)} = \sigma(-b_j)$; $P_j^{(2PL)} = \sigma(a_j(-b_j))$; $P_j^{(3PL)} = c_j + (1 - c_j) \sigma(a_j(-b_j))$. Per slice, items are ranked by I_j and the top 50% retained (avg. 125 items per MGSM slice; 33 per MMLU slice; sensitivity to this threshold in [Appendix K](#)); pairwise agreement is measured by Jaccard similarity $J(A, B) = |A \cap B| / |A \cup B|$, with an expected baseline of $J \approx 0.33$ for two independent random draws ([Figure 6](#)). Across both benchmarks, 1PL–2PL agreement (median $J = 0.40$ for MGSM, $J = 0.53$ for MMLU) and 1PL–3PL (median $J = 0.41$ for MGSM, $J = 0.45$ for MMLU) both exceed the random-selection baseline ($J \approx 0.33$). 2PL–3PL shows the strongest concordance overall (median $J = 0.92$ for MGSM, $J = 0.59$ for MMLU), confirming that 2PL and

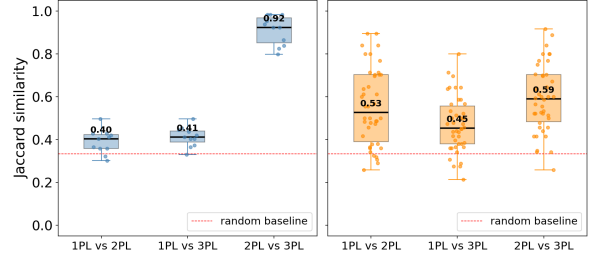


Figure 6: Jaccard similarity of the top-50% items selected by IIF at $\theta = 0$ for each model pair (**left: MGSM; right: MMLU**), across all benchmark slices (11 MGSM languages; 42 MMLU language $\times$ domain combinations). Each dot is one slice; annotated values are medians. Dashed line marks the random-selection baseline ($J \approx 0.33$).

Selection strategy	MGSM		MMLU	
	In-samp.	Held-out	In-samp.	Held-out
1PL (difficulty)	0.982	0.972	1.000	1.000
2PL difficulty only	0.978	0.970	0.932	0.932
2PL (diff+disc)	0.971	0.953	0.878	0.831
3PL difficulty only	0.979	0.972	0.922	0.889
3PL (diff + disc)	0.975	0.950	0.850	0.778
3PL (diff + disc + guess)	0.973	0.951	0.764	0.691

Table 3: Median Spearman ρ (subset vs. full benchmark) under progressively richer IIF selection strategies; the gap between rows isolates that parameter’s marginal contribution to fidelity loss.

3PL agree very strongly with each other on which items are most informative. MMLU shows lower 2PL–3PL concordance, reflecting greater variation across the heterogeneous domain slices.

Thus, switching from 1PL to 2PL or 3PL modifies item selection but does not discard all shared signal, with 2PL and 3PL producing closely related subsets. The low 1PL–2PL/3PL concordance reflects that 1PL selects items by difficulty alone, while 2PL and 3PL IIFs are dominated by the discrimination term, prioritising different, and more language-idiosyncratic, items. The strong 2PL–3PL concordance on MGSM ($J = 0.92$) could reflect near-zero guessing rates on open-ended math items, so the guessing correction in 3PL barely shifts selection from 2PL; the lower concordance on MMLU ($J = 0.59$) confirms that guessing becomes consequential for multiple-choice items where chance-level correctness is non-negligible.

Ranking Fidelity Since IIF concordance shows that IRT model choice determines which items are selected, we now ask whether those selections preserve *model rankings* on the full benchmark. We evaluate four conditions: (i) in-sample

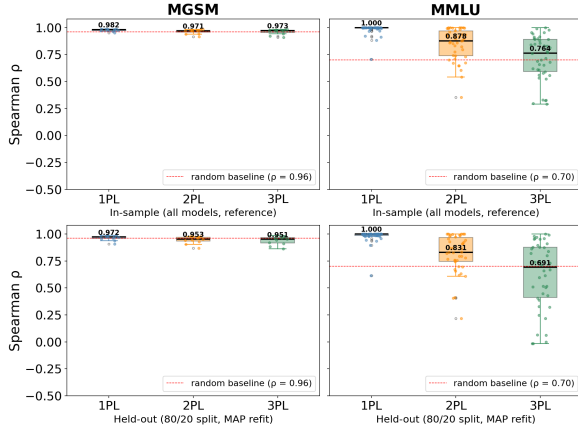


Figure 7: Ranking fidelity (Spearman ρ , subset vs. full benchmark) for MGSM (left) and MMLU (right). **Top row:** in-sample (all models, upper bound). **Bottom row:** held-out (IRT re-fitted on 80% of models, evaluated on remaining 20%). Dashed line marks the random-item baseline ($\rho \approx 0.96$ for MGSM, $\rho \approx 0.7$ for MMLU). 1PL fidelity is stable across both splits.

same-language; (ii) held-out same-language (IRT fitted on 80% of models, evaluated on held-out 20%); (iii) in-sample cross-lingual; and (iv) held-out cross-lingual. The primary metric is Spearman ρ between per-model subset accuracy and full-benchmark accuracy, compared against a 200-draw random-item baseline⁴ (Figure 7). Under 1PL, same-language fidelity is high and stable across both splits: in-sample median $\rho = 0.982$ (MGSM) and 1.000 (MMLU), dropping only marginally on held-out models (0.972 and 1.000), well above the random baselines. On MGSM, 2PL and 3PL also preserve strong held-out fidelity ($\rho = 0.953$ and 0.951), showing that with stable parameter estimation all three variants are viable. On MMLU, fidelity degrades progressively with model complexity: 2PL held-out median $\rho = 0.831$ and 3PL $\rho = 0.691$, with 3PL showing the widest spread and several low-fidelity slices (per-slice distributions and accuracy scatter plots in Appendix L).

To isolate which parameter drives this degradation, Table 3 decomposes fidelity by applying progressively richer IIF formulas, also shown in Figure 10: on MGSM all six strategies yield near-identical fidelity, confirming that difficulty alone drives the signal; on MMLU fidelity drops mono-

⁴A mean-accuracy selection baseline achieves comparable fidelity to 1PL (cross-lingual held-out: MGSM $\rho = 0.954$ vs. 0.949; MMLU $\rho = 0.879$ vs. 0.877), but this does not bear on our research question, which concerns whether IRT *parameters* generalise across languages, not whether IRT outperforms alternatives.

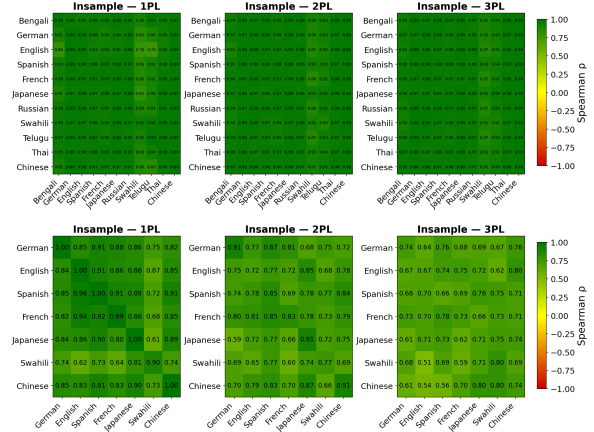


Figure 8: In-sample cross-lingual subset ranking fidelity. Cell (i, j) is Spearman ρ between subset accuracy (items from language i) and full-benchmark accuracy in language j . **Top:** MGSM. **Bottom:** MMLU (avg. over six domains). Columns = 1PL, 2PL, 3PL.

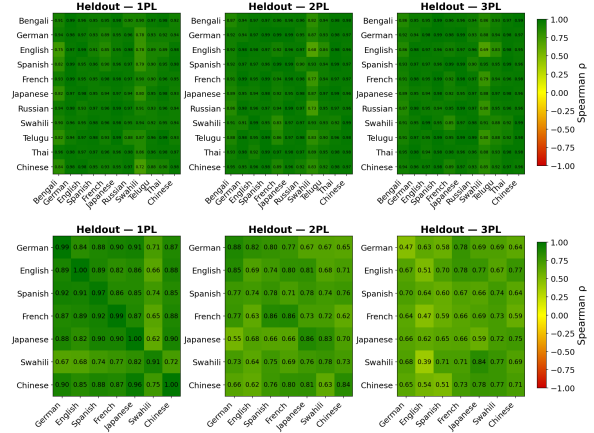


Figure 9: Held-out cross-lingual subset ranking fidelity (IRT re-fitted on 80% of models via MAP, evaluated on 20%). Same layout as Figure 8.

tonically with each parameter added, with discrimination causing the largest single step down. Selecting items by *difficulty alone* preserves high fidelity across both benchmarks and *discrimination* introduces the primary degradation: -0.017 to -0.022 on MGSM and -0.101 to -0.111 on MMLU held-out, consistent with discrimination being language-idiosyncratic. Adding *guessing* further reduces MMLU held-out fidelity by 0.087 ($0.778 \rightarrow 0.691$) while leaving MGSM virtually unchanged ($0.950 \rightarrow 0.951$). Together, within-language fidelity loss is driven primarily by *discrimination* rather than model complexity per se; difficulty-based or 1PL selection remains the safest strategy.

Cross-lingual transfer is shown in Figure 8 (in-

Selection strategy	MGSM		MMLU	
	In-samp.	Held-out	In-samp.	Held-out
1PL (difficulty)	0.968	0.949	0.854	0.877
2PL difficulty only	0.973	0.960	0.789	0.795
2PL (diff+disc)	0.973	0.962	0.765	0.777
3PL difficulty only	0.974	0.961	0.778	0.787
3PL (diff+disc)	0.972	0.958	0.761	0.749
3PL (diff+disc+guess)	0.972	0.959	0.714	0.729

Table 4: Median cross-lingual Spearman ρ (off-diagonal pairs) under progressively richer IIF selection strategies. Items are selected using the source language IRT; fidelity is evaluated in the target language.

Figure 10: Median Spearman ρ under progressively richer IIF selection strategies, isolating each parameter’s marginal contribution to fidelity.

sample) and Figure 9 (heldout), with a parameter-wise decomposition in Table 4. On MGSM, all three variants transfer well (held-out median $\rho = 0.949$ – 0.962), with 2PL and 3PL matching or slightly exceeding 1PL: arithmetic difficulty is translation-invariant and discrimination adds negligible noise. On MMLU, 1PL is the only reliable cross-lingual variant (held-out $\rho = 0.877$); fidelity degrades steadily with model complexity (2PL: 0.777; 3PL: 0.729). The parameter decomposition reveals that switching to 2PL *difficulty alone* already accounts for most of this drop ($0.877 \rightarrow 0.795$), with discrimination contributing a further but smaller step ($0.795 \rightarrow 0.777$). Guessing contributes an additional but smaller -0.020 loss. Inspecting the language-pair breakdown in the figures, this degradation is not uniform: Swahili is consistently the most isolated source and target, mirroring its outlier status in the concordance analysis (§3.1).

3.3 DIF Analysis

DIF effects are consistently strong across both benchmarks (Figure 11). On MGSM, an average of 48.3% of tested items (range: 34–63% across 10 languages) show significant DIF relative to English, with a mean anchor-corrected shift of $|\Delta b| = 1.14$ logits. On MMLU, 31.8% of items show significant DIF on average (range: 21–69% across 6 focal languages), with a mean shift of $|\Delta b| = 1.50$ logits. Pooled across both benchmarks, this corresponds to an average shift in probability of a correct response of roughly 22 percentage points at the item’s own difficulty level, rising to roughly 38 points among items flagged as DIF. Swahili

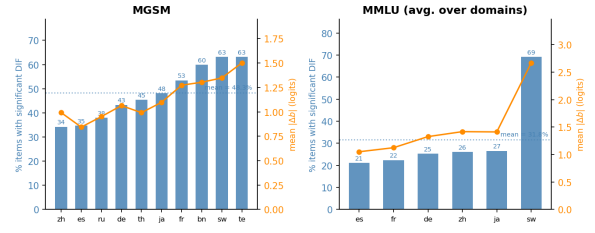


Figure 11: Percentage of items showing significant DIF relative to English (bars, left axis) and mean anchor-corrected difficulty shift $|\Delta b|$ (line, right axis), by language; MMLU averaged over its six domains. Dashed lines mark the benchmark-wide mean.

shows the highest rate on MMLU (69%) and is tied with Telugu for the highest MGSM rate (63%), consistent with its isolation in the clustering results (section 3.1). Full per-language and per-domain results are in Appendix G.

4 Discussion

Cross-lingual Concordance. We find that difficulty is consistently the most stable IRT parameter across languages, and this concordance is maintained when we move to 2PL and 3PL models. Discrimination is substantially less stable across all variants and both benchmarks. Guessing sits between the two. The language clustering results on MGSM reinforce this parameter hierarchy: difficulty profiles from all three models significantly track language typology, while discrimination profiles do not. Additionally, STEM domains are consistently more concordant than humanities and social sciences across all variants, suggesting that reasoning-based items are more likely to be translation-stable than culturally situated ones.

Subset Analysis. Within-language, 1PL subsets preserve model rankings reliably on both benchmarks, and 2PL and 3PL are competitive on MGSM but degrade substantially on MMLU. Cross-lingually, MGSM remains robust across all three IRT variants, while on MMLU 1PL is the only reliable variant. The parameter decomposition in Table 4 shows that most of the fidelity drop happens before discrimination even enters selection: switching from 1PL to 2PL difficulty alone already costs substantial fidelity on MMLU (held-out $\rho: 0.877 \rightarrow 0.795$). Discrimination contributes a further but smaller drop ($0.795 \rightarrow 0.777$). Cross-lingually, therefore, 1PL’s advantage lies not only in avoiding discrimination but in producing inherently more transferable difficulty estimates. Finally,

the near-identical item selections of 2PL and 3PL ($J = 0.92$) on MGSM but not MMLU show that the additional complexity of 3PL does not diversify the selected subset on reasoning-heavy tasks where guessing is negligible.

DIF Analysis. DIF is widespread and highly item-specific: on MGSM an average of 48.3% of items and on MMLU 31.8% of items show significant difficulty shifts relative to English and DIF rates change (highest for Swahili and Telugu on MGSM, Swahili on MMLU), indicating that the same item-level translation pressures accumulate more heavily for languages least related to the source. We restrict this test to *difficulty*: discrimination is already the least cross-lingually stable parameter in §3.1, so no discrimination-invariant anchor set can be identified to equate its scale, and guessing’s lower-asymptote estimates require substantially larger per-item response counts than our model pool provides to support individual-item significance testing. Since difficulty is the most cross-lingually stable of the three parameters and still shows this much item-level heterogeneity once tested directly, it is plausible that discrimination and guessing, already less stable in aggregate, conceal comparable or greater item-level heterogeneity that a suitably powered DIF test would need to confirm. The practical implication is that this instability is item-specific rather than an inherent property of a language, and is in principle addressable through targeted re-translation or removal with item-level auditing.

More concretely, for using IRT-based data selection in benchmark filtering on multilingual translated benchmarks: (i) we must not assume that a subset which works well within one language will generalise to others, i.e., cross-lingual validity should be verified explicitly; (ii) we must account for domain composition, as STEM domains are reliably more cross-lingually stable than non-STEM, and a single overall fidelity check will mask important variation within mixed-domain benchmarks; (iii) a subset can look valid by an aggregate concordance check while still containing individual items that behave inconsistently, so item-level DIF screening should complement aggregate validity checks before reusing a subset across languages. Simpler models that rely only on the most stable parameters will produce more transferable subsets.

5 Conclusion

We set out to test a basic but largely unexamined assumption in multilingual LLM evaluation: whether IRT, an established tool for benchmark filtering, behaves consistently when the same benchmark is translated across languages. The answer is that it does, but only for some parameters, and even for the parameter that transfers best, only for some items rather than all of them. IRT parameters form a cross-lingual stability hierarchy: difficulty transfers reliably, discrimination does not, and guessing sits in between. Even difficulty, the most stable parameter, still shows significant translation-induced shifts in roughly a third to half of items once each language’s overall difficulty is accounted for, showing that cross-lingual instability is not simply a property of a language but of specific items within it.

As translation-based multilingual benchmarks become common, the psychometric properties of translated items cannot be assumed to carry over across languages. Our findings show that the choice of IRT model, the stability of its parameters, and the presence of item-level differences determine whether a compressed benchmark remains valid across the languages it is meant to evaluate. Cross-lingual validity should be demonstrated, not assumed, and we hope this work, combining aggregate concordance checks with item-level auditing, provides a concrete empirical basis for that standard.

Limitations

We discuss some limitations of our work here. *First*, our model set is restricted to instruction-tuned models with fewer than 20 billion parameters, due to compute constraints. This means our IRT parameter estimates and subset fidelity results may not generalize to larger frontier models, whose ability profiles could alter item difficulty and discrimination rankings in ways not captured here.

Second, our experiments are conducted on two benchmarks, MGSM and Global MMLU, which, while covering a broad range of languages and domains, are both derived from English source items via translation. Extending this analysis to benchmarks constructed natively in non-English languages, or to tasks beyond multiple-choice and arithmetic reasoning, remains an important direction for future work. Additionally, the Global MMLU subset used here includes four European

languages (German, English, Spanish, French) out of seven, which may inflate cross-lingual concordance estimates relative to a more typologically diverse sample.

Ethical Considerations

We do not anticipate any risks in our work. Any data used does not contain any personal details or offensive content. In this study, our use of all artifacts is consistent with the license and intended use of all datasets and models. The benchmarks used, [Multilingual GSM \(Shi et al., 2023\)](#) and [Global MMLU \(Singh et al., 2025\)](#) are available on HuggingFace for free use. All large language models used are also open from HuggingFace. The [py-irt library](#) used for 1PL model fitting is open source and freely available under MIT License. Our MAP-based implementation for 2PL and 3PL is implemented in PyTorch and will be released alongside the codebase upon publication.

References

- Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, and 1 others. 2023. Mega: Multilingual evaluation of generative ai. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267.
- Avi Allalouf, Ronald K Hambleton, and Stephen G Sireci. 1999. Identifying the causes of dif in translated verbal items. *Journal of educational measurement*, 36(3):185–198.
- Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. 2019. [Pyro: Deep universal probabilistic programming](#). *J. Mach. Learn. Res.*, 20:28:1–28:6.
- Shamil Chollampatt, Minh Quang Pham, Sathish Reddy Indurthi, and Marco Turchi. 2025. Cross-lingual evaluation of multilingual text generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7766–7777.
- Sumanth Doddapaneni, Mohammed Safi Ur Rahman Khan, Dilip Venkatesh, Raj Dabre, Anoop Kunchukuttan, and Mitesh M Khapra. 2025. Cross-lingual auto evaluation for assessing multilingual llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29297–29329.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. [The language model evaluation harness](#).
- Aletta Grisay, JHAL De Jong, Eveline Gebhardt, Alla Berezner, and Beatrice Halleux-Monseur. 2007. Translation equivalence across pisa countries. *Journal of Applied Measurement*, 8(3):249.
- Ronald K Hambleton, Peter F Merenda, Charles D Spielberger, and 1 others. 2005. Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. *Adapting educational and psychological tests for cross-cultural assessment*, 1:3–38.
- Robert J Harvey and Allen L Hammer. 1999. Item response theory. *The Counseling Psychologist*, 27(3):353–383.
- Valentin Hofmann, David Heineman, Ian Magnusson, Kyle Lo, Jesse Dodge, Maarten Sap, Pang Wei Koh, Chun Wang, Hannaneh Hajishirzi, and Noah A. Smith. 2025. [Fluid language model benchmarking](#). In *Second Conference on Language Modeling*.
- Paul W Holland and Howard Wainer. 2012. *Differential item functioning*. Routledge.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International conference on machine learning*, pages 4411–4421. PMLR.
- Viet Dac Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hiu Mn, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 13171–13189.
- John P Lalor, Hao Wu, and Hong Yu. 2016. Building an evaluation scale using item response theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 648–657.
- John P Lalor, Hao Wu, and Hong Yu. 2019. Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4249–4259.
- Frederic M Lord. 2012. *Applications of item response theory to practical testing problems*. Routledge.

Taciano L Milfont and Ronald Fischer. 2010. Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of psychological research*, 3(1):111–130.

Prathiba Natesan, Ratna Nandakumar, Tom Minka, and Jonathan D Rubright. 2016. Bayesian prior choice in irt estimation using mcmc and variational bayes. *Frontiers in psychology*, 7:1422.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P Lalor, Robin Jia, and Jordan Lee Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change nlp leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503.

Pedro Rodriguez, Phu Mon Htut, John P Lalor, and João Sedoc. 2022. Clustering examples in multi-dataset benchmarks with item response theory. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 100–112.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and 1 others. 2021. Xtreme-r: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. [Language models are multilingual chain-of-thought reasoners](#). In *The Eleventh International Conference on Learning Representations*.

Shivalika Singh, Angelika Romanou, Cl  mentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, and 1 others. 2025. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799.

Wendy M Yen and Anne R Fitzpatrick. 2006. Item response theory. *Educational measurement*, 4:111–153.

Ziyin Zhang, Yikang Liu, Weifang Huang, Junyu Mao, Rui Wang, and Hai Hu. 2024. Mela: Multilingual evaluation of linguistic acceptability. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2658–2674.

Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, and 1 others. 2026. Lost in

benchmarks? rethinking large language model benchmarking with item response theory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 35085–35093.

A Related Work

We discuss some relevant work here that grounds and motivates our research.

IRT for NLP Benchmark Filtering. One of the earliest works in Item Response Theory for NLP by [Lalor et al. \(2016\)](#) applied it to annotation quality estimation to identify uninformative items in crowdsourced datasets. Subsequent work used IRT to characterise the difficulty and discriminating power of NLP test sets ([Lalor et al., 2019](#); [Rodriguez et al., 2021](#)) and to cluster benchmark items by their psychometric profiles ([Rodriguez et al., 2022](#)). The `py-irt` library ([Natesan et al., 2016](#)) made SVI-based IRT estimation tractable for NLP-scale response matrices. Most recently, IRT has been applied directly to LLM benchmarks: [Hofmann et al. \(2025\)](#) showed that IRT-selected subsets substantially reduce evaluation cost while preserving validity, and [Zhou et al. \(2026\)](#) examined ranking fidelity loss under aggressive compression. Our work extends this line to multilingual settings, asking whether item parameters estimated in one language transfer to others, which is a question none of the prior NLP work addresses.

Multilingual LLM Evaluation. Large-scale multilingual evaluation suites have proliferated in recent years, including XTREME ([Hu et al., 2020](#)), XTREME-R ([Ruder et al., 2021](#)), and MEGA ([Ahuja et al., 2023](#)). Translation-based benchmark construction, exemplified by MGSM ([Shi et al., 2023](#)) and Global MMLU ([Singh et al., 2025](#)), is the dominant strategy for cross-lingual coverage. Prior work has documented performance gaps across languages and attributed them to pretraining data distributions ([Lai et al., 2023](#); [Zhang et al., 2024](#)), but the question of whether the *measurement instrument itself* behaves consistently across languages has received little psychometric attention. Our work provides a principled test of this assumption at scale.

Differential Item Functioning. Differential Item Functioning (DIF) is the standard psychometric framework for testing whether an item measures the same construct at the same difficulty level across different groups of equal ability ([Holland](#)

and Wainer, 2012). An item exhibiting DIF functions differently across groups not because the groups differ in latent ability, but because the item carries group-specific bias or ambiguity that is independent of the underlying construct. IRT-based approaches compare ICC parameters across groups and test for statistically significant differences (Lord, 2012). Our work relates to but differs from classical DIF analysis in several ways. Traditional DIF tests whether a *single item* functions differently across two groups conditional on latent ability, using a significance test. We instead assess whether *rank orderings of all items* by their estimated IRT parameters are consistent across all language versions of a benchmark simultaneously which is a global concordance question rather than an item-level significance test. Furthermore, in our setting the “groups” are not distinct human subpopulations but distinct language versions of the same benchmark administered to the same set of LLMs. A formal per-item DIF analysis across language groups would complement our concordance approach; however, classical DIF requires equating groups on latent ability, which is ill-defined here because language itself affects overall model performance distributions. Our global concordance framework sidesteps this confound by assessing rank-order agreement across all items simultaneously, rather than conditioning on matched-ability subgroups.

In the context of translated assessments, DIF has been studied extensively in educational measurement. Allalouf et al. (1999) showed that cross-lingual DIF in translated scholastic aptitude tests arises from changes in linguistic difficulty, item format, and cultural familiarity under translation. Hambleton et al. (2005) and Grisay et al. (2007) extend this to large-scale international assessments such as PISA and TIMSS, providing systematic procedures for identifying cross-lingual DIF and removing items whose difficulty shifts under translation the same validity concern that motivates our cross-lingual concordance analysis. Cross-lingual measurement invariance has been applied to psychological scales and survey instruments (Milfont and Fischer, 2010), but to our knowledge has not been systematically applied to LLM evaluation benchmarks.

B Implementation Details

Artifacts. We use LLM models publicly available on HuggingFace, and two multilingual benchmarks also sourced from HuggingFace: Multilingual GSM (MGSM; Shi et al. 2023) and Global MMLU (Singh et al., 2025). We create and release two codebases: one for evaluating LLMs on both benchmarks, and one for fitting IRT models on the resulting evaluation data and further analysis based on that data. No new models or datasets are created.

LLM evaluation. Evaluations are run using lm-evaluation-harness v0.4.10 (Gao et al., 2024) with vllm v0.13.0 as the inference backend. We evaluate 102 models on MGSM and 103 models on Global MMLU, selected from the HuggingFace Open LLM Leaderboard as the highest-ranked instruction-tuned open-weight models with fewer than 20 billion parameters and native compatibility with lm-evaluation-harness. Each model is called in inference-only mode with a maximum context length of 12,000 tokens and a batch size of 8. Results are stored as binary correctness labels (item–response pairs), yielding 568,900 evaluated pairs in total.

IRT training. One model is fitted per benchmark slice (11 MGSM language slices, 42 MMLU language×domain slices), independently for each of the three IRT variants (1PL, 2PL, 3PL), giving 159 fitted models in total. **1PL** is fitted using py-irt v0.6.6 (Natesan et al., 2016) on top of Pyro v1.9.1 (Bingham et al., 2019) with PyTorch v2.9.1 as the backend: SVI with Adam optimiser, Trace-ELBO loss, exponential learning rate schedule ($\eta = 0.1$, $\gamma = 0.9999/\text{epoch}$), dropout $p = 0.5$, vague priors, 2,000 epochs; parameters at minimum-ELBO epoch retained. **2PL and 3PL** are fitted via MAP using our own PyTorch implementation: Adam optimiser with cosine annealing; 4,000 epochs for 2PL and 6,000 for 3PL; with best MAP objective retained. Held-out fidelity evaluations use an 80/20 model train/test split, with IRT refitted on the training split only; random-item baselines are averaged over 200 independent draws per slice. Statistical analyses use scipy v1.15.3, numpy v2.2.6, pandas v2.3.3, and matplotlib v3.10.3.

Compute. All LLM model evaluations are run on a high performance computing cluster using one NVIDIA A100 GPU per job (120 GB RAM, CUDA

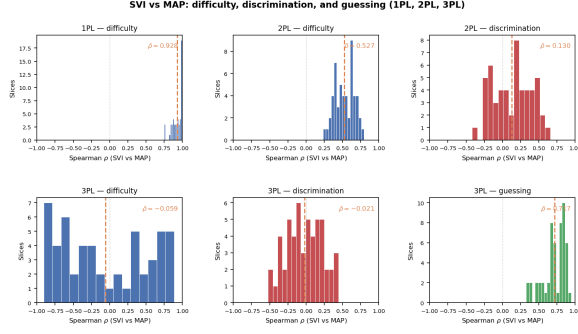


Figure 12: Per-slice Spearman ρ between SVI and MAP estimates for each (model, parameter) combination across all 53 benchmark slices. Dashed orange lines mark the mean ρ per panel. 1PL difficulty shows strong SVI–MAP agreement; discrimination collapses under SVI for 2PL and 3PL.

11.8). IRT fitting and all statistical analyses run on CPU. Total GPU compute was approximately **280 A100-GPU-hours** for LLM inference. IRT fitting across all 159 models (53 slices $\times$ 3 variants) required under one CPU-hour in total; statistical analyses are similarly negligible.

C SVI vs. MAP Inference Comparison

To verify that our choice of inference method does not confound cross-model comparisons, we fit a MAP-based 1PL to all 53 benchmark slices alongside MAP-based 2PL and 3PL, and compare difficulty, discrimination, and guessing estimates against their SVI counterparts (Figure 12).

For **1PL**, SVI and MAP difficulty estimates are highly concordant (mean $\rho = 0.928$, min = 0.749, max = 0.996), confirming that using `py-irt`’s SVI for 1PL does not affect our results.

For **2PL and 3PL**, the comparison directly illustrates why MAP is necessary. Discrimination estimates are near-uncorrelated across the two methods (2PL: mean $\rho = 0.130$; 3PL: mean $\rho = -0.021$): SVI drives discrimination toward zero (2PL) or negative values (3PL), while MAP with softplus-constrained discrimination recovers psychometrically meaningful positive values. Under 3PL the distortion is severe enough to corrupt even difficulty estimates (mean $\rho = -0.059$). The guessing parameter is relatively robust across inference methods (mean $\rho = 0.717$), consistent with its role as a lower-asymptote constraint less sensitive to discrimination interactions.

D Split-Half Analysis

To verify that discrimination and guessing parameters are reliably estimated under MAP rather than under-identified given our model pool size, we conduct a within-language split-half reliability check. For each of the 11 MGSM languages, we randomly partition the full $N=102$ model pool into two disjoint halves ($N=51$ each) and independently refit 2PL/3PL via MAP on each half, repeating this for 3 random splits per language (33 fits per model in total).

Agreement between the two halves is strong: mean Spearman $\rho = 0.614$ for 2PL discrimination ($[0.424, 0.754]$), $\rho = 0.621$ for 3PL discrimination ($[0.382, 0.762]$), and $\rho = 0.708$ for 3PL guessing ($[0.517, 0.872]$), despite each half containing roughly half as many models as the full sample used in the paper. Each half also correlates strongly with the full-sample ($N=102$) estimates reported in the main text: $\rho \in [0.776, 0.931]$ (mean 0.866) for 2PL discrimination, $\rho \in [0.748, 0.917]$ (mean 0.860) for 3PL discrimination, and $\rho \in [0.708, 0.973]$ (mean 0.874) for 3PL guessing.

Guessing spread at $N=51$ is nearly identical to the full sample (mean $\sigma = 0.049$ vs. 0.051); discrimination spread is somewhat reduced at $N=51$ (mean $\sigma = 0.445$ vs. 0.578 for 2PL, 0.418 vs. 0.549 for 3PL), consistent with expected shrinkage toward the MAP prior under a smaller model pool, though item rankings by discrimination are preserved, as shown by the correlations above. Together, these results indicate that discrimination and guessing are reliably recoverable within a language at half the model pool size, and that their low cross-lingual concordance (Appendix F) reflects cross-lingual instability rather than under-identified noise.

E Correlated Raters Analysis

Kendall’s W aggregates agreement across raters (here, languages); a tight cluster of correlated raters could inflate the statistic even if the remaining raters disagree substantially, analogous to non-independent raters inflating inter-rater reliability. section 3.1 shows that European languages form exactly such a tight cluster under item difficulty. To test whether this cluster is responsible for the concordance values reported in §3.1, we recompute W after removing all Indo-European languages: German, English, Spanish, French, Bengali, and Russian for MGSM (5 languages remain: Japanese,

	W_{full}	$W_{\text{restr.}}$	Δ
MGSM 1PL	0.544	0.560	+0.016
MGSM 2PL	0.526	0.553	+0.027
MGSM 3PL	0.519	0.547	+0.028
MMLU 1PL	0.592	0.606	+0.014
MMLU 2PL	0.540	0.574	+0.034
MMLU 3PL	0.582	0.593	+0.011

Table 5: Kendall’s W for item difficulty before (W_{full}) and after ($W_{\text{restricted}}$) removing Indo-European languages. MGSM: 11 $\rightarrow$ 5 languages. MMLU: 7 $\rightarrow$ 3 languages (mean across domains). Concordance increases in every case.

Swahili, Telugu, Thai, Chinese), and German, English, Spanish, and French for MMLU (3 languages remain: Japanese, Swahili, Chinese), the latter averaged over its six domains (results in Table 5).

Removing the Indo-European cluster *increases* W in every case: MGSM rises from 0.544 $\rightarrow$ 0.560 (1PL), 0.526 $\rightarrow$ 0.553 (2PL), and 0.519 $\rightarrow$ 0.547 (3PL); MMLU rises from 0.592 $\rightarrow$ 0.606 (1PL), 0.540 $\rightarrow$ 0.574 (2PL), and 0.582 $\rightarrow$ 0.593 (3PL). If the Indo-European cluster were propping up the reported concordance, its removal should have lowered W ; instead, the more typologically distinct remaining languages reproduce the same signal on their own. The MMLU restricted set has only three languages and its estimates therefore carry more sampling variance than the seven-language originals, but the same pattern holds on MGSM’s larger five-language restricted set. These results indicate that the reported concordance values are not an artifact of rater redundancy.

F Spearman Rank Correlation Heatmaps

In main text, Figure 2 shows pairwise Spearman rank correlations for item *difficulty* while Figure 3 provides the corresponding heatmaps for *discrimination* (a, available from 2PL onward) and *guessing* (c, 3PL only). Here, we provide a more detailed discussion of findings:

Discrimination (a). Discrimination parameters show low cross-lingual agreement under both 2PL and 3PL. For MGSM 2PL, the mean pairwise $\rho = 0.103$ (range $[-0.068$ (English–Japanese) to 0.275 (German–Russian)]: while the estimates are positive and psychometrically meaningful within each language, they do not agree strongly across languages, indicating that how sharply an item discriminates between models is a language-specific property of item formulation.

Under MGSM 3PL the mean pairwise $\rho = 0.098$ (range $[-0.085, 0.241]$), consistent with the additional guessing parameter absorbing some residual shared signal and leaving discrimination with little cross-lingual structure. MMLU 2PL and 3PL show a similarly low average discrimination agreement ($\bar{\rho} \approx 0.10$ for both), indicating that even with larger per-domain item pools, discrimination does not generalize cross-lingually. This provides clear evidence that discrimination captures language-idiosyncratic surface features of item formulation rather than a shared psychometric property.

Guessing (c). The guessing parameter tells a markedly different story. For MGSM 3PL, $\bar{\rho} = 0.349$ (range $[0.186, 0.537]$), with Bengali–German achieving the strongest agreement ($\rho = 0.537$), a plausible finding if both language versions impose similar chance-level accessibility on the same underlying items. English–Telugu is the weakest pair ($\rho = 0.186$); all pairs remain positive. For MMLU 3PL, guessing agreement is comparably stable: $\bar{\rho} = 0.359$ (range $[0.138, 0.538]$), with every language pair yielding a positive correlation and Japanese–Chinese the most consistent ($\rho = 0.538$). The consistently positive guessing correlations, in contrast to near-zero or negative discrimination correlations under the same model, suggest that lower-bound response behavior is a genuine language-invariant property: some items simply impose a floor on the probability of a correct response regardless of the language of presentation, while discrimination is driven by language-specific surface features of the item formulation.

Science vs Non-Science Domain Figure 13 shows the 2×3 grid of averaged Spearman difficulty heatmaps for science vs non-science domains across the three IRT model variants. Under 1PL, science heatmaps show a predominantly warm palette (Kendall’s $W = 0.647$, $\bar{\rho} = 0.588$), while non-science is already lower ($\bar{\rho} = 0.492$, $W = 0.564$). Under MAP-fitted 2PL and 3PL, the science advantage is maintained: science domains achieve $\bar{\rho} = 0.559$, $W = 0.620$ (2PL) and $\bar{\rho} = 0.581$, $W = 0.640$ (3PL), while non-science remains somewhat lower at $\bar{\rho} = 0.420$, $W = 0.501$ (2PL) and $\bar{\rho} = 0.479$, $W = 0.553$ (3PL). Both domain types maintain moderate cross-lingual concordance across all model variants, with science items consistently showing stronger agreement. This reflects that science items test well-defined reasoning steps or unambiguous facts whose difficulty is

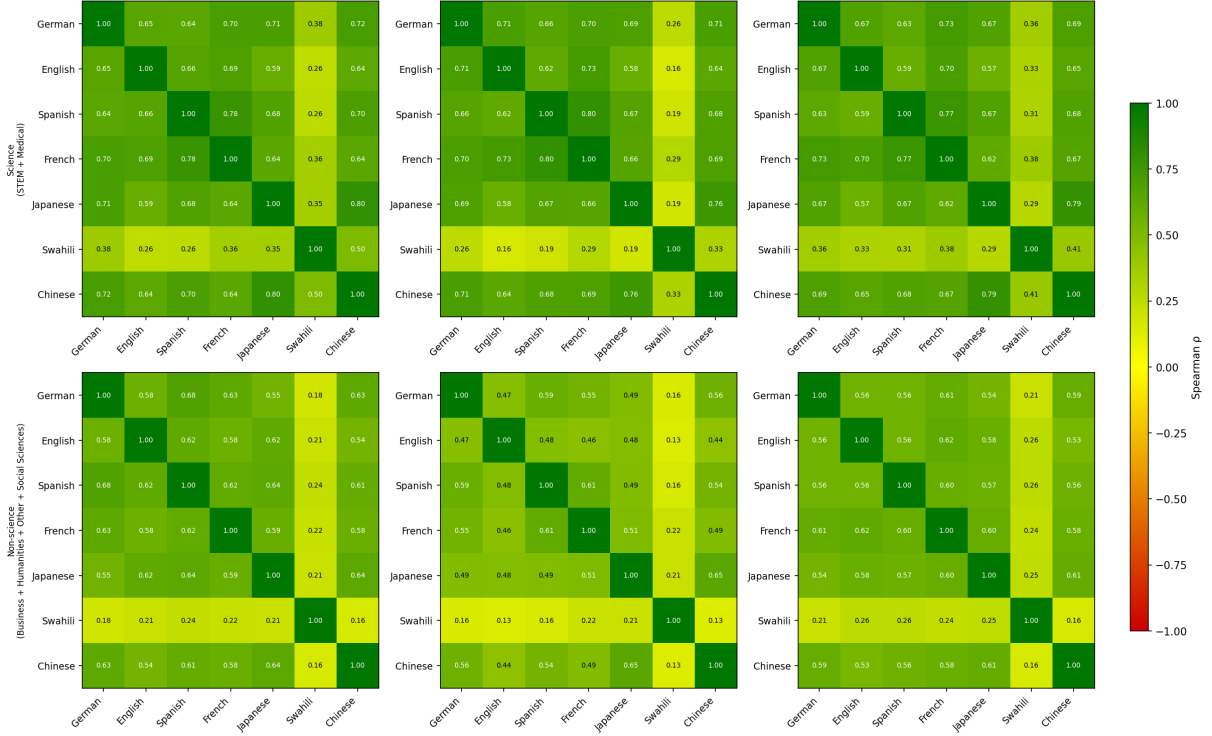


Figure 13: Pairwise Spearman rank correlation of item *difficulty* across languages for MMLU, split by domain type (rows) and IRT model variant (columns). **Top row:** Science domains (STEM + Medical), averaged element-wise. **Bottom row:** Non-science domains (Business + Humanities + Other + Social Sciences), averaged element-wise. Colour encodes $\rho \in [-1, +1]$: red = -1 , yellow = 0 , green = $+1$.

partly preserved under translation, whereas non-science difficulty is dominated by culturally situated knowledge and nuanced phrasing that shifts substantially across languages.

G DIF Analysis

Table 6 and Table 7 report the per-language DIF rates and mean anchor-corrected difficulty shifts underlying subsection 3.3, and Table 8 breaks the MMLU results down by domain. For MGSM, the anchor set is fixed at 63 items (bottom quartile of items by cross-lingual rank standard deviation) and 187 items are tested per language pair. For MMLU, both the anchor and tested set sizes vary by domain, following each domain’s item count.

H Per-Item Rank Stability Distributions

Figure 14 and Figure 15 shows per-item rank standard deviation sorted ascending (most to least consistent left to right). Each bar is one item; bar height is the standard deviation of its rank position across all languages. Multi-parameter models show one sub-panel per parameter.

Lang.	% DIF	mean $ \Delta b $
zh	34.2	0.992
es	34.8	0.843
ru	38.0	0.951
de	43.3	1.064
th	45.5	0.991
ja	48.1	1.097
fr	53.5	1.273
bn	59.9	1.303
sw	63.1	1.347
te	63.1	1.501

Table 6: MGSM: % of items showing significant DIF relative to English (BH-FDR $q = 0.05$), and mean anchor-corrected difficulty shift $|\Delta b|$, by language. $n_{\text{anchor}} = 63$, $n_{\text{tested}} = 187$ for every language pair.

I Within-Language Model Fit

To assess within-language fit quality for each IRT variant, we report the MAP negative log-likelihood per observation ($-\ell/\text{obs}$) at convergence, where lower values indicate better fit. Table 9 reports these values for every slice.

As expected under correct model fitting (2PL nests 1PL; 3PL nests 2PL), 2PL achieves better fit than 1PL on all 53 slices. 3PL achieves better fit than 2PL on 43/53 slices; on the remaining 10,

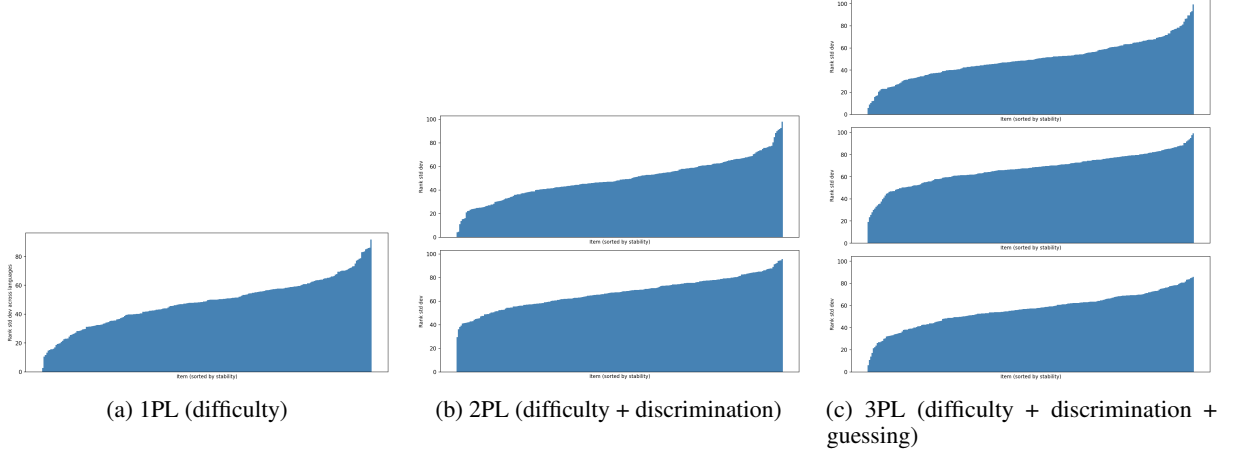


Figure 14: MGSM per-item rank standard deviation across 11 languages (250 items). Each subfigure shows all parameters for that model variant stacked vertically.

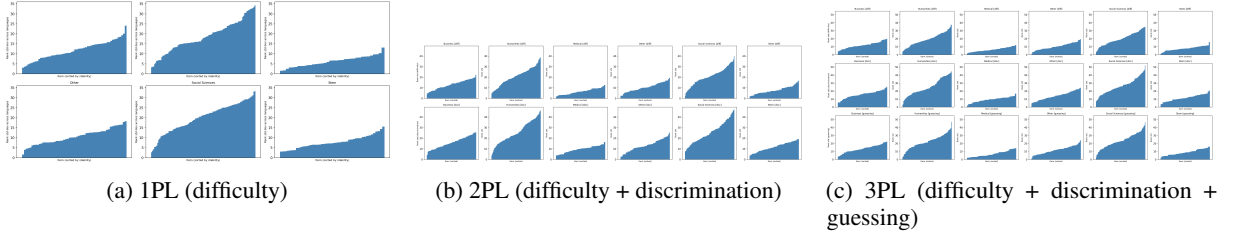


Figure 15: MMLU per-item rank standard deviation across 7 languages, pooled over six domains ($n \leq 102$ items per domain).

Lang.	% DIF	mean $ \Delta b $
es	21.0	1.051
fr	22.4	1.124
de	25.3	1.326
zh	26.1	1.417
ja	26.6	1.412
sw	69.1	2.668

Table 7: MMLU: % of items showing significant DIF relative to English, and mean anchor-corrected difficulty shift, by language, each averaged across the six MMLU domains.

Domain	% DIF	mean $ \Delta b $
Humanities	23.5	1.190
Medical	25.9	1.441
Other	28.2	1.248
Business	32.6	1.842
Social Sciences	35.3	1.621
STEM	45.1	1.656

Table 8: MMLU: % of items showing significant DIF and mean anchor-corrected difficulty shift, by domain, averaged across the six focal languages relative to English.

3PL’s fit is marginally worse (by $0.0001 - -0.008$ nats/obs), consistent with the added guessing parameter being harder to identify at MAP than genuinely reducing fit. The improvement is larger on MMLU (mean 1PL $-\ell/\text{obs} = 0.180$ vs 2PL 0.127 vs 3PL 0.124) than on MGSM (mean 1PL 0.426 vs 2PL 0.389 vs 3PL 0.389), consistent with MMLU’s more heterogeneous item difficulty permitting greater discrimination by 2PL, and the guessing parameter capturing floor-level response probabilities in knowledge-intensive domains.

J IIF Theta Sensitivity

The main analysis evaluates IIF at $\theta = 0$ (median ability). To assess sensitivity to this choice, we recompute item selections at $\theta \in \{-1, 0, +1\}$ and measure the Jaccard similarity of the resulting top-50% subsets across theta values within each model, and cross-model Jaccard at each theta.

Within-model sensitivity. Under MAP, 2PL item selection shows meaningful theta sensitivity on both benchmarks: MGSM 2PL mean $J = 0.639$ between $\theta = -1$ and $\theta = 0$; $J = 0.429$ between $\theta = 0$ and $\theta = +1$; $J = 0.232$ between $\theta = -1$

and $\theta = +1$. This reflects that properly-identified discrimination parameters vary across items, so IIF rankings depend on theta. 3PL shows similar sensitivity (MGSM: $J = 0.617, 0.400, 0.208$ for the same pairs). 1PL shows moderate theta sensitivity on MGSM (mean $J = 0.727$ between $\theta = -1$ and $\theta = 0$; $J = 0.835$ between $\theta = 0$ and $\theta = +1$) but near-invariance on MMLU (mean $J \geq 0.929$). The greater sensitivity on MGSM reflects the wider difficulty range of arithmetic items relative to the narrower MMLU domain slices.

Cross-model concordance across theta. Table 10 shows median Jaccard for each model pair at each θ . The qualitative ordering is preserved across all theta values: 1PL–2PL and 1PL–3PL agreement both remain above the random baseline ($J \approx 0.33$) at every θ , and 2PL–3PL agreement is consistently high ($J \geq 0.87$). The conclusion that 2PL and 3PL select overlapping but not identical subsets relative to 1PL therefore holds across all evaluated ability levels.

K Compression Ratio Sensitivity

The main analysis uses a top-50% IIF cutoff. To assess sensitivity to this choice, Table 11 reports in-sample mean Spearman ρ (subset vs. full benchmark) at top-30%, top-50%, and top-70% for each IRT variant and benchmark.

The qualitative ordering is preserved across all ratios. 1PL is reliable at 50% and above on both benchmarks; at 30%, fidelity dips slightly (MMLU mean $\rho = 0.931$). 2PL is competitive on MGSM at all compression levels (mean $\rho \geq 0.939$), but shows substantially lower fidelity on MMLU at all thresholds (mean $\rho = 0.771$ – 0.875), remaining below 1PL. 3PL follows a similar pattern, with MMLU fidelity lower than 2PL and 1PL at all ratios (mean $\rho = 0.674$ – 0.721). The 50% default is the point where 1PL clears the random baseline reliably across all slices while achieving meaningful compression.

L Subset Ranking Fidelity

Here, we provide: per-slice fidelity across all benchmark variants (Figure 16), and per-model accuracy-prediction scatter for in-sample (Figure 17) and held-out (Figure 18) evaluations.

Per-slice Fidelity (Figure 16). Each dot is one benchmark slice — 11 MGSM language slices and 42 MMLU language $\times$ domain slices. For

MGSM, 1PL is uniformly strong across all 11 languages (median $\rho = 0.982$, IQR entirely above $\rho = 0.95$), with no language slice falling below the random baseline. 2PL is similarly strong (median $\rho = 0.971$, IQR 0.957 – 0.976), with the minimum at $\rho = 0.915$, well above the random baseline; no slice fails badly. 3PL is comparable (median $\rho = 0.973$, IQR 0.955 – 0.976 , min 0.911). For MMLU, the contrast between 1PL and the multi-parameter models sharpens considerably. 1PL achieves median $\rho = 1.000$ — the majority of the 42 slices reach perfect rank correlation — with all slices well above the random baseline ($\rho \approx 0.70$). 2PL falls to median $\rho = 0.878$ (IQR 0.740 – 0.970 , min 0.354): most MMLU slices retain useful fidelity, but the lower tail extends below the random baseline on a minority of slices. 3PL is further reduced (median $\rho = 0.764$, IQR 0.594 – 0.892 , min 0.290), with more slices falling below the reference line than under 2PL.

In-sample accuracy prediction (Figure 17).

Each point is one (model, benchmark-slice) pair ($n = 1,122$ for MGSM; $n = 4,320$ for MMLU). The dashed $y = x$ line marks perfect accuracy prediction; deviations from it indicate systematic bias in how the subset estimates a model’s full-benchmark score. Under 1PL, MGSM points cluster tightly near the diagonal (MAE = 0.086 , bias = $+0.026$): the subset very slightly overestimates full accuracy, consistent with IIF at $\theta = 0$ selecting items near median difficulty, which form a marginally easier sub-pool than the benchmark as a whole. Under 1PL, MMLU shows the opposite bias (MAE = 0.098 , bias = -0.065): the IIF-selected items are slightly harder than the MMLU average because IIF peaks for items with difficulty $b \approx 0$, which are harder than the typical MMLU item (most items are easier than chance for these models). Consequently, subset accuracy understates full-benchmark accuracy by about six percentage points on average. Under 2PL and 3PL, MGSM accuracy prediction is close to 1PL (MGSM 2PL MAE = 0.103 , bias = $+0.073$; 3PL MAE = 0.101 , bias = $+0.079$). On MMLU, 2PL IIF introduces a moderate positive bias ($+0.080$, MAE = 0.117): the discrimination-weighted selection still over-selects items where models score high, inflating subset accuracy estimates relative to the full benchmark by about 8 percentage points.

Held-out accuracy prediction (Figure 18). The held-out evaluation re-fits IRT on 80% of models

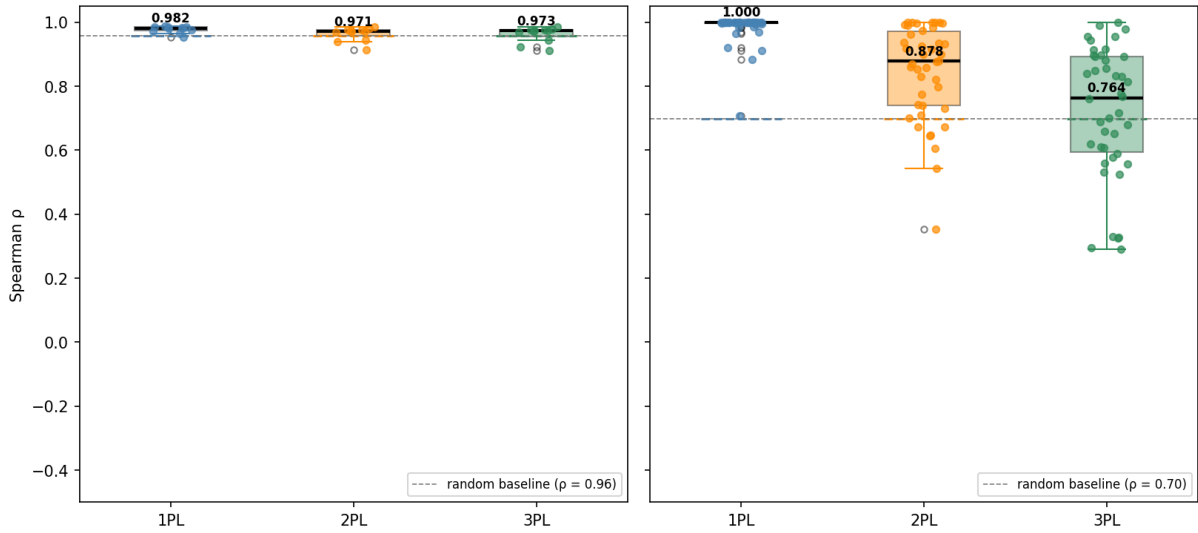


Figure 16: Distribution of Spearman ρ (subset vs. full benchmark) across all in-sample benchmark slices: 11 MGSM language slices (left) and 42 MMLU language $\times$ domain slices (right). Each dot is one slice; annotated values are medians. Dashed lines mark for the random-item baseline ($\rho \approx 0.96$ for MGSM, ≈ 0.70 for MMLU).

and measures accuracy prediction on the remaining 20%. For 1PL, the held-out MAE is nearly identical to in-sample on both benchmarks (MGSM: 0.086 held-out vs. 0.086 in-sample; MMLU: 0.090 vs. 0.098), and the directional bias is unchanged, confirming that 1PL item selection generalizes to unseen models without any degradation in accuracy-prediction quality. For MGSM 2PL, held-out MAE (0.108) is comparable to in-sample (0.103 in-sample is for all models; held-out evaluates on the 20% unseen split), reflecting that MAP discrimination estimates generalize reasonably to unseen models within MGSM. For MMLU, 2PL and 3PL retain moderate-to-large positive bias (+0.100 and +0.181 respectively) in held-out evaluation, confirming that the positive accuracy bias is a structural property of discrimination-weighted item selection on knowledge-intensive benchmarks rather than an in-sample artifact.

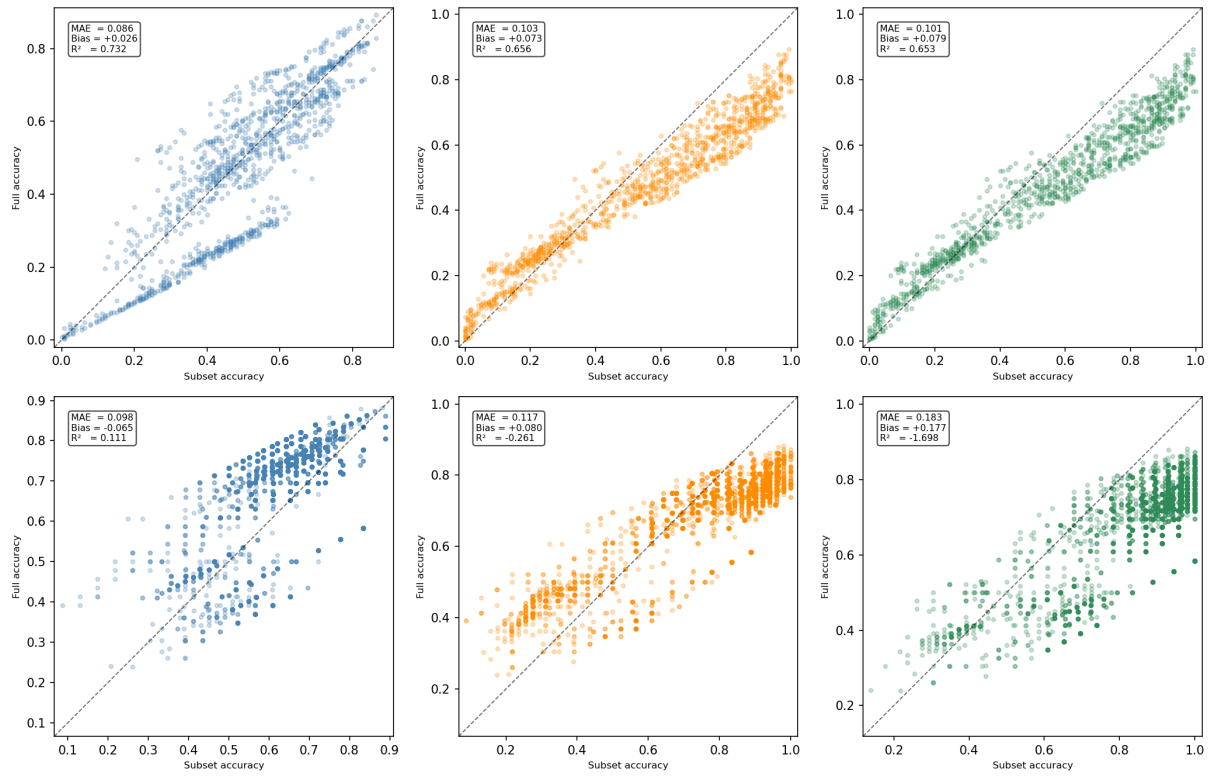


Figure 17: In-sample subset accuracy vs. full-benchmark accuracy for each (model, benchmark-slice) pair. Rows: MGSM (top) and MMLU (bottom). Columns: 1PL, 2PL, 3PL. Dashed line is $y = x$ (ideal); each panel is annotated with mean absolute error (MAE), Spearman ρ , and mean signed bias (subset – full). 1PL points track the diagonal tightly with low bias; 2PL and 3PL scatter widely and show systematic overestimation on MMLU.

Benchmark	Slice	1PL	2PL	3PL
MGSM	bn	0.360	0.311	0.313
	de	0.453	0.416	0.415
	en	0.420	0.383	0.381
	es	0.492	0.464	0.463
	fr	0.454	0.420	0.420
	ja	0.466	0.436	0.435
	ru	0.479	0.445	0.442
	sw	0.405	0.366	0.367
	te	0.284	0.236	0.244
	th	0.445	0.412	0.412
	zh	0.427	0.391	0.391
MMLU	de_business	0.176	0.137	0.138
	de_humanities	0.139	0.104	0.104
	de_medical	0.122	0.077	0.077
	de_other	0.150	0.102	0.101
	de_social_sciences	0.166	0.132	0.129
	de_stem	0.222	0.151	0.146
	en_business	0.108	0.066	0.062
	en_humanities	0.112	0.078	0.081
	en_medical	0.113	0.092	0.095
	en_other	0.155	0.122	0.119
	en_social_sciences	0.151	0.113	0.112
	en_stem	0.178	0.125	0.122
	es_business	0.163	0.124	0.119
	es_humanities	0.126	0.085	0.088
	es_medical	0.153	0.088	0.078
	es_other	0.186	0.137	0.133
	es_social_sciences	0.137	0.109	0.106
	es_stem	0.213	0.151	0.145
	fr_business	0.174	0.134	0.135
	fr_humanities	0.123	0.093	0.093
	fr_medical	0.141	0.091	0.095
	fr_other	0.175	0.124	0.117
	fr_social_sciences	0.146	0.113	0.111
	fr_stem	0.231	0.170	0.165
	ja_business	0.162	0.086	0.077
	ja_humanities	0.140	0.095	0.093
	ja_medical	0.158	0.086	0.083
	ja_other	0.182	0.128	0.120
	ja_social_sciences	0.184	0.137	0.134
	ja_stem	0.201	0.121	0.114
	sw_business	0.332	0.260	0.255
	sw_humanities	0.319	0.246	0.241
	sw_medical	0.251	0.161	0.155
	sw_other	0.294	0.233	0.231
	sw_social_sciences	0.337	0.258	0.251
	sw_stem	0.305	0.224	0.219
	zh_business	0.137	0.079	0.078
	zh_humanities	0.136	0.095	0.090
	zh_medical	0.125	0.077	0.071
	zh_other	0.198	0.132	0.123
	zh_social_sciences	0.140	0.105	0.103
	zh_stem	0.178	0.113	0.109

Table 9: MAP negative log-likelihood per observation ($-\ell/\text{obs}$) for each IRT model variant and benchmark slice. Lower values indicate better within-language fit.

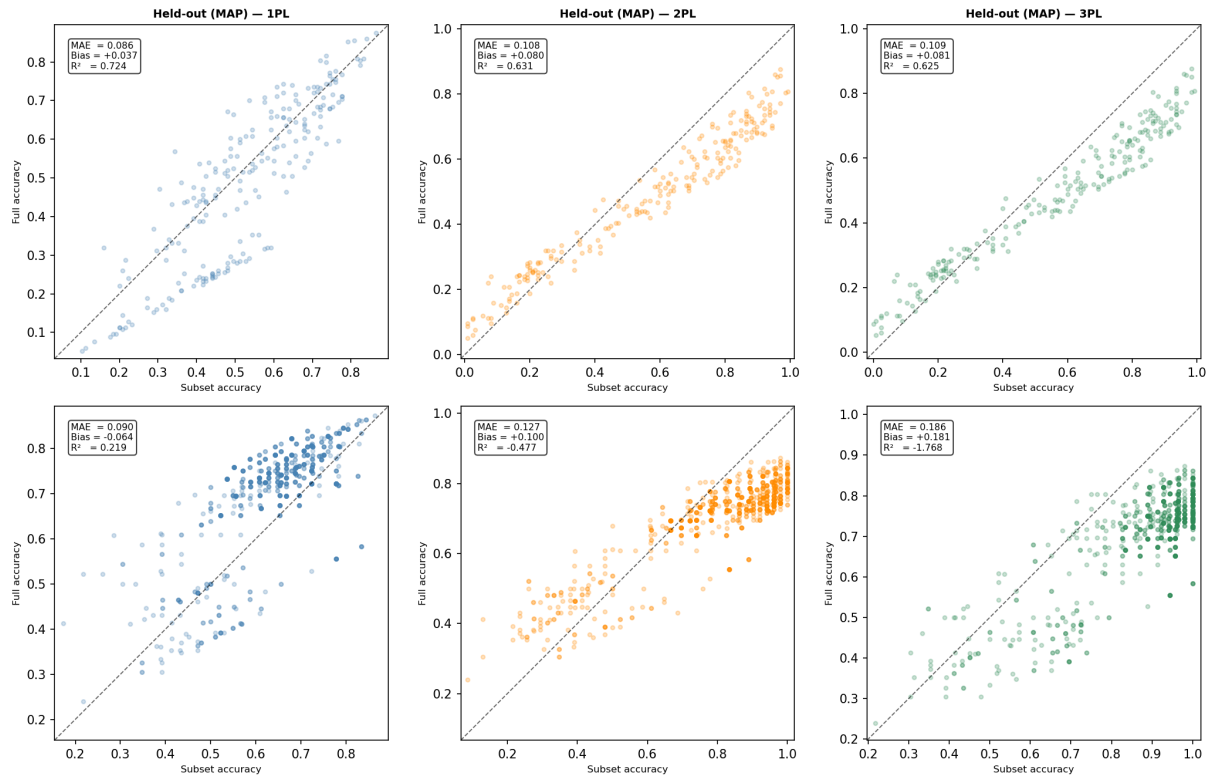


Figure 18: Held-out subset accuracy vs. full-benchmark accuracy. IRT re-fitted on 80% of models; each point is one (held-out model, benchmark-slice) pair ($n = 220$ for MGSM; $n = 876$ for MMLU). Layout and annotation as in Figure 17. 1PL MAE and bias are essentially unchanged from in-sample; 2PL and 3PL retain their large positive bias on MMLU.

Pair	θ	MGSM median J	MMLU median J
1PL vs 2PL	-1	0.553	0.268
	0	0.404	0.528
	+1	0.506	0.672
1PL vs 3PL	-1	0.524	0.273
	0	0.412	0.454
	+1	0.543	0.672
2PL vs 3PL	-1	0.923	0.688
	0	0.923	0.590
	+1	0.866	0.895

Table 10: Median Jaccard similarity of top-50% IIF-selected subsets across model pairs, evaluated at three theta values. The random-selection baseline is $J \approx 0.33$. Qualitative orderings are stable across all theta values.

Model	Top-k	MGSM mean ρ	MMLU mean ρ
1PL	30%	0.939	0.931
	50%	0.977	0.983
	70%	0.994	0.999
2PL	30%	0.939	0.771
	50%	0.963	0.844
	70%	0.974	0.875
3PL	30%	0.939	0.674
	50%	0.961	0.721
	70%	0.975	0.717

Table 11: In-sample ranking fidelity (mean Spearman ρ , subset vs. full benchmark) at three compression ratios. The random-selection baseline is $\rho = 0.96$ for MGSM and $= 0.70$ for MMLU. 1PL conclusions are stable across all ratios; 2PL and 3PL fail on MMLU at every threshold.