

Can We Trust Item Response Theory for AI Evaluation?

Han Jiang¹ Sunbeom Kwon² Jinwen Luo³ Ziang Xiao¹ Susu Zhang²

¹Johns Hopkins University ²University of Illinois Urbana-Champaign

³University of California, Los Angeles

hjiang66@jh.edu, szhan105@illinois.edu

Abstract

AI benchmarks increasingly leverage item-level statistical models, particularly item response theory (IRT), to estimate capabilities, rank systems, select informative examples, and diagnose benchmark quality. However, AI benchmark data often departs from the data regimes encountered in human testing, for which standard IRT estimation tools were originally developed: these benchmarks typically involve fewer evaluated models, far more items, and capability distributions that may be skewed, clustered, or multimodal. We examine how these regime mismatches challenge the reliability of IRT modeling for AI evaluation. Using item parameters and capability distributions derived from six widely used LLM benchmarks, we simulate response matrices under three common IRT models and compare four estimators used in recent benchmark studies: marginal maximum likelihood, Markov chain Monte Carlo, variational inference, and a neural pseudo-Siamese estimator. Across 18,000 simulation runs, we systematically evaluate computational feasibility, scalability, and the reliability of IRT inferences about model rankings, predicted performance, and item characteristics. Results show that classical estimators can become computationally infeasible for large benchmarks, whereas scalable estimators can produce unreliable item-level and ranking inferences with small or non-normally distributed model sets. This study identifies when latent trait models reliably support or risk distorting AI benchmarking claims, and what sample sizes and diagnostics are needed for trustworthy use.

1 Introduction

AI benchmarks have become the primary instrument for evaluating model capabilities, comparing AI systems, and informing deployment decisions. Yet the dominant practice of relying on aggregate metrics such as mean accuracy is deceptively simple: it reduces a rich response matrix to a single number, treating all items as equally informative regardless of their difficulty or discriminative power [50, 17]. This simplification dilutes informative signals in evaluation data, renders model rankings sensitive to benchmark composition, and hinders the accumulation of comparable findings across evaluation efforts, reflecting a deep tension between validity, reliability, and efficiency [25, 19].

This tension naturally directs attention to psychometrics, the discipline devoted to the science of human testing and assessment, for well-established measurement tools. *Item response theory (IRT)*, the most widely adopted statistical framework, offers a principled alternative. Instead of treating each item equally, IRT jointly estimates parameters of item characteristics and examinee ability on a shared latent scale, enabling in-depth diagnosis of item quality, efficient testing through adaptive item selection, and cross-benchmark comparison of ability estimates [40]. Motivated by these advantages, a growing body of work has adopted IRT for AI evaluation, applying it to produce finer-grained model and item diagnostics (e.g., [31, 54]) and enable more efficient evaluation strategies, such as adaptive AI testing (e.g., [26, 61]) and benchmark compression (e.g., [47]).

Nevertheless, the adoption has outpaced validation. A crucial concern is that *AI benchmark data fundamentally depart from the human testing regime for which IRT estimation methods were developed and validated*. As illustrated in Fig. 1, the number of evaluated models is typically small (fewer than 100) while the number of items can reach tens of thousands. This differs from typical psychometric settings, where hundreds to thousands of examinees respond to tests of 20-100 items [14]. Furthermore, the model capability distributions are frequently non-normal, exhibiting multimodality or skewness that conflicts with the normality assumptions underlying standard estimation procedures [67, 58]. *These departures have important implications*, as many IRT estimation methods’ statistical validity directly depends on the sample-to-parameter ratio and the match between the assumed and the actual trait distribution. Violations of these assumptions may systematically bias inferences on both item characteristics and model capabilities.

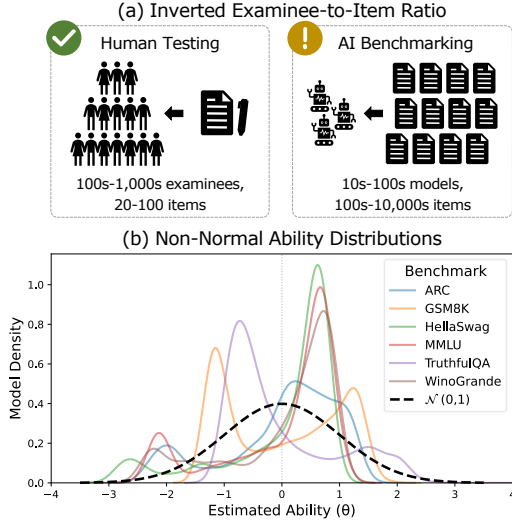


Figure 1: Illustration of the data regime mismatch between human testing and AI benchmarking.

Existing studies that apply IRT do not focus on verifying whether the chosen IRT models and estimation methods produce reliable estimates under this regime mismatch, leaving critical questions unanswered. It remains unknown how small model populations, non-normal capability distributions, and long benchmarks affect parameter recovery and downstream inferences such as model rankings and item-level diagnostics. Practical guidance for conducting such studies is equally lacking: for instance, when constructing an item-level benchmark repository (e.g., [27]), how many evaluated models will be sufficient to fit a given IRT model reliably?

To the best of our knowledge, no prior work has systematically addressed these questions. This paper fills this gap through three contributions:

- A **literature review** surveying 19 studies that apply IRT to AI benchmarking, characterizing the current landscape of IRT usage and empirically identifying the range of application conditions that inform the design of our simulation study (Sec. 3).
- A **simulation study**¹ calibrated to representative AI evaluation data regimes, systematically evaluating computational feasibility, parameter recovery, and downstream inference validity across six benchmarks, three IRT models, four estimators, and five sample sizes (Secs. 4 & 5).
- A **Practical guide** based on the findings, including recommendations on minimum sample sizes, IRT model selection, and estimation methods, clarifying when IRT reliably supports AI benchmarking claims and when it risks distorting them (Sec. 6).

Sec. 2 provides the background on IRT and simulation methods, and Sec. 7 gives concluding remarks.

2 Background

2.1 Item response theory

The limitations of aggregate metrics for AI evaluation naturally motivate the adoption of *item response theory* (IRT), a family of statistical models from psychometrics. Unlike classical test theory which primarily relies on aggregate scores [21, 20], IRT explicitly accounts for differences in item difficulty by jointly modeling item characteristics and examinee ability on a common latent scale [40, 22].

Specifically, IRT models the probability of the correct response to an item as a function of a latent ability parameter θ and one or more item-level parameters capturing item characteristics [41]. Its simplest form, the *one-parameter logistic model* (**1PL**, also known as the *Rasch* model [48]), characterizes each item solely by its *difficulty* b : $p(y=1|\theta, b) = \sigma(\theta - b)$, where $y \in \{0, 1\}$ denotes

¹Code and data: <https://github.com/Salomeeeee/IRT-Simulation>.

the response correctness and $\sigma(\cdot)$ is the sigmoid function. The *two-parameter logistic model (2PL)* introduces a *discrimination* parameter a , allowing items to differ in how effectively they distinguish between high- and low-ability examinees: $p(y = 1|\theta, a, b) = \sigma(a(\theta - b))$. The *three-parameter logistic model (3PL)* further adds a *pseudo-guessing* parameter c , representing the probability that an extremely low-ability examinee responds correctly: $p(y = 1|\theta, a, b) = c + (1 - c)\sigma(a(\theta - b))$ [42]. Extensions such as the 4PL [3] and the graded response model [53] incorporate additional parameters into the response function.

In practice, IRT models are estimated using observed response data from settings such as educational testing and AI benchmarking. Given a response matrix $Y \in \{0, 1\}^{N \times J}$ where each entry $y_{i,j}$ records whether examinee i answered item j correctly, an IRT model is fit to jointly estimate θ_i for each examinee and the item parameter(s) $\zeta_j = \{b_j, \dots\}$, for each item. Standard estimators include marginal maximum likelihood via Expectation-Maximization (**MML-EM** [5]), Markov chain Monte Carlo (**MCMC** [1, 2]), and variational inference (**VI** [64]), which we elaborate on in Sec. 3. Estimated parameters inform downstream analyses, including flagging problematic items, constructing shorter yet equally informative tests, and ranking examinees on a common scale with associated uncertainty.

Despite these benefits, we highlight two characteristics of AI benchmarking data that challenge the assumptions of common IRT estimation methods, which *potentially compromise the reliability of IRT-based AI benchmark analysis*. First, standard IRT estimators were developed primarily for settings with relatively large examinee samples and comparatively few items. Accurate estimation typically requires at least $N = 500$ for 2PL and $N = 1,000$ for 3PL [14]. Statistical consistency of MML-EM further assumes fixed J and $N \rightarrow \infty$. In contrast, AI benchmarks often evaluate fewer than 100 models on up to tens of thousands of items [18].

Second, standard estimators commonly assume normally distributed abilities, either explicitly (MML-EM) or through prior specification (Bayesian and VI). However, model capability distributions may exhibit substantial skewness or multimodality, bearing little resemblance to the assumed normal distributions [67, 58]. Evaluated models rarely constitute a random sample from a well-defined population. Instead, AI benchmarks typically comprise community-contributed convenience samples, in which many models share foundation models and adaptation pipelines [6, 60], forming a non-random, lineage-dependent convenience set with uneven representation across architecture families, training regimes, and capability strata.

2.2 Simulation studies: evaluating the evaluator

Simulation studies, also known as Monte Carlo studies, are the standard methodology for evaluating the performance of statistical methods under controlled conditions [44]. The general procedure follows a well-established workflow:

- specify true values for parameters that are ordinarily unobservable,
- generate simulated response data from these known parameters and model,
- fit the model of interest to the simulated data, and
- compare the estimated parameters against the data-generating ground truth,

which has been formalized as the ADEMP framework – Aims of the study, Data generation mechanisms, Estimated parameters, Methods being evaluated, and Performance measures – for planning and reporting simulation studies [57]. By directly manipulating simulation conditions and evaluating estimation accuracy against known ground truth, we can isolate the effects of specific conditions on method performance. An IRT simulation study thus serves as a means of evaluating the evaluator itself: by systematically varying simulation conditions such as the number of evaluated models, the distribution of model capabilities, the size of AI benchmarks, and the choice of IRT models, we can directly examine whether the resulting estimates (i.e., model capabilities and item parameters) reliably support their intended uses.

In psychometrics, a substantial body of simulation research has investigated the conditions under which IRT estimates for human testing are trustworthy (e.g., [28, 56]), establishing the practical guidelines and assumptions outlined in Sec. 2.1. For example, recommended minimum sample sizes are 250-500 for 1PL and 2PL [14, 15, 55], and 1,000-3,000 for 3PL [14]. These studies also document how assumption violations, such as non-normal ability distributions or short test lengths, degrade estimation quality (e.g. [59, 63, 32]). However, these guidelines were developed under the human testing conditions described above. AI evaluation setting departs from these conditions, yet

no prior work has conducted simulation studies in this context. This paper thus addresses this gap with a comprehensive simulation study calibrated to representative AI evaluation conditions observed in our literature review.

3 Item response theory for AI evaluation

Taking the simple average score on a benchmark ignores that not all benchmark items are equally informative [50]. Items with poor discriminative power, such as those on which models perform similarly regardless of overall capability, do not contribute capability-relevant signal [23]. As models rapidly evolve and more items become trivially easy, this becomes increasingly common [46, 16]. Additionally, some items may capture shortcuts such as spurious correlations or format heuristics rather than the intended capability, jeopardizing the validity of score interpretations. Without principled methods to identify and down-weight such uninformative items, they contribute equally to aggregate scores alongside more diagnostic items. Furthermore, benchmark-specific average score lacks mechanisms to place models evaluated on different benchmarks on a common scale [37], hindering the accumulation of comparable findings across evaluation efforts.

IRT emerged as a viable approach to addressing these AI benchmarking challenges. As a survey of current IRT applications to address these limitations, we conducted a scoping review. We searched arXiv using keywords related to *item response theory (IRT)*, *AI evaluation*, and *AI benchmarking*, covering publications from 2016 to April, 2026. Two authors independently screened the results against predefined inclusion criteria: studies must apply IRT or related item-factor models to model evaluation in AI, ML, or NLP contexts; studies that merely cite IRT without applying it were excluded. This process identified 19 studies.

Across these studies, IRT-style probabilistic models have been used to support (1) *latent capability scoring* that accounts for item difficulty or discrimination in place of simple aggregate scores [35, 43, 50, 31]; (2) *item and benchmark analysis*, including the identification of difficult, informative, saturated or weakly discriminating examples [34, 36, 50, 8, 54, 67]; and (3) *efficient evaluation* through smaller item subsets, adaptive item administration or generation, and prediction of performance on unseen items or tasks, reducing the need to run every model on full benchmarks [68, 47, 31, 61, 25, 26, 38, 11]. IRT model specifications ranged from standard binary 1PL, 2PL, 3PL and 4PL, with either unidimensional or multidimensional latent ability spaces, to adaptations for data formats including pairwise comparisons (e.g., [49]), continuous scores (e.g., [38]), response-length information [65], and item text features (e.g., [11]). The *unidimensional 2PL with item-specific difficulty and discrimination parameters* was the most commonly adopted model.

Most of these studies fitted IRT models to binary response matrices, in which rows represent evaluated models and columns represent benchmark examples, or items. However, the row and column dimensions differed substantially from typical human-testing contexts and varied substantially across studies. Many included tens to low hundreds of evaluated models, including cases below 50, with analyses based on OpenLLM Leaderboard [45] as the main exception, where model counts exceeded 5,000. Benchmark length, by contrast, ranged from a few hundred to more than 80,000 examples. Estimation methods were also heterogeneous, including marginal or full-information maximum likelihood via the EM algorithm [61, 31], MCMC [25, 38, 54], variational inference [62, 50, 47, 26], and neural-network representations of IRT models fit with gradient-based optimization [67, 26].

This variation motivates a methodological study that examines how evaluated model count, benchmark length, capability distribution, IRT model specification and estimation method affect the reliability of item- and model-level inferences used to construct and interpret AI benchmark claims.

4 Simulation design

The simulation aims to examine *whether existing IRT estimation approaches provide reliable item- and model-level inferences under data regimes observed in AI evaluation*. The data conditions, estimation methods, and evaluation metrics were chosen to reflect common practices in recent applications of IRT to AI benchmark analysis.

4.1 Data generation

IRT model, sample size, and benchmark. For IRT model, we focus on unidimensional 1PL, 2PL, and 3PL as the most commonly adopted specifications in the reviewed studies.

Sample sizes, i.e., the number of evaluated models, were chosen from the empirical range observed across the studies in Sec. 3: $N = 30$, similar to sample sizes found in [67, 47, 61]; $N = 100$, similar to [26, 62, 47, 25]; $N = 180$, similar to [50, 61]; $N = 400$, similar to [47]; and $N = 1000$, as a reference condition closer to sample-size recommendations for 3PL estimation in human testing.

Benchmark conditions were based on the OpenLLM Leaderboard v1 benchmarks analyzed in [47, 25, 31]. The leaderboard aggregates results from over 6,000 community-contributed models across six benchmark datasets: ARC-CHALLENGE [12], HELLASWAG [66], MMLU [24], TRUTHFULQA [39], WINOGRANDE [51], and GSM8K [13], spanning tasks from commonsense reasoning to mathematical problem solving, with benchmark lengths ranging from approximately 800 to 15,000 items. In total, this yields $3 \times 5 \times 6 = 90$ unique simulation conditions.

True parameters and response matrices. The true item parameters and model capability distributions were derived from the preprocessed OpenLLM response matrices used in [31], which contain binary responses from over 6,000 models across the six benchmarks. Following their preprocessing, models with the lowest 0.1% of scores were removed to reduce noise, and items with standard deviations below 1% or mean accuracies above 95% were excluded as uninformative or nearly saturated. The retained matrices' dimensions (N models $\times J$ items) were as follows: $5,222 \times 644$ for TRUTHFULQA, $5,221 \times 844$ for ARC-CHALLENGE, $6,550 \times 1,045$ for WINOGRANDE, $6,068 \times 1,306$ for GSM8K, $5,221 \times 5,711$ for HELLASWAG, and $5,219 \times 12,508$ for MMLU.

For each benchmark, we fit the 1PL, 2PL and 3PL models to the empirical response matrix using variational inference (VI), given its fast computation in high- N , high- J settings. The resulting item and model capability parameters were first rescaled so that the capability distribution has mean 0 and variance of 1. Then, the item parameter estimates were treated as the *true item parameters* in the simulation. Figs. 6–11 in App. B.1 present the distributions of the estimated model capabilities and true difficulty parameters for each benchmark, illustrating the preserved distributional features such as skewness, clustering, and multimodality. Additionally, to assess whether using the VI-estimated parameters as generating values favors VI in estimator comparisons, we conducted additional simulations with PSN-IRT-estimated parameters. The results in App. B.2 showed similar relative estimator performance, except for item difficulty recovery.

For each of the 90 conditions, we generated 50 replications. In each replication, N *true capability values* were sampled from the empirical distribution of the estimated capabilities for the corresponding condition. The sampled capability values and the item parameters were then used to generate a full binary response matrix, which was used as the input data for IRT estimation.

4.2 Estimation

We evaluated four IRT estimators spanning the major paradigms used in recent IRT-based AI evaluation studies (Sec. 3): MML-EM, MCMC, VI, and PSN. Full technical details, including formulations, prior specifications, and implementation settings, are provided in App. A.

MML-EM estimates item parameters by maximizing the marginal likelihood, integrating out latent abilities under an assumed population distribution, typically the normal distribution. It is the conventional approach in psychometrics, implemented here via the Metropolis-Hastings Robbins-Monro algorithm, a stochastic variant of EM, in the `mirt` R package [10]. MML assumes a fixed, known ability distribution and its statistical consistency guarantees require $N \rightarrow \infty$ with J fixed, which is inverted in AI benchmarking. **MCMC** performs full Bayesian inference, sampling from the joint posterior over item parameters and examinee abilities. We used Stan's Hamiltonian Monte Carlo via the `brms` R package [9]. MCMC iteratively samples parameters from the posterior distribution with computational demand scaling in N , J and chain length. **VI** approximates the Bayesian posterior with a factorized variational distribution, optimized via KL divergence minimization. We used the implementation in the Python package, `py-irt` [33]. VI is computationally attractive for larger response matrices as it replaces the posterior sampling with gradient-based optimization on the approximate distribution. However, it inherits the prior assumptions of the Bayesian model and adds approximation error from the chosen variational family. **PSN** is a neural-network estimator that learns

ability and item parameter embeddings, which are passed through an IRT response function to predict binary responses, trained with gradient-based optimization [67]. Unlike the other estimators, PSN does not impose a parametric distribution on abilities, but it does not directly provide uncertainty quantification or consistency guarantees for its estimates.

4.3 Evaluation metrics

The evaluation covers computational feasibility and the three IRT use cases identified in Sec. 3: latent capability scoring, item analysis, and efficient benchmarking, resulting in four corresponding subsections in Sec. 5.

Computational feasibility. Computational feasibility was summarized by *failure rate* (the proportion of replications that failed to produce parameter estimates) and per-replication *runtime* for completed runs. For MML-EM, failures also included non-convergence when the algorithm terminated without satisfying the convergence criterion after 2,000 cycles.

Latent capability scoring. For model capability, *ranking recovery* was evaluated using Kendall’s rank (τ) correlation between true and estimated capability values. Additionally, we measured *aggregate score error*: the absolute difference between expected full-benchmark scores under true and estimated parameters, $\left| \frac{1}{J} \sum_{j=1}^J \pi_{ij} - \frac{1}{J} \sum_{j=1}^J \hat{\pi}_{ij} \right|$, averaged across models, where $\pi_{ij} = P(Y_{ij} = 1 \mid \theta_i, \zeta_j)$ and $\hat{\pi}_{ij} = P(Y_{ij} = 1 \mid \hat{\theta}_i, \hat{\zeta}_j)$ denote model i ’s correct response probability on item j based on true and estimated parameters, respectively.

Item analysis. Recovery of item-level properties was evaluated using Pearson correlations between true and estimated item parameters, including *difficulty recovery* and *discrimination recovery* (2/3PL), and the average absolute error in estimated guessing probability (3PL), i.e., $\frac{1}{J} \sum_{j=1}^J |\hat{c}_j - c_j|$. Similar to aggregate score error, we measured *item mean error* as absolute deviation between item-level correct response probabilities computed from estimated and true parameters, $\left| \frac{1}{N} \sum_{i=1}^N \pi_{ij} - \frac{1}{N} \sum_{i=1}^N \hat{\pi}_{ij} \right|$, averaged across items.

Efficient benchmarking. To assess IRT’s effectiveness in AI benchmark compression, Fisher information $\mathcal{F}$ [40] was computed for each item j given each model’s estimated capability value $\hat{\theta}_i$: $\hat{\mathcal{F}}_j(\hat{\theta}_i) = \hat{a}_j^2 \cdot \frac{(\hat{\pi}_{ij} - \hat{c}_j)^2}{(1 - \hat{c}_j)^2} \cdot \frac{1 - \hat{\pi}_{ij}}{\hat{\pi}_{ij}}$. For 1PL and 2PL, $\hat{c}_j = 0$, and for 1PL, $\hat{a}_j = 1$. The top 10%, 20%, and 100 items with largest Fisher information averaged across $\hat{\theta}_i$ formed the short benchmark. Using 4,000 held-out models not used for IRT estimation, *short-form ranking recovery* was computed as Kendall correlation between average scores on the short benchmark and the full benchmark. This measures whether IRT-based item selection can support benchmark compression while preserving full-benchmark model rankings.

5 Results

5.1 Which IRT estimators are computationally feasible?

Across all conditions and replications, MML-EM exhibited the highest failure rate (69.45%). Specifically, for HELLASWAG and MMLU with $J > 5,000$, MML-EM consistently ran out of memory on 32GB CPU RAM. The out-of-memory error persisted on one tested replication with 128GB CPU RAM with HELLASWAG, 1PL, $N = 30$, rendering it infeasible for large-scale benchmarks. 3PL conditions consistently showed estimation or convergence failures. For smaller benchmarks under 1PL and 2PL, MML-EM often ran into convergence failures at small sample sizes. The distributions of MML-EM and MCMC exit status are detailed in App. C.

For MCMC, estimation became computationally prohibitive for the larger sample sizes and benchmarks: With a 72-hour limit, no replications finished for MMLU when $N = 1000$ (under any IRT model) and $N = 400$ (3PL), and for HELLASWAG when $N = 1000$ (3PL). 5.8% of replications under other MMLU and HELLASWAG conditions also hit the 72-hour limit. Less than 0.1% of the finished replications within 72 hours failed due to algorithm divergence.

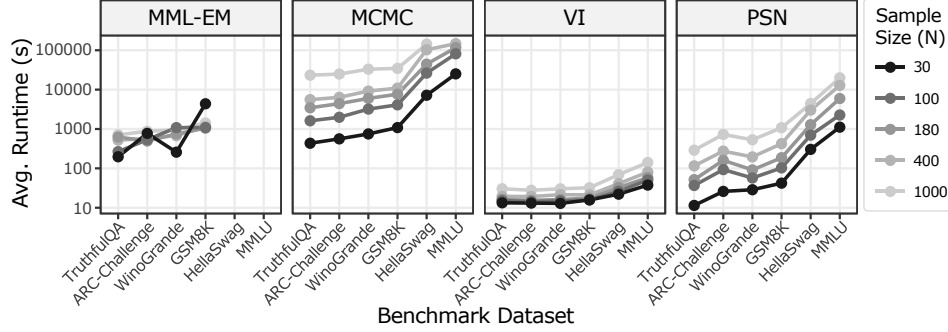


Figure 2: Per-replication runtime (in seconds, log scale) averaged over 1PL, 2PL, & 3PL, across IRT estimators, benchmarks, and sample sizes. Missing points for MML-EM and MCMC on HELLASWAG ($J=5,711$) and MMLU ($J=12,508$) indicate computational infeasibility.

VI had an overall failure rate of 10.71%. All failures were estimation algorithm failures, primarily in the 3PL conditions (27.56% of replications), with slightly higher rates for larger benchmarks. PSN achieved 0% failure across all conditions.

Runtime differed substantially across estimators (Fig. 2). As 1PL, 2PL, and 3PL did not differ meaningfully in runtime within each estimator, we report averages across IRT models and replications. Note that these runtimes reflect practical feasibility under the hardware configurations tested rather than a controlled comparison of computational efficiency: MML-EM, MCMC, and VI were run on CPU, while PSN was run primarily on a single A100 GPU, with a small number of replications on H100 and V100. VI was the fastest estimator with minimal sensitivity to sample size and benchmark length; MCMC was the slowest, often exceeding 10,000 seconds per replication; and PSN runtime scaled with both benchmark length and sample size, growing steeply for large benchmarks; while manageable in our conditions, its scalability to very large benchmarks and samples warrants caution.

5.2 How do different IRT estimators recover model capabilities and rankings?

Ranking recovery (Kendall’s τ) fell below .60 in some conditions, with variation depending primarily on the specific benchmark–IRT model combination. Within a benchmark–IRT model combination, there was comparatively little effect of sample sizes or estimator. Recall that each benchmark–IRT model combination shared the same true item parameters and capability distribution. To understand what specific item or capability distribution characteristics predicted poor ranking recovery, we fitted a multilevel regression model predicting Fisher-transformed Kendall’s τ between true and estimated capabilities. The regression model included four continuous predictors: (i) absolute sample skewness statistic of the true capabilities [29]; (ii) insufficient difficulty coverage, defined as the proportion of evaluated models with 5 or fewer items whose difficulties were within 0.2 of their true capabilities; (iii) marginal reliability, where a lower value implies limited item discrimination, jointly reflecting difficulty mismatch and low item discrimination; and (iv) sample size, $\log(N)$. Continuous predictors were standardized, and estimator was included as a categorical predictor. Detailed definitions, model specifications, and results are provided in App. D; the full model had $R^2 = .798$.

As shown in Fig. 3 (left), *insufficient difficulty coverage was the strongest predictor of poor ranking recovery* (coefficient = $-.280$). Adding it to a base model with only estimator and log sample size as covariates increased marginal R^2 from .039 to .760. *Absolute capability skewness was also strongly predictive when added separately to the base model*, increasing marginal R^2 to .437. Capability skewness was moderately correlated with insufficient difficulty coverage (Kendall’s $\tau = .40$) and retained incremental predictive value for poor ranking (coefficient = $-.103$) after accounting for insufficient coverage and the other predictors. Fig. 3 (right) also shows how ranking recovery declines as the true capability distribution becomes more skewed. Marginal reliability had a small positive association with ranking recovery but had limited incremental predictive value.

Comparing estimators, MCMC and VI performed above the estimator mean, whereas PSN performed below it. However, the effects of sample size and estimator were both relatively small, suggesting that when too few items target the difficulty range where the evaluated models differ in capability, or when

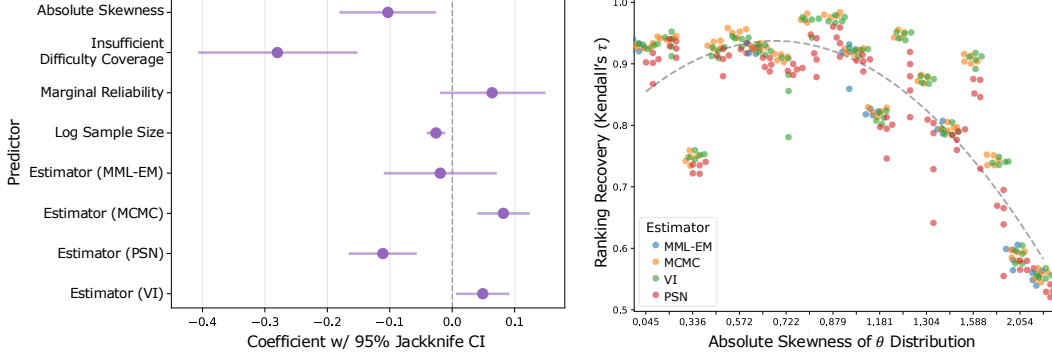


Figure 3: Factors (predictors) associated with ranking recovery. Left: Regression coefficients with 95% jackknife CIs with intercept omitted. Right: Ranking recovery versus absolute capability skewness; points represent simulation conditions, colored by estimator, with sample sizes not visually distinguished. Dashed lines indicate zero (left) and a quadratic fit (right).

true capability distribution is irregular, *increasing sample size or choosing a particular estimator does not substantially improve ranking.*

Poor recovery of rankings does not necessarily imply practically consequential errors in estimated performance. Exact rankings may be less meaningful when models have nearly identical capabilities. In these settings, small estimation errors can reverse model ordering without materially affecting estimated performance. Consistent with tinyBenchmarks [47], aggregate score error (Fig. 16) remained below .025 across most conditions, including skewed conditions such as HELLASWAG and TRUTHFULQA where ranking recovery degraded sharply.

Nevertheless, rankings matter when benchmark results support comparative decisions, such as selecting a model for deployment [30] or determining whether a new model improves on a baseline [23]. This is particularly relevant for general-purpose benchmarks whose results are interpreted as evidence of broader capabilities and guide downstream model selection [52]. Moreover, a commonly stated benefit of IRT in the studies reviewed in Sec. 3 is that accounting for item difficulty and discrimination can better distinguish models with similar aggregate scores. Our ranking analysis therefore cautions against capability comparisons from the ordering of point estimates alone, especially when items provide limited information for distinguishing models and capability distribution is skewed.

5.3 How do different IRT estimators recover item and benchmark characteristics?

Here, we present results under 2PL as the commonly adopted IRT specification in AI benchmarking literature. Fig. 4 presents the difficulty recovery, discrimination recovery, and item mean error across estimators, benchmarks, and sample sizes. Full results for 1PL and 3PL are reported in App. E.2.

Reliable estimation of item parameters depended heavily on sample size, i.e., the number of evaluated models: $N = 30$ (the darkest lines) was insufficient for item parameter recovery under any estimator. At this sample size, difficulty recovery dropped below 0.50 (e.g., MML-EM on ARC-CHALLENGE and PSN on WINOGRANDE) and discrimination recovery below 0.60 (e.g., MML-EM and PSN on GSM8K). Item mean error was also substantially higher compared to all larger sample sizes. A marked improvement occurred at $N = 100$, though further increasing sample size led to incremental gains. This suggests that the small model counts ($N \approx 30$) commonly observed in AI benchmarking (Sec. 3) could risk item-level inference reliability.

A remaining question is whether $N = 30$ would suffice for reliable item parameter recovery under more regular capability distributions. We conducted two supplementary analyses: (i) a comparison of $N = 30$ replications with near-normal empirical distributions of true capabilities against their $N = 100$ counterparts, and (ii) controlled 3PL simulations on ARC-CHALLENGE, using capabilities drawn from a standard normal distribution or resampled from the empirical capabilities within $[-1, 1]$. The parameter recovery gap between $N = 30$ and $N = 100$ persisted in both analyses. Even under normal capability distributions, the examined estimators showed poor item-parameter recovery at $N = 30$, with clear improvements at $N = 100$. More details are provided in App. E.3.

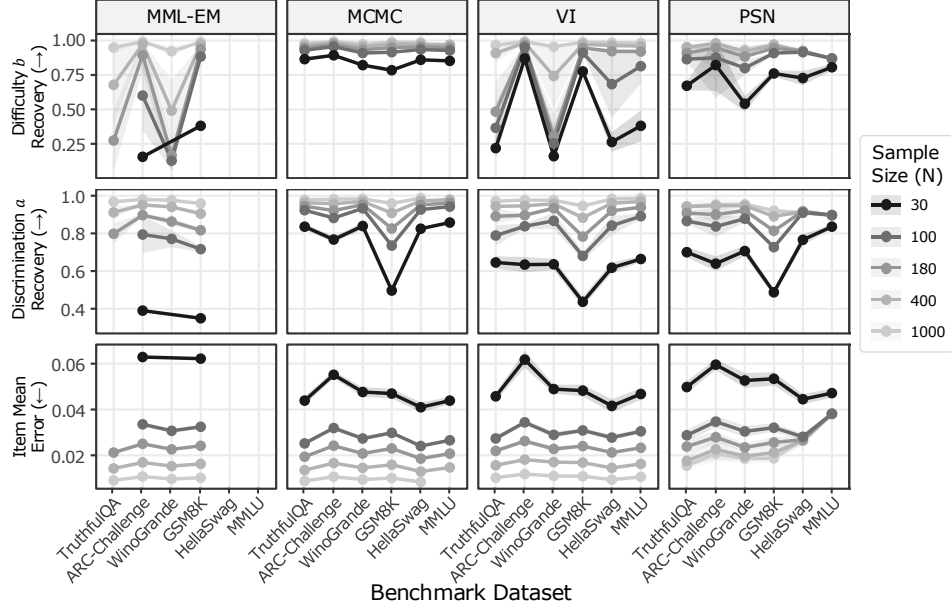


Figure 4: Item parameter recovery under 2PL across estimators, benchmarks, and sample sizes. Higher is better for b and a recovery; lower is better for item mean error. Shaded regions indicate ± 1 standard deviation across 50 replications. Missing points indicate computational infeasibility.

Among the IRT estimators, MCMC generally achieved the most reliable item parameter recovery when computationally feasible. VI, despite being one of the most widely used estimators in recent IRT-based AI evaluation studies, showed notably unreliable difficulty recovery in several conditions: For 2PL, its difficulty recovery was below .50 for $N \leq 180$ on TRUTHFULQA and GSM8K, indicating that VI-estimated item difficulties may diverge substantially from true item difficulty. Similarly, MML-EM, the conventional approach to IRT estimation in human testing, produced unreliable estimates for small samples and certain benchmarks, even when it converged. PSN, in general, showed more reliable item parameter recovery than MML-EM and VI, with slightly lower accuracy and higher variability across replications than MCMC in some conditions.

5.4 Can estimated IRT parameters support efficient benchmarking?

Finally, we examined whether short benchmarks, constructed based on expected Fisher information using estimated item parameters, could accurately predict aggregate score on the full test for held-out samples. We evaluated subsets containing 10% or 20% of the full benchmark items, or a fixed length of 100 items. Each was compared with a random-selection baseline averaged over 30 item subsets of the same size. Across these subset sizes, IRT-based short benchmark generally outperformed the random baseline across benchmarks and estimators, with pronounced improvements under 1PL but only marginal improvements under 2PL and 3PL (Fig. 22). For 1PL, larger samples ($N \geq 100$) improved short-form ranking recovery, as shown in Figs. 23 & 24.

Fig. 5 shows the ranking recovery under 2PL for the 10% item subsets. Unlike full-benchmark ranking recovery (Fig. 3), short-form ranking recovery showed no monotonic relationship with the skewness of capability distribution. To assess whether the modest gains under 2PL were due to the small subset sizes, we also examined larger subsets across all six benchmarks, including subsets of up to

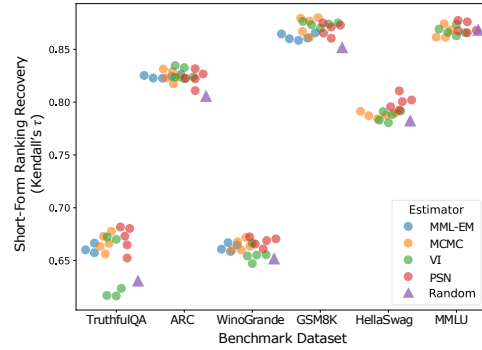


Figure 5: Short-form ranking recovery across benchmarks and estimators (2PL). Each point represents one simulation condition.

2,500 items for MMLU. IRT-based short benchmarks continued to perform similarly to the random baselines. Detailed results for all short benchmarks are provided in App. F.

6 Practical recommendations and implications

Our results suggest that IRT-based AI benchmark analysis requires different safeguards depending on the intended use of the fitted model. For estimator choice, MML-EM is the conventional approach in human testing, but in our simulations it became computationally burdensome for large-item benchmarks and did not always yield reliable estimates even when runs converged. MCMC provided the most consistent parameter recovery, but its computational cost limited its feasibility for combinations of large benchmarks, large sample sizes, and more complex IRT models. VI was fast and recovered model rankings slightly better, but was found unreliable sometimes for aggregate score recovery and item-parameter recovery. PSN-IRT remained computationally manageable and showed competitive performance in many settings (except with small sample sizes or in ranking recovery), suggesting that neural estimators may be useful for very large benchmarks.

Sample size requirements depend on the inference target. Item analysis claims are less reliable for small model pools. *Regardless of capability distribution, $N = 30$ was insufficient for recovering item difficulty, discrimination, or mean scores for any estimator examined*, and recovery improved clearly at $N \geq 100$. This finding cautions against use of item parameters to characterize item difficulty, informativeness or benchmark quality when sample size is very small. Model-level aggregate score was less sensitive to sample size when benchmarks were long, as averaging across many items stabilized expected scores. Thus, small model pools may support aggregate score prediction but could be insufficient for item-quality analysis or benchmark auditing.

Model ranking recovery was strongly related to difficulty coverage and capability-distribution skewness, with poorer recovery when few items had difficulties near the evaluated models’ capabilities or when capabilities were highly skewed. *Larger sample sizes and alternative estimators did not resolve these ranking challenges*. For ranking claims, researchers may consider running simulation checks using their estimated capability and item parameters as a true data-generating model to assess whether rankings are recoverable under the observed sample size, item pool and capability distribution.

7 Limitations and Conclusion

This paper investigated the reliability of IRT when applied under the data regimes typical of AI benchmarking. Through a structured review of existing studies, we identified substantial heterogeneity in IRT model choices, estimation methods, and data conditions, with many studies operating in regimes far from those for which IRT was originally validated. Our simulation study, spanning 360 conditions and 18,000 runs across six benchmarks, three IRT models, five sample sizes, and four estimators, revealed several findings with practical implications.

Several limitations remain. Computational feasibility was not assessed on uniform hardware, and convergence monitoring was limited to MML-EM. We also did not examine sample sizes between 30 and 100, where item parameter recovery improved substantially, or models beyond unidimensional 1, 2, and 3PL, or estimators beyond the four evaluated. The conditions under which IRT-based selection improves short-form ranking recovery also require further study. Capability skewness and insufficient difficulty coverage were also moderately correlated in the conditions calibrated from the OpenLLM data, which did not allow fully isolating their effects on ranking recovery.

Our simulations were calibrated to six OpenLLM Leaderboard benchmarks, with conditions informed by the reviewed studies. This empirical grounding supports their relevance to similar applications, but estimator comparisons and sample-size guidance remain conditional on the settings examined. Therefore, **when assessing the reliability of IRT-based inferences in a particular application, we suggest researchers to evaluate recovery through simulations based on their own data**: Specifically, they can use estimated item and capability parameters to simulate responses under the same sample size and IRT model, refit the model, and evaluate recovery of the quantities of interest. This provides a best-case assessment of reliability of inferences assuming the fitted IRT model is correctly specified. We release our simulation and evaluation source code to facilitate this practice.

Acknowledgments and Disclosure of Funding

This work was supported by Schmidt Sciences.

References

- [1] J. H. Albert. Bayesian estimation of normal ogive item response curves using gibbs sampling. *Journal of Educational Statistics*, 17(3):251–269, 1992.
- [2] J. H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.
- [3] M. A. Barton and F. M. Lord. An upper asymptote for the three-parameter logistic item-response model. *ETS Research Report Series*, 1981(1):i–8, 1981.
- [4] T. M. Bechger, G. Maris, H. H. F. M. Verstralen, and A. A. B’eguin. Using classical test theory in combination with item response theory. *Applied Psychological Measurement*, 27(5):319–334, 2003.
- [5] R. D. Bock and M. Aitkin. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46(4):443–459, 1981.
- [6] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Muniyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang. On the opportunities and risks of foundation models, 2022.
- [7] P.-C. Bürkner. Bayesian item response modeling in r with brms and stan. *Journal of statistical software*, 100:1–54, 2021.
- [8] M. Byrd and S. Srivastava. Predicting difficulty and discrimination of natural language questions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 119–130, Stroudsburg, PA, USA, 2022. Association for Computational Linguistics.
- [9] P.-C. Bürkner. brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1):1–28, 2017.
- [10] R. P. Chalmers. mirt: A multidimensional item response theory package for the r environment. *Journal of Statistical Software*, 48(6):1–29, 2012.
- [11] J. Chen, C. Wang, G. Zhang, P. Ye, L. Bai, W. Hu, Y. Qu, and S. Hu. Learning compact representations of LLM abilities via item response theory. *arXiv [cs.AI]*, Oct. 2025.
- [12] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018.
- [13] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training verifiers to solve math word problems, 2021.
- [14] R. J. De Ayala. *The theory and practice of item response theory*. Guilford Publications, 2013.

- [15] C. DeMars. *Item Response Theory*. Understanding Statistics: Measurement. Oxford University Press, New York, NY, 2010.
- [16] İ. E. Deveci and D. Ataman. The ouroboros of benchmarking: Reasoning evaluation in an era of saturation. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2025.
- [17] K. Ethayarajh and D. Jurafsky. Utility is in the eye of the user: A critique of NLP leaderboards. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online, Nov. 2020. Association for Computational Linguistics.
- [18] S. Frick, A. Krivosija, and A. Munteanu. Scalable learning of item response theory models. In S. Dasgupta, S. Mandt, and Y. Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 1234–1242. PMLR, 02–04 May 2024.
- [19] A. Gill, A. Ravichander, and A. Marasovic. What has been lost with synthetic evaluation? In C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9902–9945, Suzhou, China, Nov. 2025. Association for Computational Linguistics.
- [20] H. Gulliksen. *Theory of Mental Tests*. Wiley Publications in Psychology. John Wiley & Sons, Hoboken, NJ, 1950.
- [21] R. K. Hambleton and R. W. Jones. Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3):38–47, 1993.
- [22] R. K. Hambleton, H. Swaminathan, and H. J. Rogers. *Fundamentals of Item Response Theory*, volume 2 of *Measurement Methods for the Social Sciences*. Sage Publications, Thousand Oaks, CA, 1991.
- [23] D. Heineman, V. Hofmann, I. Magnusson, Y. Gu, N. A. Smith, H. Hajishirzi, K. Lo, and J. Dodge. Signal and noise: A framework for reducing uncertainty in language model evaluation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.
- [24] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
- [25] V. Hofmann, D. Heineman, I. Magnusson, K. Lo, J. Dodge, M. Sap, P. W. Koh, C. Wang, H. Hajishirzi, and N. A. Smith. Fluid language model benchmarking. In *Second Conference on Language Modeling*, 2025.
- [26] H. Jiang, X. Yi, Z. Wei, Z. Xiao, S. Wang, and X. Xie. Raising the bar: Investigating the values of large language models via generative evolving testing. In *Forty-second International Conference on Machine Learning*, June 2025.
- [27] H. Jiang, S. Zhang, X. Yi, X. Xie, and Z. Xiao. Position: Science of ai evaluation requires item-level benchmark data, 2026.
- [28] S. Jiang, C. Wang, and D. J. Weiss. Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in Psychology*, Volume 7 - 2016, 2016.
- [29] D. N. Joanes and C. A. Gill. Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 47(1):183–189, 1998.
- [30] M. Jung, M. Lee, Y. Kim, S. Choi, and M. Kahng. Who defines “best”? towards interactive, user-defined evaluation of llm leaderboards. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency*, pages 4733–4755, 2026.

- [31] A. Kipnis, K. Voudouris, L. M. Schulze Buschoff, and E. Schulz. metabench - a sparse benchmark of reasoning and knowledge in large language models. In *The Thirteenth International Conference on Learning Representations*, Oct. 2024.
- [32] L. Kirisci, T.-c. Hsu, and L. Yu. Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement*, 25(2):146–162, 2001.
- [33] J. P. Lalor and P. Rodriguez. py-irt: A scalable item response theory library for python. *INFORMS J. on Computing*, 35(1):5–13, Jan. 2023.
- [34] J. P. Lalor, H. Wu, T. Munkhdalai, and H. Yu. Understanding deep learning performance through an examination of test set difficulty: A psychometric case study. *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2018:4711–4716, Oct. 2018.
- [35] J. P. Lalor, H. Wu, and H. Yu. Building an evaluation scale using item response theory. In J. Su, K. Duh, and X. Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 648–657, Austin, Texas, Nov. 2016. Association for Computational Linguistics.
- [36] J. P. Lalor, H. Wu, and H. Yu. Learning latent parameters without human response patterns: Item response theory with artificial crowds. *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2019:4240–4250, Nov. 2019.
- [37] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. Cosgrove, C. D. Manning, C. Re, D. Acosta-Navas, D. A. Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. WANG, K. Santhanam, L. Orr, L. Zheng, M. Yuksekgonul, M. Suzgun, N. Kim, N. Guha, N. S. Chatterji, O. Khattab, P. Henderson, Q. Huang, R. A. Chi, S. M. Xie, S. Santurkar, S. Ganguli, T. Hashimoto, T. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang, and Y. Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. Featured Certification, Expert Certification, Outstanding Certification.
- [38] L. Liao, Q. Zhang, R. Wu, and G. Fang. Toward a unified framework for data-efficient evaluation of large language models. *arXiv [cs.AI]*, Oct. 2025.
- [39] S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring how models mimic human falsehoods. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [40] F. M. Lord. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1980.
- [41] F. M. Lord and M. R. Novick. *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA, 1968.
- [42] F. M. Lord, M. R. Novick, and A. Birnbaum. Some latent trait models and their use in inferring an examinee’s ability. 1966.
- [43] F. Martínez-Plumed, R. B. C. Prudêncio, A. Martínez-Usó, and J. Hernández-Orallo. Item response theory in AI: Analysing machine learning classifiers at the instance level. *Artif. Intell.*, 271:18–42, June 2019.
- [44] T. P. Morris, I. R. White, and M. J. Crowther. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102, 2019.
- [45] A. Myrzakhan, S. M. Bsharat, and Z. Shen. Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena, 2024.
- [46] S. Ott, A. Barbosa-Silva, K. Blagec, J. Brauner, and M. Samwald. Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications*, 13(1):6793, 2022.

- [47] F. M. Polo, L. Weber, L. Choshen, Y. Sun, G. Xu, and M. Yurochkin. tinyBenchmarks: evaluating LLMs with fewer examples. In *International Conference on Machine Learning*, pages 34303–34326. PMLR, July 2024.
- [48] G. Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests*. Nielsen & Lydiche, Copenhagen, Denmark, 1960.
- [49] Z. Robertson. Identity-link IRT for label-free LLM evaluation: Preserving additivity in TVD-MI scores. *arXiv [cs.LG]*, Oct. 2025.
- [50] P. Rodriguez, J. Barrow, A. Hoyle, J. P. Lalor, R. Jia, and J. Boyd-Graber. Evaluation examples are not equally informative: How should that change NLP leaderboards? In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *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, Online, Aug. 2021. Association for Computational Linguistics.
- [51] K. Sakaguchi, R. L. Bras, C. Bhagavatula, and Y. Choi. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106, Aug. 2021.
- [52] O. Salaudeen, A. Reuel, A. Ahmed, S. Bedi, Z. Robertson, S. Sundar, B. Domingue, A. Wang, and S. Koyejo. Measurement to meaning: A validity-centered framework for ai evaluation. *arXiv preprint arXiv:2505.10573*, 2025.
- [53] F. Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 34(4, Pt. 2):1–97, 1969. Psychometrika Monograph Supplement No. 17.
- [54] M. Schilling-Wilhelmi, N. Alampara, and K. M. Jablonka. Lifting the benchmark iceberg with item-response theory. In *AI for Accelerated Materials Design - ICLR 2025*, 2025.
- [55] U. Schroeders and T. Gnambs. Sample-size planning in item-response theory: A tutorial. *Advances in Methods and Practices in Psychological Science*, 8(1), 2025.
- [56] S. Sen and A. S. Cohen. The impact of sample size and various other factors on estimation of dichotomous mixture IRT models. *Educ. Psychol. Meas.*, 83(3):520–555, June 2023.
- [57] B. S. Siepe, F. Bartoš, T. P. Morris, A.-L. Boulesteix, D. W. Heck, and S. Pawel. Simulation studies for methodological research in psychology: A standardized template for planning, preregistration, and reporting. *Psychological Methods*, 2024.
- [58] C. Siska, K. Marazopoulou, M. Ailem, and J. Bono. Examining the robustness of LLM evaluation to the distributional assumptions of benchmarks. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10406–10421, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics.
- [59] C. A. Stone. Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement*, 16(1):1–16, 1992.
- [60] C. Touts, R. Bommasani, K. Creel, S. Bana, D. Jurafsky, and P. S. Liang. Ecosystem-level analysis of deployed machine learning reveals homogeneous outcomes. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 51178–51201. Curran Associates, Inc., 2023.
- [61] S. T. Truong, Y. Tu, P. Liang, B. Li, and S. Koyejo. Reliable and efficient amortized model-based evaluation. In *Forty-second International Conference on Machine Learning*, June 2025.
- [62] C. Vania, P. M. Htut, W. Huang, D. Mungra, R. Y. Pang, J. Phang, H. Liu, K. Cho, and S. R. Bowman. Comparing test sets with item response theory. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *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 1141–1158, Online, Aug. 2021. Association for Computational Linguistics.

- [63] C. M. Woods and D. Thissen. Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika*, 71(2):281–301, 2006.
- [64] M. Wu, R. L. Davis, B. W. Domingue, C. Piech, and N. Goodman. Variational item response theory: Fast, accurate, and expressive, 2020.
- [65] Z. Xu, J. Liu, Y. Wang, and Y. Gu. Latency-response theory model: Evaluating large language models via response accuracy and chain-of-thought length. *arXiv [stat.ME]*, Dec. 2025.
- [66] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. HellaSwag: Can a machine really finish your sentence? In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics.
- [67] H. Zhou, H. Huang, Z. Zhao, L. Han, H. Wang, K. Chen, M. Yang, W. Bao, J. Dong, B. Xu, C. Zhu, H. Cao, and T. Zhao. Lost in benchmarks? rethinking large language model benchmarking with item response theory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(41):35085–35093, Mar. 2026.
- [68] Y. Zhuang, Q. Liu, Y. Ning, W. Huang, R. Lv, Z. Huang, G. Zhao, Z. Zhang, Q. Mao, S. Wang, and E. Chen. Efficiently measuring the cognitive ability of LLMs: An adaptive testing perspective. <https://openreview.net/forum?id=s6X3s3rBPW>, Oct. 2023. Accessed: 2026-4-16.

A Estimator introduction

MML-EM. Marginal maximum likelihood (MML) is the conventional approach to IRT parameter estimation for human testing. It treats latent traits as random draws from a population distribution, often standard normal for computational convenience, and estimates item parameters, $(\zeta_j, j = 1, \dots, J)$ by maximizing the marginal likelihood given the response matrix:

$$L_{MML}(\zeta) = \prod_{i=1}^N \int \left[\prod_{j=1}^J p(Y_{ij} \mid \theta, \zeta_j) \right] f(\theta) d\theta. \quad (1)$$

Optimization is done with the EM algorithm. We used the `mirt` package and adopted the stochastic variant, the Metropolis-Hastings Robbins-Monro algorithm, replacing fixed quadrature with Monte Carlo integration when standard EM is too costly for large item pools. Statistical consistency theory guarantees the MML estimator to recover true item parameters when the number of items (J) is fixed and sample size (N) grows to infinity. Although in human testing, often the assumption of trait normality is empirically plausible and $J < N$, MML’s behavior is less certain in AI benchmark regimes, where the number of evaluated systems may be small, the number of items large, and the capability distribution nonnormal.

MCMC. Bayesian MCMC specifies prior distributions for item parameters and model capabilities, and then samples from the posterior distribution after observing the response data ($\mathbf{Y}$),

$$p(\boldsymbol{\theta}, \boldsymbol{\zeta} \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \boldsymbol{\theta}, \boldsymbol{\zeta}) p(\boldsymbol{\theta}, \boldsymbol{\zeta}). \quad (2)$$

We implemented Bayesian estimation in `brms`, which uses Stan’s Hamiltonian Monte Carlo sampler. HMC uses gradient information to reduce random-walk behavior, although sampling efficiency still depends on posterior geometry and model size. Prior settings and implementation details are reported in Appendix A. Bayesian posterior estimates can be sensitive to prior specification when empirical information is limited, especially when few evaluated systems are available for estimating many item parameters. Further, runtime grows with both sample size and benchmark length as each MCMC iteration evaluates the full likelihood over all model and item parameters.

The prior specifications for the MCMC algorithm in this study followed [7] and are summarized in Table 1. For each model, the MCMC algorithm was run using a single chain with 2,000 total iterations, where the first 1,000 iterations were discarded as burn-in.

Variational inference. VI in this study was fitted with `py-irt`, which approximates the Bayesian posterior in Eq. (2) with a tractable variational distribution, e.g., for the 1PL model,

$$q_\phi(\boldsymbol{\theta}, \mathbf{b}, \mu, \tau) = q(\mu) q(\tau) \prod_{i=1}^N q(\theta_i) \prod_{j=1}^J q(b_j), \quad (3)$$

where μ and τ are the location and precision parameters of the normal latent capability. The variational parameters are estimated by minimizing $D_{KL}(q_\phi \parallel p)$, the KL divergence between (3) and (2). VI is computationally attractive for long benchmark matrices as it replaces numerical integration or posterior sampling with gradient-based optimization. However, it inherits the prior assumptions of the Bayesian model and adds approximation error from the chosen variational family, which can affect inferences in small-sample AI benchmark settings and when capability distributions is non-normal.

PSN-IRT. PSN [67] is a neural-network representation of an IRT model, which assigns each evaluated model a learned capability embedding (θ_i) and each benchmark item item-parameter embeddings (a_j, b_j, c_j). These embeddings are passed through an item response function layer to predict binary response. We adapted the authors’ original 4PL implementation to the 1 - 3PL models used in our simulation. PSN is computationally efficient with gradient-based optimization and does not assuming a normal distribution for latent capability. However, it is trained as a predictive neural model and does not directly provide uncertainty quantification or parameter consistency guarantees for parameter estimates, and its estimates are specific to the trained network and item set.

Table 1: Prior Specifications for MCMC Algorithm. Normal^+ denotes a half-normal distribution.

IRT Model	Parameter	Prior	Hyperprior
1PL	Ability (θ)	$\theta_i \sim \text{Normal}(0, \sigma_\theta)$	$\sigma_\theta \sim \text{Normal}^+(0, 1)$
	Difficulty (b)	$b_j \sim \text{Normal}(\mu_b, \sigma_b)$	$\sigma_b \sim \text{Normal}^+(0, 3)$
2PL & 3PL	Ability (θ)	$\theta_i \sim \text{Normal}(0, 1)$	$\sigma_\theta = 1$ (Fixed scale)
	Difficulty (b)	$b_j \sim \text{Normal}(\mu_b, \sigma_b)$	$\mu_b \sim \text{Normal}(0, 5)$ $\sigma_b \sim \text{Normal}^+(0, 3)$
	Discrimination (a)	$\log a_j \sim \text{Normal}(\mu_{\log a}, \sigma_{\log a})$	$\mu_{\log a} \sim \text{Normal}(0, 1)$ $\sigma_{\log a} \sim \text{Normal}^+(0, 1)$
3PL	Pseudo-Guessing (c)	$\text{logit } c_j \sim \text{Normal}(\mu_{\text{logit } c}, \sigma_{\text{logit } c})$	$\mu_{\text{logit } c} \sim \text{Logistic}(0, 1)$ $\sigma_{\text{logit } c} \sim \text{Normal}^+(0, 1)$

B Data-generating parameters

B.1 True capability and difficulty distributions

In Figs. 6–11, we present distributions of true capability (θ) and difficulty (b) parameters for the six benchmarks under 1PL, 2PL, and 3PL models, estimated from the full OpenLLM response matrices via variational inference. All distributions exhibit non-normality, with capability values showing multimodality across all three IRT models.

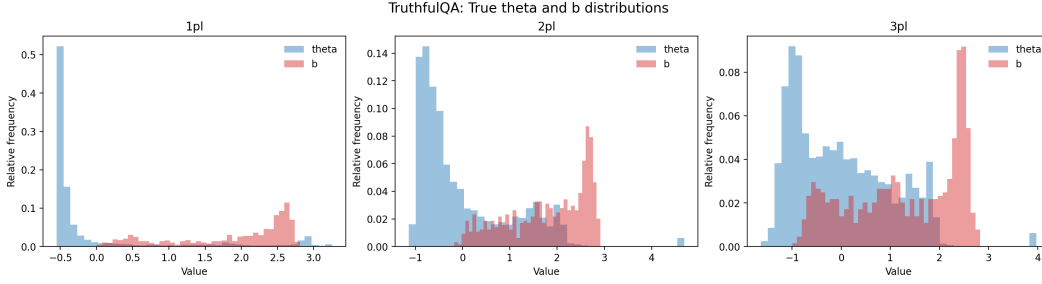


Figure 6: Distributions of true capability (θ) and difficulty (b) parameters for TRUTHFULQA.

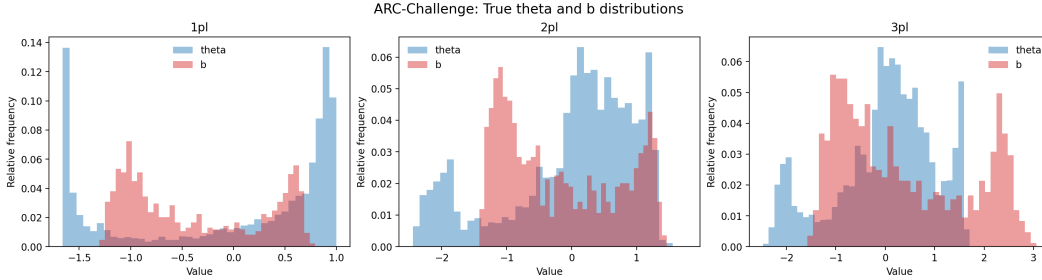


Figure 7: Distributions of true capability (θ) and difficulty (b) parameters for ARC-CHALLENGE.

B.2 Reliability of the data-generating parameters

Our primary simulations introduced in Sec. 4 used parameters estimated by VI from the empirical OpenLLM response matrices as data-generating parameters. These estimates should be understood as pseudo-ground truth rather than verified true parameters. Their suitability raises two related questions:

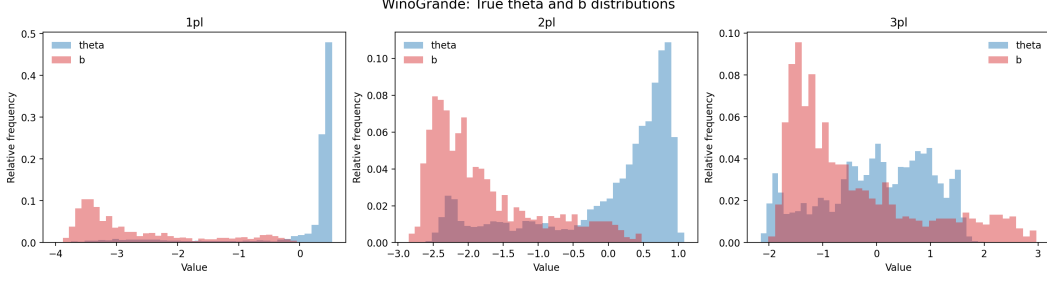


Figure 8: Distributions of true capability (θ) and difficulty (b) parameters for WINOGRANDE.

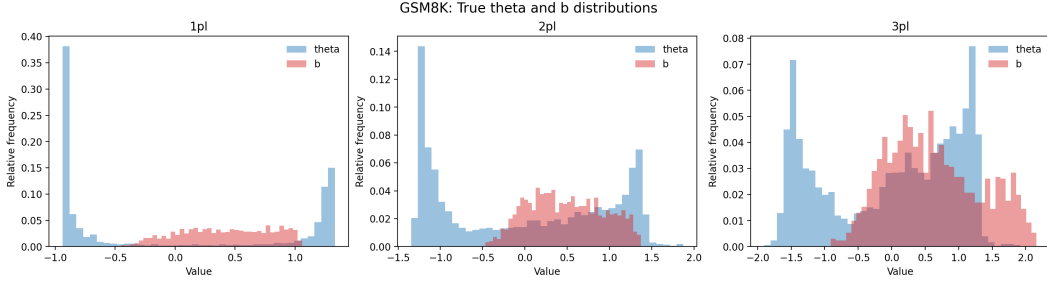


Figure 9: Distributions of true capability (θ) and difficulty (b) parameters for GSM8K.

Whether VI estimation errors could make the simulated data unrealistic. The strongest VI instability in our simulations occurred at small sample sizes, whereas each empirical response matrix used to derive the data-generating parameters contained more than 5,000 models. This is well beyond the small-sample regime in which VI exhibited poor parameter recovery, providing some support for the stability of the data-generating parameters. Exact capability rankings remain a caveat, because irregular capability distributions can make latent ordering difficult to recover even with many models. Nevertheless, the overall shape of the capability distribution is less sensitive than the ordering of closely matched models and was broadly consistent with the distribution of observed aggregate scores. Thus, the empirical fits provide a plausible basis for simulation, although they should not be interpreted as known ground truth.

Whether using VI both to generate the parameters and as a competing estimator could favor VI in the subsequent recovery analysis. After fitting VI, we transformed the parameters so that the data-generating capabilities had mean 0 and standard deviation 1, with the corresponding transformations applied to the item parameters. This removes arbitrary location and scale differences across fitted models. However, rescaling alone cannot rule out estimator-specific bias arising from VI’s dual role in parameter generation and evaluation. Notably, VI did not consistently outperform the other estimators in our primary experiments and often showed poorer item-parameter recovery at small sample sizes, providing no evidence of a general advantage from this dual role.

Therefore, to further assess this possibility, we conducted a sensitivity analysis on ARC-CHALLENGE under 1PL, comparing MCMC, VI, and PSN at $N \in 30, 100, 180, 400$. We fitted PSN to the same empirical response matrix and generated 10 replications per condition using the resulting parameters. We then evaluated difficulty recovery, item mean error, aggregate score error, and ranking recovery, comparing the relative performance of the three estimators under PSN- and VI-derived data-generating parameters.

As shown in Fig. 12, changing the source of the data-generating parameters did not substantially alter the relative performance of MCMC, VI, and PSN across most metrics. Difficulty recovery was the main exception, with PSN performing more favorably relative to VI under PSN-derived parameters. This exception does not reverse our broader finding in Sec. 5.3 that VI can exhibit unreliable item-parameter recovery at small sample sizes. In particular, VI did not gain a consistent advantage when evaluated under VI-derived parameters across the outcomes. Table 2 reports the absolute performance of VI under the two data-generating parameter sources.

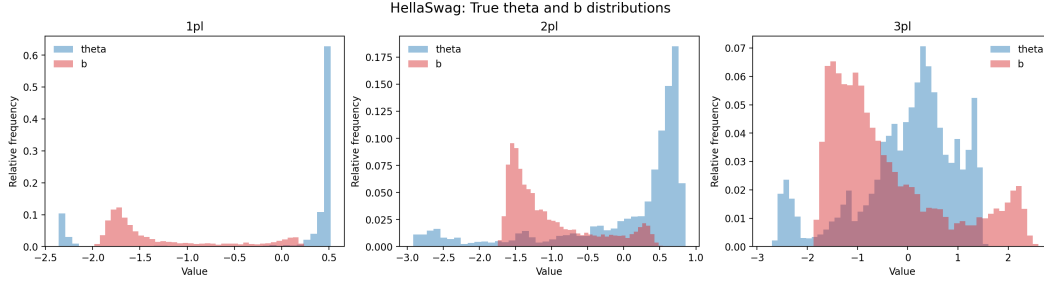


Figure 10: Distributions of true capability (θ) and difficulty (b) parameters for HELLASWAG.

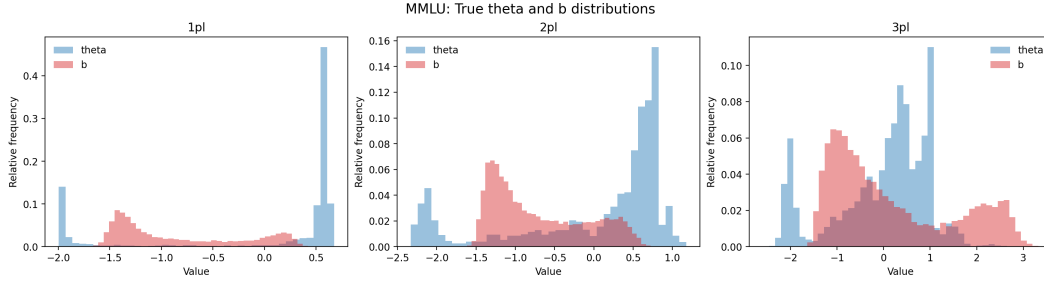


Figure 11: Distributions of true capability (θ) and difficulty (b) parameters for MMLU.

Overall, this sensitivity analysis suggests that our main comparisons are not driven solely by the use of VI-derived data-generating parameters.

C Exit statuses of MML-EM and MCMC

The distributions of the exit statuses of MML-EM and MCMC are displayed in Figs. 13 & 14.

D Failure mode analysis for ranking recovery

To examine the factors associated with the poor ranking recovery in Sec. 5.2, we first fitted a multilevel linear regression model predicting the Fisher-transformed ranking-recovery correlation. All continuous predictors were standardized, such that their coefficients represent the expected change in the outcome associated with a one-standard-deviation increase in the predictor. The model included the following predictors:

- *Absolute skewness* $|\gamma|$, defined in Sec. 5.2 as the absolute sample skewness of the true model capability θ ;
- *Insufficient difficulty coverage*, defined as the proportion of models having ≤ 5 items whose difficulty is in the $\theta \pm .2$ neighborhood;
- *Marginal reliability* [4], with lower values indicating less precise capability measurement due to limited test information, which may reflect difficulty mismatch, low item discrimination, or both;
- *Log sample size*, defined as $\log(N)$; and
- *Estimator*, effect-coded such that each estimator's coefficient reflects its deviation from the mean across all four estimators introduced in Sec. 4.2.

Table 3 reports the regression coefficients. For the four continuous predictors, a larger absolute coefficient indicates a stronger conditional association with ranking recovery per standard-deviation change, holding the other variables fixed. The estimator coefficients instead represent effect-coded deviations from the mean across estimators. This regression model accounted for approximately 80% of variance in ranking recovery ($R^2 = .798$).

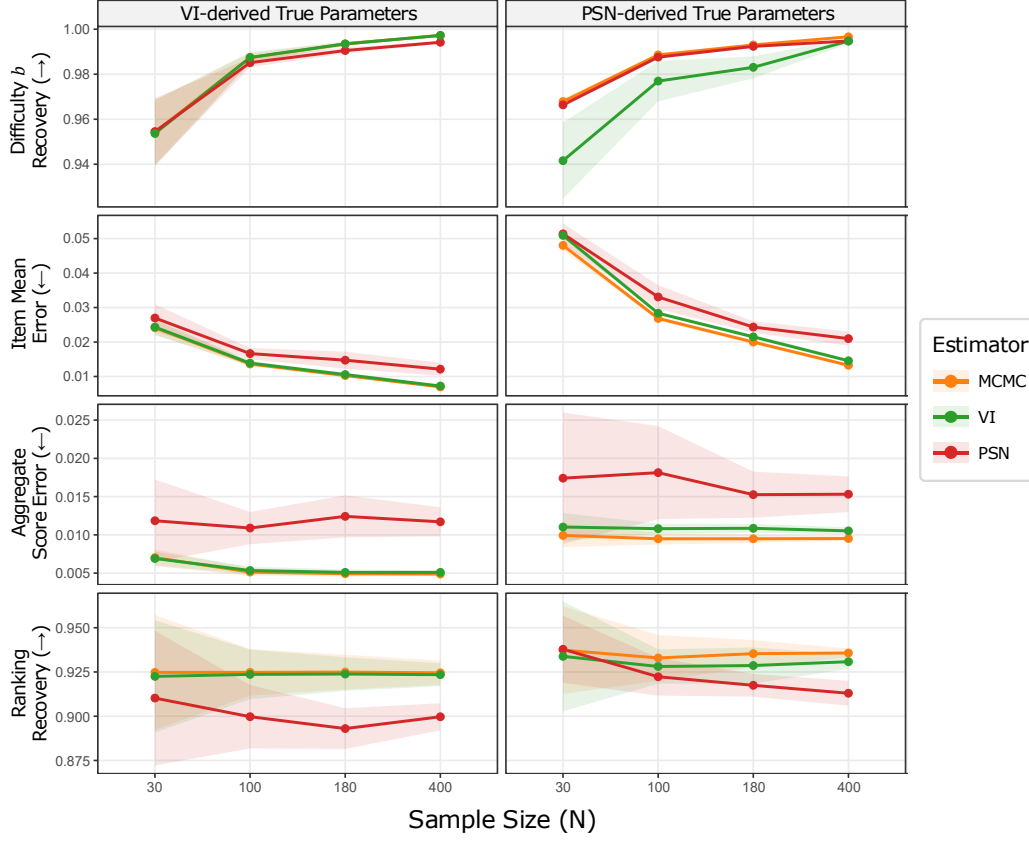


Figure 12: Performance of MCMC, VI, and PSN under VI- and PSN-derived data-generating parameters for ARC-Challenge with IPL.

Fig. 15 presents bivariate scatter plots for the three data-characteristic predictors with larger coefficients. We additionally assessed their incremental predictive value by fitting three regression models that separately added each predictor to a base model containing only estimator and log sample size, and reporting the resulting marginal R^2 in Table 4.

Together, these analyses helped clarify the factors associated with ranking failure. insufficient difficulty coverage was the most predictive factor of poor ranking recovery: adding it to the base model increased the marginal R^2 from .039 to .760. Absolute skewness was the second most predictive. It was moderately correlated with insufficient difficulty coverage (Kendall’s correlation = .40) but also increased incremental predictive power. On the estimation side, larger sample size did not improve ranking recovery and instead showed a small negative association, consistent with Fig. 3. Relative to the mean across estimators, MCMC and VI achieved better ranking recovery, PSN performed worse, and MML-EM showed no clear difference.

E Detailed results for parameter recovery

E.1 Aggregate score error

Please refer to Fig. 16 for detailed results.

E.2 Item parameter recovery

Please refer to Figs. 17 & 18 for detailed results.

Table 2: VI Performance under VI- and PSN-derived data-generating parameters for ARC-Challenge with 1PL. Convergence rates are based on 50 replications per sample size under VI-derived parameters and 10 under PSN-derived parameters. Metric values are means with standard deviations across converged replications in parentheses.

Metric	True Params.	Sample Size			
		$N = 30$	$N = 100$	$N = 180$	$N = 400$
Convergence (%)	VI-derived	96	98	98	100
	PSN-derived	90	90	90	100
Difficulty Recovery ($\uparrow$)	VI-derived	.954 (.015)	.987 (.002)	.993 (.001)	.997 (.0003)
	PSN-derived	.942 (.017)	.977 (.009)	.983 (.005)	.995 (.0012)
Item Mean Error ($\downarrow$)	VI-derived	.0243 (.0023)	.0139 (.0007)	.0106 (.0005)	.0072 (.0002)
	PSN-derived	.0509 (.0012)	.0283 (.0005)	.0215 (.0006)	.0146 (.0005)
Aggregate Score Error ($\downarrow$)	VI-derived	.0069 (.0009)	.0053 (.0005)	.0051 (.0004)	.0051 (.0002)
	PSN-derived	.0110 (.0018)	.0108 (.0007)	.0109 (.0006)	.0105 (.0005)
Ranking Recovery ($\uparrow$)	VI-derived	.923 (.032)	.924 (.014)	.924 (.010)	.923 (.006)
	PSN-derived	.934 (.031)	.928 (.010)	.929 (.010)	.931 (.004)

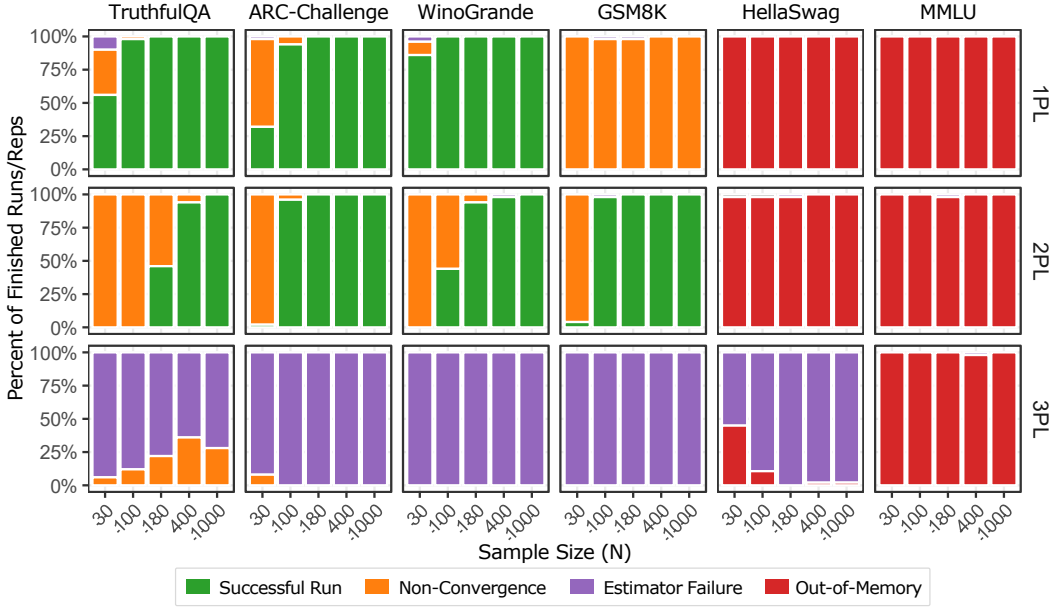


Figure 13: Distribution of MML-EM exit statuses across IRT models, benchmarks, and sample sizes. HELLASWAG and MMLU failed entirely due to out-of-memory errors. For smaller benchmarks, non-convergence was the primary failure mode, particularly at small sample sizes and under 3PL. Success rates improved with increasing sample size for 1PL and 2PL but remained near zero for 3PL across all conditions.

E.3 Disentangling sample size from distributional skewness

As discussed in Sec. 2.1, the empirical capability distributions observed in AI evaluation context are frequently skewed, multimodal, or otherwise irregular. Because these properties may themselves impair IRT estimation, the poor parameter recovery observed at $N = 30$ could reflect both insufficient sample size and capability-distribution shape. The large item sets common in AI benchmarks can improve capability estimation because each model capability is informed by many item responses. However, they do not increase the number of observations available for each item parameter, which remains informed by only N model responses. Thus, a large item set cannot necessarily compensate for a small model sample in item-parameter recovery.

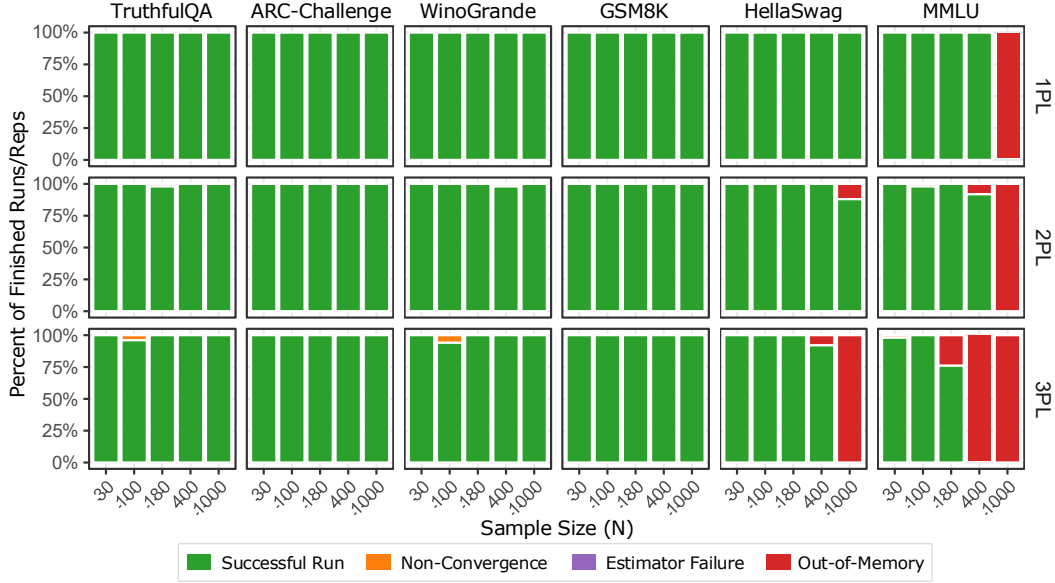


Figure 14: Distribution of MCMC exit statuses across IRT models, benchmarks, and sample sizes.

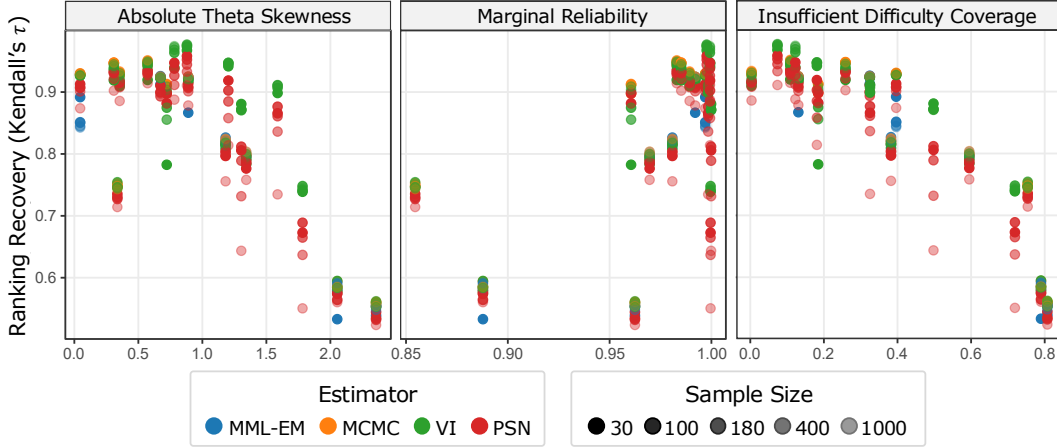


Figure 15: Bivariate relationships between ranking recovery and absolute theta skewness, marginal reliability, and insufficient difficulty coverage, defined in App. D. Each point represents a simulation condition; color indicates the estimator, and opacity indicates sample size.

We therefore conducted two complementary analyses to assess whether the recovery improvement from $N = 30$ to $N = 100$ persisted under more regular capability distributions.

Post-hoc analysis of near-normal replications. At $N = 30$, we retained replications whose distance did not exceed the 95th percentile of the distances obtained from standard-normal samples of the same size, as measured by the 2-Wasserstein distance. This criterion identifies empirical samples whose departures from normality are comparable to those arising from finite random samples drawn from a normal distribution.

Among the $6 \times 3 \times 50 = 900$ existing $N = 30$ replications, 119 met this criterion. Of these, 112 were from the 3PL simulations: 36 for WINOGRANDE, 27 for ARC-CHALLENGE, 17 each for TRUTHFULQA and HELLASWAG, 12 for MMLU, and 3 for GSM8K; we focused on these 112 replications. For each benchmark $\times$ estimator condition, we recomputed the mean and standard deviation of item parameter recovery, item mean error, and aggregate score error.

Table 3: Multilevel regression predicting Fisher-transformed ranking recovery. Continuous predictors were standardized, while estimator was effect-coded such that each estimator coefficient represents its deviation from the mean across all four estimators. Bold and italics indicate the largest and second-largest absolute coefficients among the continuous predictors, respectively.

Predictor	Coefficient	95% Jackknife CI
Intercept	1.380	[1.297, 1.462]
Absolute Theta Skewness	<i>-0.103</i>	[-0.181, -0.026]
Marginal Reliability	0.064	[-0.020, 0.149]
Insufficient Difficulty Coverage	-0.280	[-0.407, -0.152]
Log Sample Size	-0.026	[-0.041, -0.011]
Estimator (MML-EM)	-0.019	[-0.110, 0.071]
Estimator (MCMC)	0.082	[0.040, 0.124]
Estimator (PSN)	-0.111	[-0.166, -0.057]
Estimator (VI)	0.049	[0.006, 0.091]

Table 4: Marginal R^2 values for models predicting Fisher-transformed ranking recovery. The base model included estimator and log sample size, and each extended model added one standardized predictor in Table 3. Marginal R^2 represents the variance explained by the fixed effects. Bold indicates the model with the highest marginal R^2 .

Model	Marginal R^2
Base (<i>Estimator + Log Sample Size</i>)	.039
Base + <i>Absolute Theta Skewness</i>	.437
Base + <i>Marginal Reliability</i>	.325
Base + <i>Insufficient Difficulty Coverage</i>	.760

Fig. 19 compares this near-normal $N = 30$ subset with the complete results. Across metrics, the near-normal subset generally tracked the full $N = 30$ sample and remained substantially worse than the $N = 100$ results. Fig. 20 provides a matched comparison between the selected $N = 30$ replications and their corresponding $N = 100$ replications. Recovery still improved substantially at $N = 100$, indicating that restricting attention to near-normal capability samples did not close the small-sample recovery gap. The GSM8K results should be interpreted cautiously because only three replications met the near-normality criterion.

This post-hoc analysis does not fully isolate capability-distribution shape because the replications were selected rather than experimentally generated. Nevertheless, it shows that the observed $N = 30$ recovery problems were not confined to the most irregular empirical capability samples.

Controlled capability-distribution simulations. We next conducted controlled simulations for ARC-CHALLENGE under 3PL, comparing MCMC, VI, and PSN at $N \in 30, 100$. We considered three capability-sampling schemes:

- *Full resample*: capabilities were resampled from the complete empirical capability distribution used in the primary simulations;
- *Homogeneous resample*: capabilities were resampled from the approximately unimodal $[-1, 1]$ region of the empirical distribution, representing a more homogeneous model population; and
- *Normal sample*: capabilities were drawn from the standard normal distribution, preserving a wider capability range than the homogeneous resample.

The full-resample results were based on the original 50 replications, while the homogeneous and normal simulations used 10 replications per sample size. Fig. 21 reports difficulty, discrimination, and guessing-parameter recovery, mean item error, and aggregate score error under the three sampling schemes.

Across both the homogeneous and normal capability distributions, the recovery gap between $N = 30$ and $N = 100$ persisted for most estimators and metrics. Item-parameter recovery generally improved

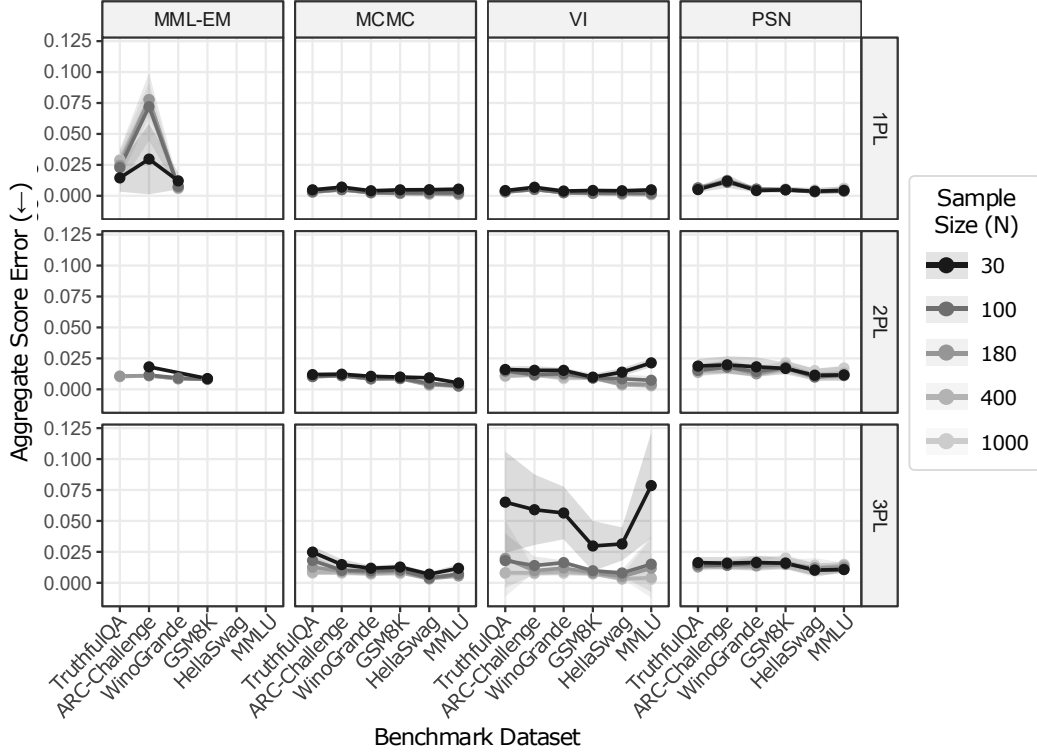


Figure 16: Aggregate score error across IRT models, estimators, benchmarks, and sample sizes. Lower is better. Shaded regions indicate ± 1 standard deviation across 50 replications. Missing points indicate computational infeasibility.

and mean item error declined at $N = 100$, consistent with the results under the full empirical distribution. VI aggregate score error was the only clear exception to this pattern. Thus, replacing the irregular empirical capability distribution with a standard normal distribution or a more homogeneous empirical subset did not eliminate the small-sample recovery problem.

Both analyses show substantial recovery improvements from $N = 30$ to $N = 100$ under more regular capability distributions, indicating that the poor item-parameter recovery at $N = 30$ cannot be explained solely by skewness or other irregularities in the empirical capability distributions. Distributional shape remains relevant, particularly for ranking recovery, but small sample size contributes separately to unreliable item-related inference. The results therefore support our recommendation that increasing the number of evaluated models is important for item-parameter and aggregate-score recovery even when the underlying capability distribution is approximately normal or comparatively homogeneous.

F Detailed results for efficient benchmarking

Please refer to Fig. 22 for the complete short-form ranking-recovery results across the three test lengths (i.e., 10%, 20%, and 100 items), IRT models, benchmarks, and estimators. Figs. 23 & 24 further break down the results by sample size. Additional results across a broader range of short-form benchmark sizes are provided in Figs. 25-30.

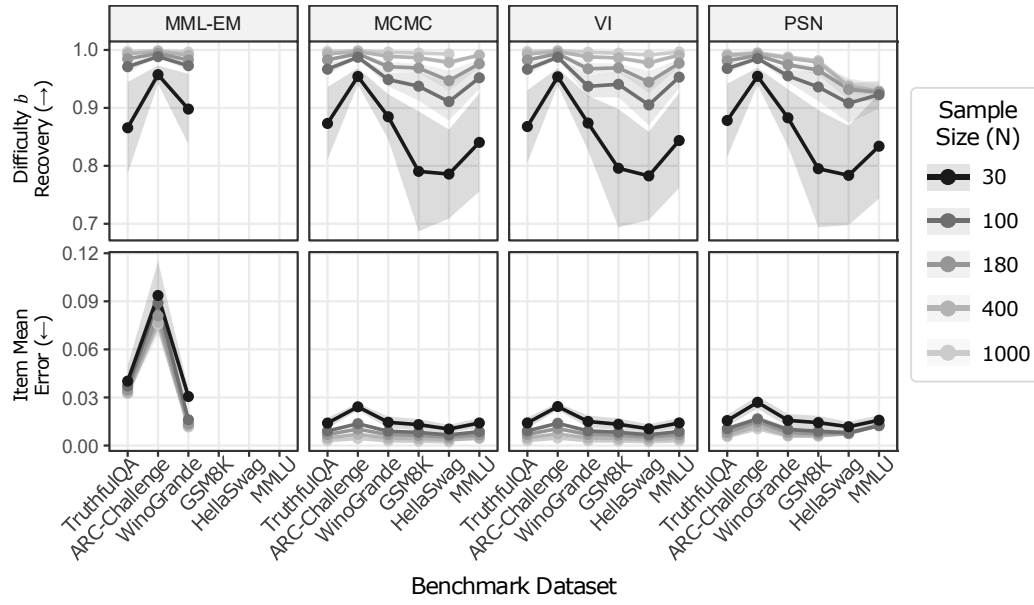


Figure 17: Item parameter recovery under 1PL across estimators, benchmarks, and sample sizes. Higher is better for difficulty recovery, and lower is better for item mean error. Shaded regions indicate ± 1 standard deviation across 50 replications. Missing points indicate computational infeasibility.

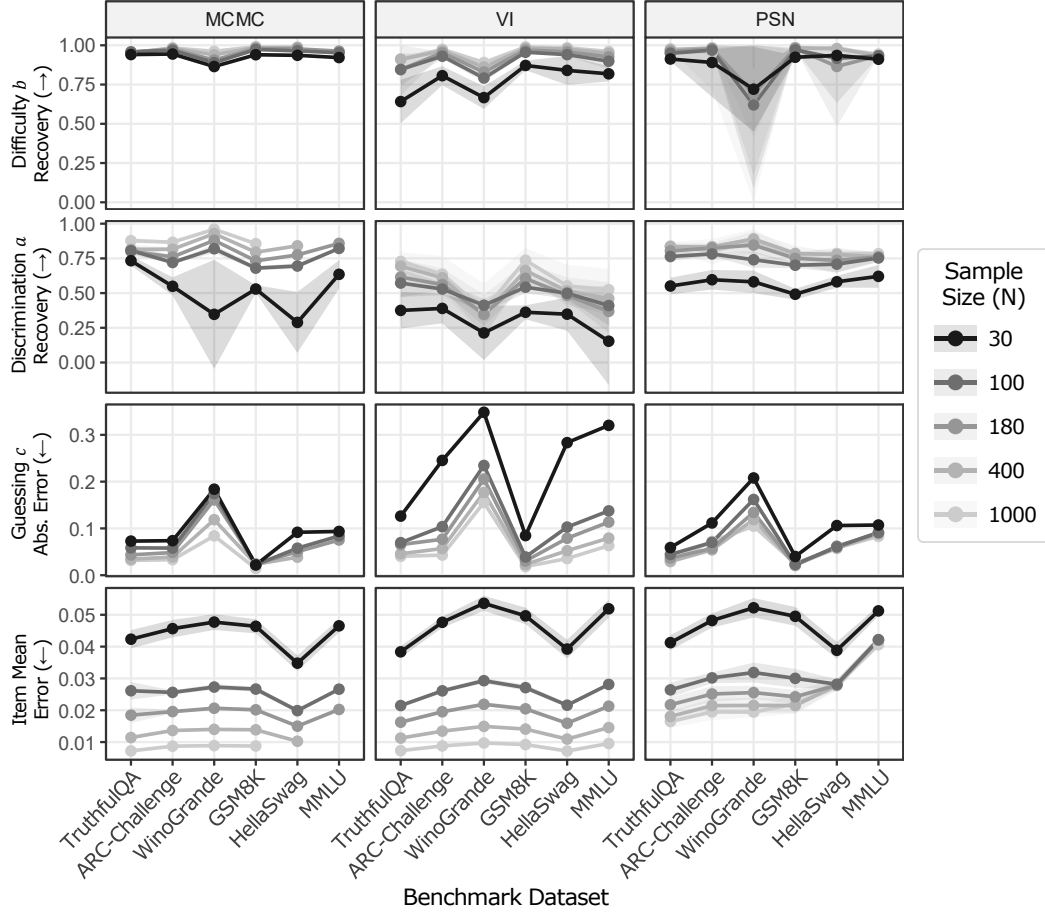


Figure 18: Item parameter recovery under 3PL across estimators, benchmarks, and sample sizes. Higher is better for difficulty recovery and discrimination recovery; lower is better for guessing absolute error and item mean error. Shaded regions indicate ± 1 standard deviation across 50 replications. Missing subplots and points indicate computational infeasibility.

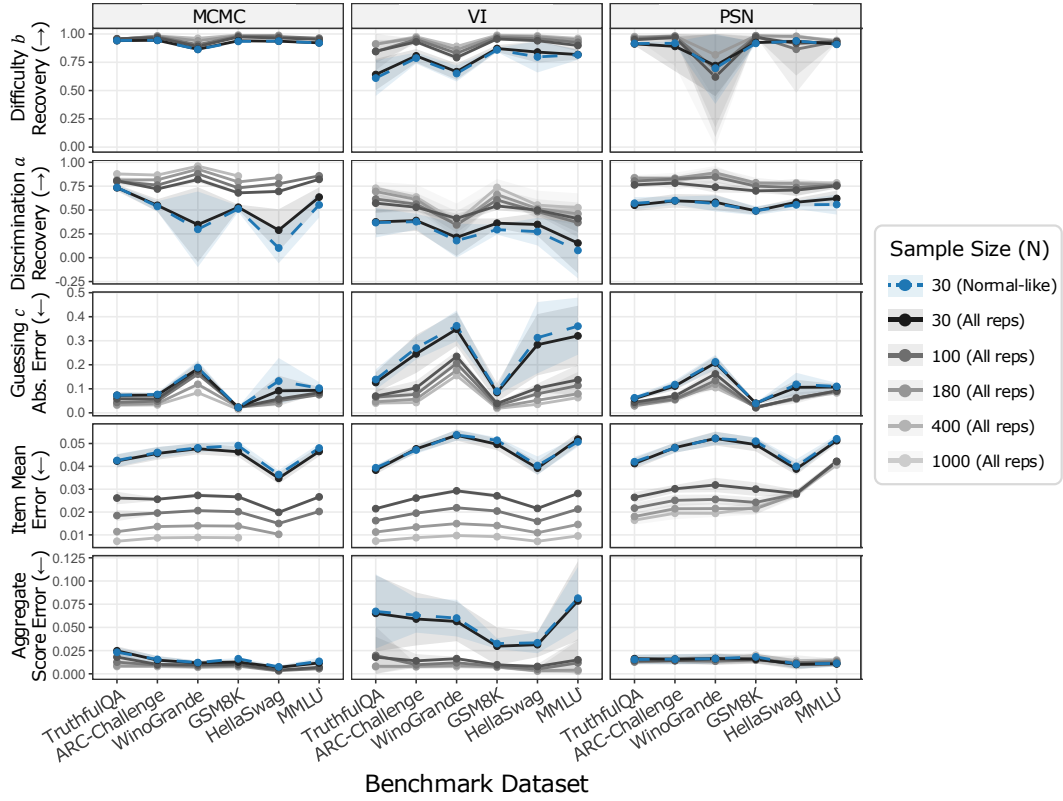


Figure 19: Parameter and aggregate-score recovery under 3PL for the near-normal $N = 30$ subset (blue), compared with all replications at each sample size (grayscale). Points and shaded bands show the mean and one standard deviation across replications, respectively.

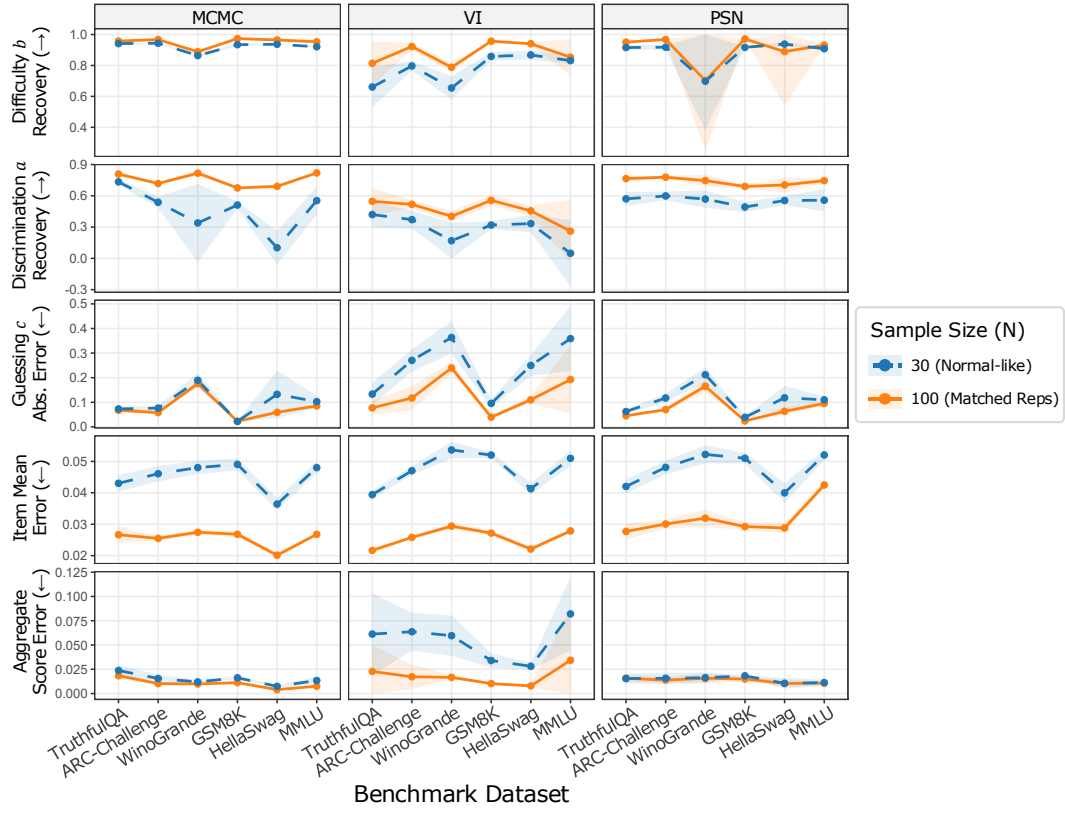


Figure 20: Parameter and aggregate-score recovery under 3PL for the near-normal $N = 30$ subset (blue), compared with the matched $N = 100$ results from the same selected replications (orange). Points and shaded bands show the mean and one standard deviation across replications, respectively.

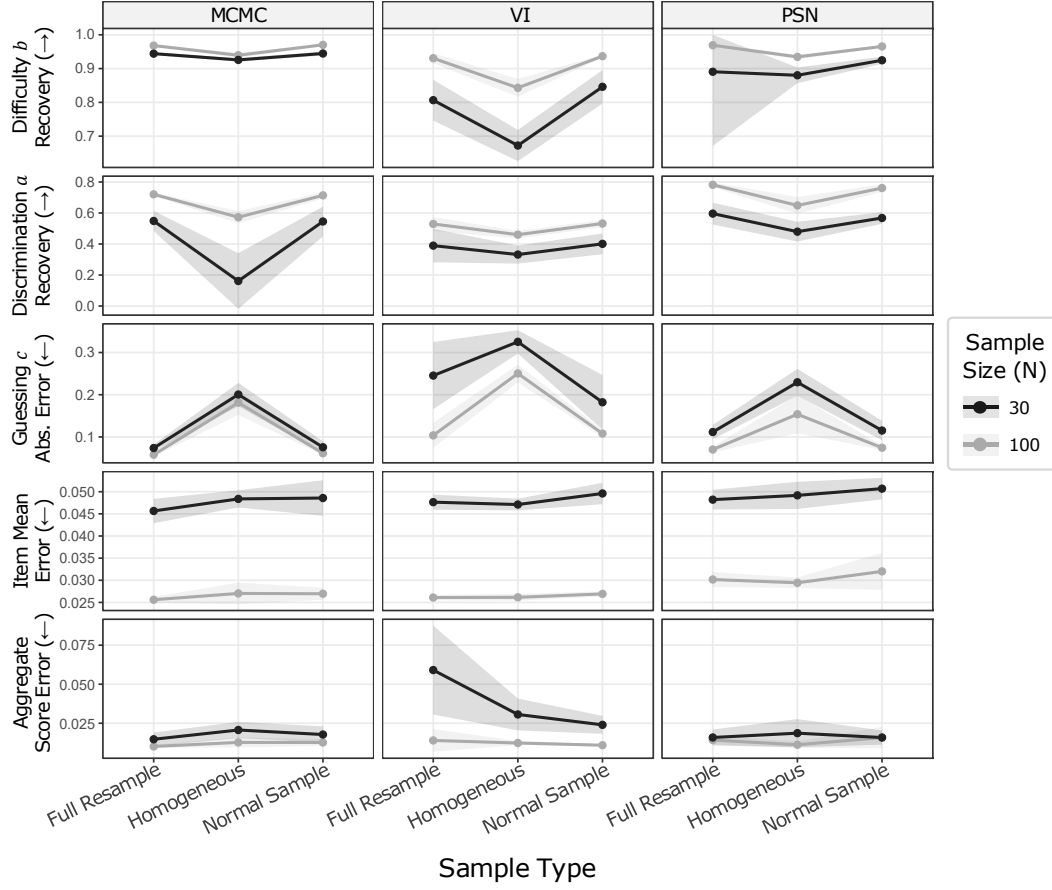


Figure 21: Parameter and aggregate-score recovery for ARC-Challenge under 3PL at $N = 30$ and $N = 100$ across three capability-sampling types: *full resample* draws from the complete empirical capability distribution, *homogeneous* resample from its $[-1, 1]$ region, and *normal sample* from the standard normal distribution. Points and shaded bands show the mean and one standard deviation across replications, respectively.

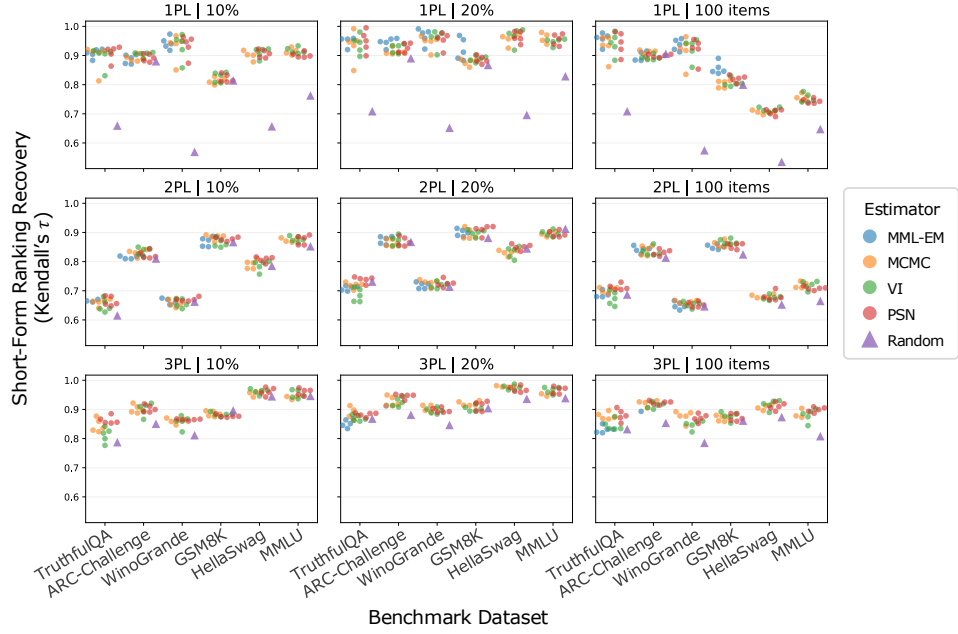


Figure 22: Short-form ranking recovery across test lengths, IRT models, benchmarks, and estimators. Rows correspond to IRT models and columns to subset sizes; sample sizes are pooled without visual distinction. Purple triangles show the random baseline averaged over 30 randomly selected subsets of the corresponding size per replication.

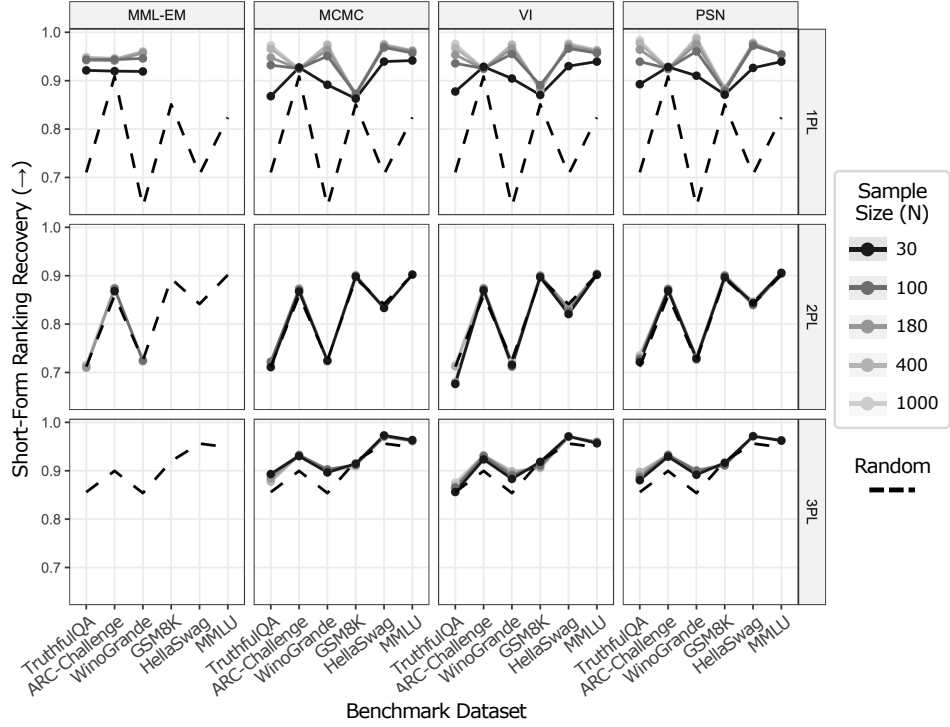


Figure 23: Short-form ranking recovery for 20%-item subsets across IRT models, benchmarks, estimators, and sample sizes.

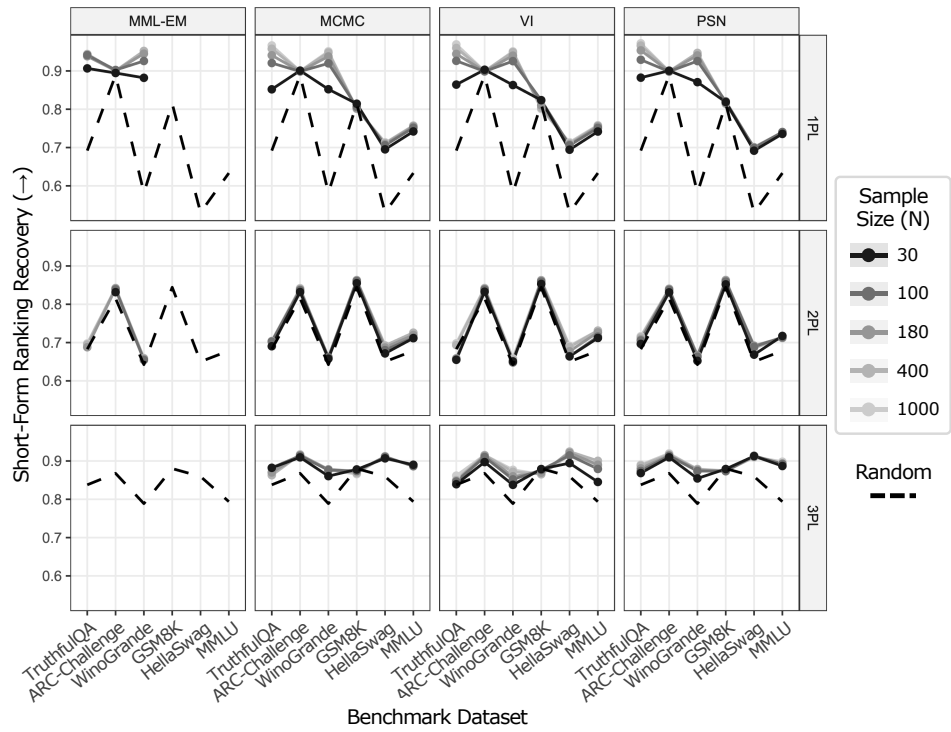


Figure 24: Short-form ranking recovery for 100-item subsets across IRT models, benchmarks, estimators, and sample sizes.

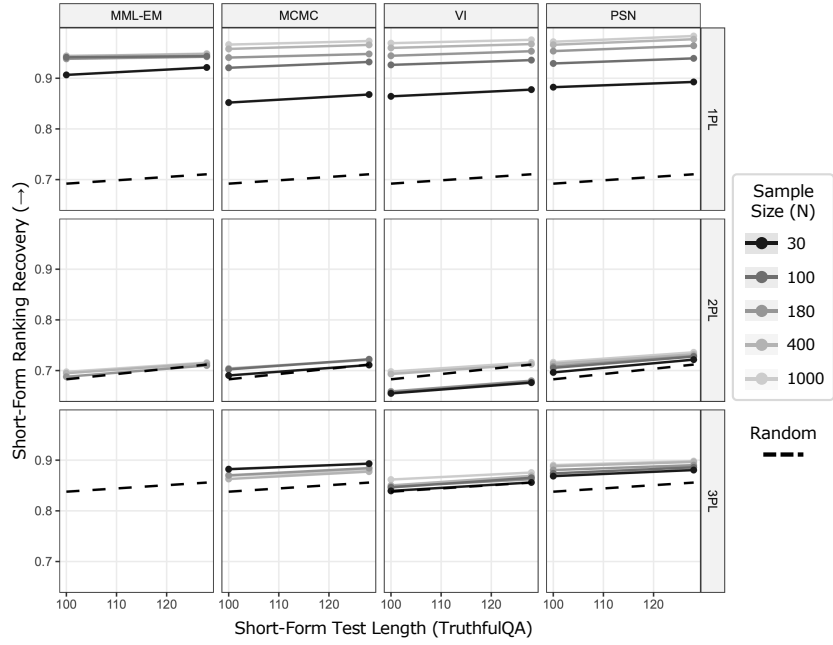


Figure 25: Short-form ranking recovery for larger subsets of TRUTHFULQA across IRT models, estimators, and sample sizes.

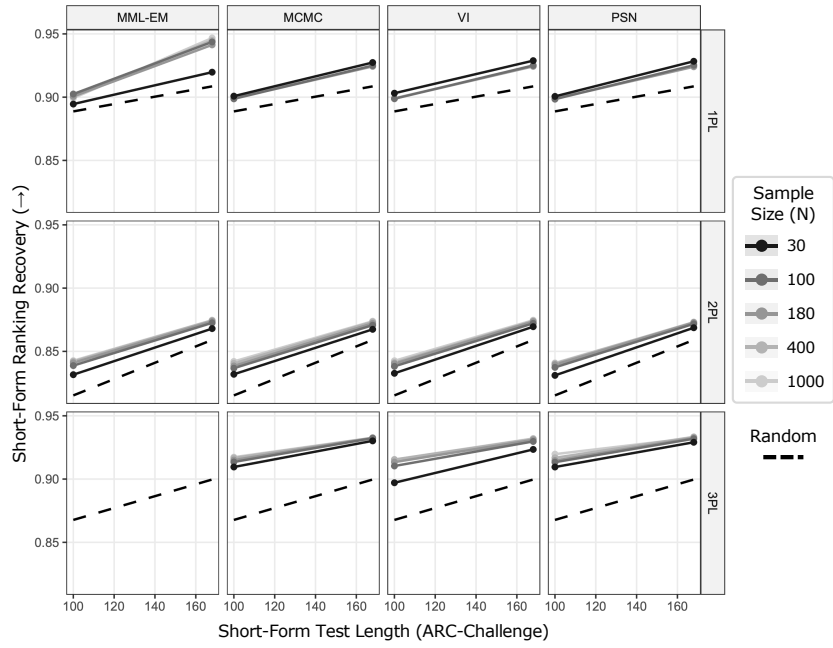


Figure 26: Short-form ranking recovery for larger subsets of ARC-CHALLENGE across IRT models, estimators, and sample sizes.

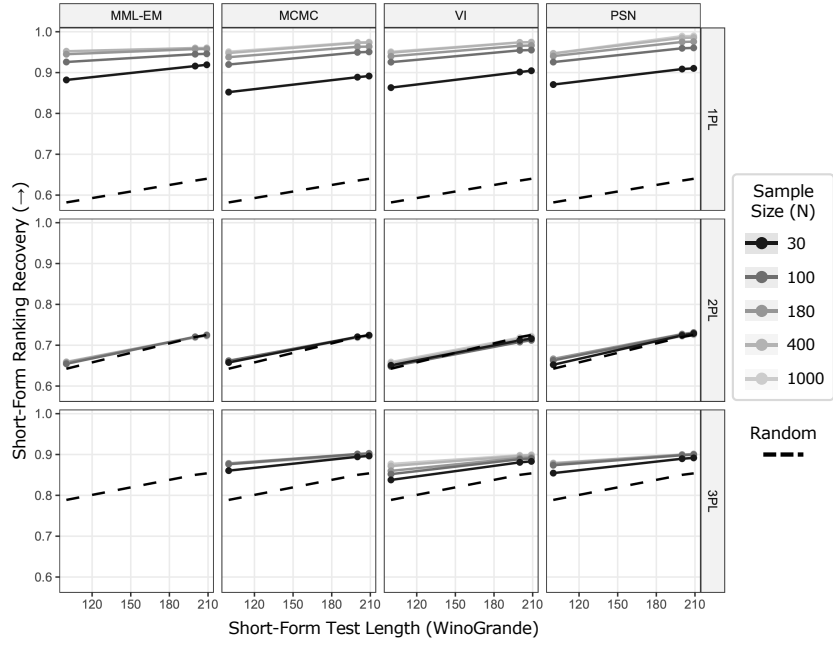


Figure 27: Short-form ranking recovery for larger subsets of WINOGRANDE across IRT models, estimators, and sample sizes.

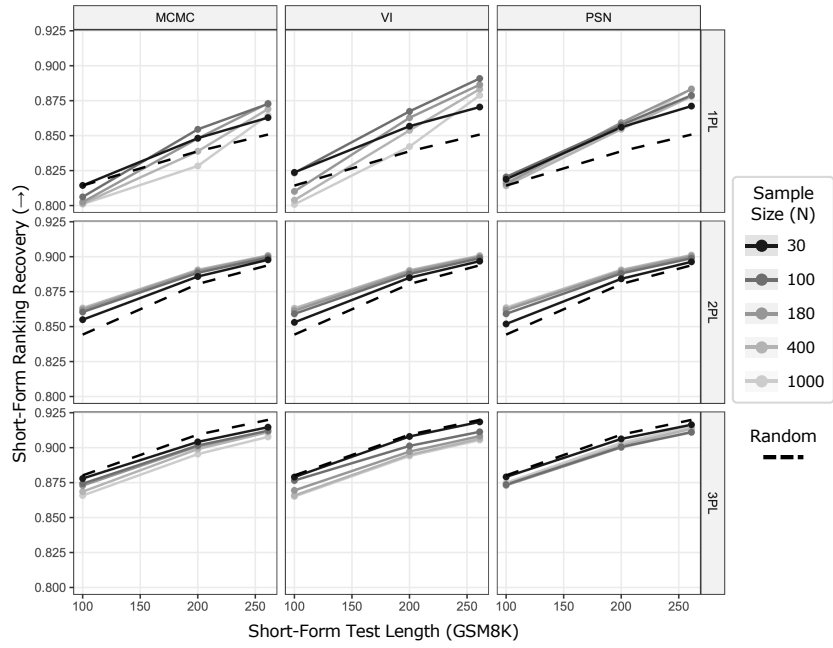


Figure 28: Short-form ranking recovery for larger subsets of GSM8K across IRT models, estimators, and sample sizes.

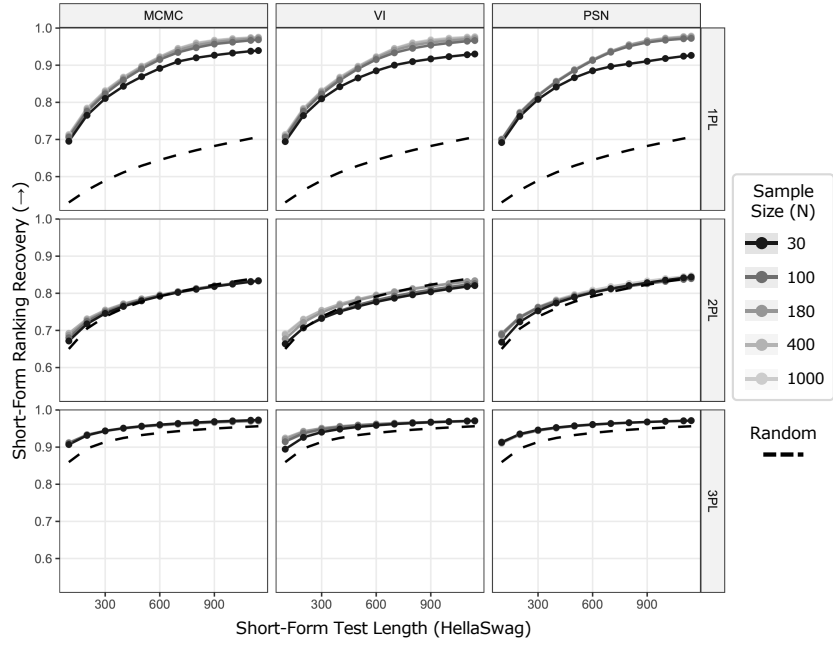


Figure 29: Short-form ranking recovery for larger subsets of HELLASWAG across IRT models, estimators, and sample sizes.

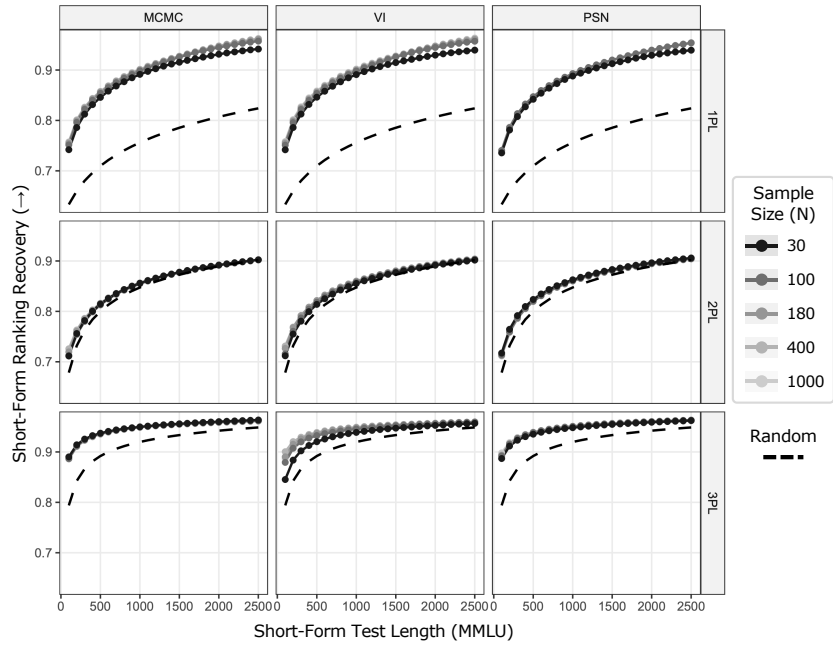


Figure 30: Short-form ranking recovery for larger subsets of MMLU across IRT models, estimators, and sample sizes.