Is the Golden Ticket Tainted? Testing Standardized Tests with AI

Vikram K Suresh

Email: krishnvv@ucmail.uc.edu

Department of Computer Science, University of Cincinnati, Cincinnati-OH, USA.

Department of Economics, University of Cincinnati, Cincinnati-OH, USA.

February 3, 2026

Abstract

Standardized test scores are signals in college admissions, yet the measurement properties of these instruments receive little empirical attention. This paper uses large language models (LLMs) as stable external benchmarks to audit item-level changes in the SAT math section from 2011 to 2023. Estimating a pooled item response theory model on 1291 math items evaluated by 20 LLMs, I find a downward drift in item difficulty beginning in 2017, with discrimination remaining broadly stable. This shift moves test information towards lower ability regions, reducing signal precision at the upper tail where selective institutions screen applicants. These diagnostics suggest that where the SAT is most informative has changed over time, with implications for efficiency of student institution matching and the distributional consequences of admissions policies.

Keywords: standardized testing; artificial intelligence; higher education; signal quality; item response theory

JEL Codes: I21, I23, I24, J24

Acknowledgments: VKS acknowledges the support of the Department of Economics at the University of Cincinnati. Especially, Saani Rawat for his extraordinary help with the data curation and feedback on earlier drafts of this manuscript. VKS is also immensely grateful to Prof. Asawari Deshmukh and her freshmen students who volunteered to take the SAT assessment for validation purposes. VKS also thanks the Department of Computer Science at the University of Cincinnati for their support during this project.

1 Introduction

Standardized admissions tests such as the SAT have long served as quantitative signals of academic preparedness in U.S. college admissions. Literature identifies test scores as noisy measures of latent skill and studies how institutions weight, down weight, or ignore these signals in admissions and scholarship decisions (Cunha and Heckman, 2007; Cunha *et al.*, 2010). At the same time, test scores are found to predict long-run educational and labor-market outcomes, making them central inputs into selective screening (Rothstein, 2004; Chetty *et al.*, 2023; Friedman *et al.*, 2025).

Since COVID-19, many institutions expanded test-optional or test-free admissions policies. Recent work documents that these changes altered application behavior, enrollment patterns, and the composition of admitted students, particularly at selective institutions (Sacerdote *et al.*, 2025; Avery *et al.*, 2025). However, this literature largely treats the SAT as a stable measurement instrument. Institutions either observe scores or do not, but little empirical work evaluates the measurement properties of the test itself. Specifically, where and how precisely it measures ability have changed over time. If the test becomes less informative in the ability ranges relevant for selective screening, then any admissions process that uses test scores faces different precision tradeoffs even holding policy fixed.

Existing work focuses primarily on whether scores are required or submitted and how participation decisions shape admissions outcomes. *Conditional on the test existing and being observed, how informative is the signal?* If the SAT's measurement properties evolve over time, then comparisons of pre- and post–test-optional outcomes conflate policy effects with changes in signal quality. This issue is particularly important in the context of generative AI, which has made other components of the application package including essays, recommendation letters, and portfolios, easier to polish and potentially game (Spence, 1973; Frankel and Kartik, 2019; Cui *et al.*, 2025). Standardized tests may play a more central role in the future admissions. It is therefore critical to understand the informativeness of the SAT as a measurement instrument.

Evolving test properties alter the signal in education markets, influencing intergenerational mobility and labor-market outcomes (Rothstein, 2004; Chetty *et al.*, 2020, 2023). This paper contributes to AI-based measurement in applied economics by using large language models (LLMs) as a common external benchmark to recover item-level parameter changes in a proprietary high-stakes standardized test. Other assessments document concurrent math performance declines (e.g., NAEP Grade 12 math proficiency reached historic lows in 2024; PISA U.S. math declined by 13 points from 2018-2022) (National Center for Education Statistics, 2024; Organisation for Economic Co-operation and Development, 2023). The American College Testing (ACT) which also tests for college readiness like the SAT has reported similar declining trends in composite scores (ACT, Inc., 2025). These changes motivate a closer analysis of the measurement properties

over time of one such standardized test, the SAT Math section.

The psychometric literature on item parameter drift (IPD) and differential item functioning (DIF) provides a conceptual framework for understanding how item properties evolve over time and across subpopulations (Wells *et al.*, 2002; Holland and Wainer, 1993). This however requires access to item-level human response data, which are proprietary for high-stakes tests such as the SAT. The LLM-benchmark approach acts as an external reference variant of the IPD analysis where a stable synthetic population substitutes for the matched human reference group in traditional drift detection procedures.

This AI-based benchmarking approach fixes a set of open-source LLMs as synthetic test-takers to diagnose changes in SAT Math items from 2011 to 2023. Estimating a pooled item response theory (IRT) model on LLM responses yields item-level difficulty, discrimination, and guessing parameters on a common latent scale (Birnbaum, 1968; Lord, 1980). Using these recovered IRT parameters, I examine how the informativeness of SAT Math scores evolved over time and what cohort ability distributions are required to reconcile observed population SAT score moments. In settings where item-level human response data are proprietary, LLMs offer an external measurement audit of the test instrument itself.

The identifying assumption is that the LLM benchmark is stable over time. All items are evaluated simultaneously by the same LLMs under a fixed prompting and decoding protocol (see Online Appendix I for prompt details). So, cross-year variation in response patterns reflects item-level differences rather than benchmark drift. To assess whether LLM-implied difficulty trends generalize to human test-takers, I conduct a supplementary validation exercise in which university business college freshmen answer a stratified sample of items drawn from items in years 2011, 2015, 2019, and 2023. This exercise is not powered to estimate item parameters. Instead, it provides a directional check on whether the rank ordering of difficulty implied by the LLM benchmark aligns with human performance.

The results show a downward trend in LLM-implied item difficulty beginning in 2016, with comparatively modest changes in discrimination. Mechanically, this shift moves test information toward lower regions of the ability scale (Lord, 1980). As a consequence, the SAT Math section becomes less precise in the upper tail, reducing its capacity to discriminate among high-ability examinees, the region most relevant for selective admissions. These findings are measurement diagnostics rather than causal estimates of policy effects. However, they highlight how the informativeness of standardized tests has changed, AI tools providing a novel lens for such measurement audits in the absence of proprietary human response data.

2 Conceptual Framework

Observed SAT score distributions in a given year result from two components: (i) the cohort latent ability distribution of test-takers, and (ii) the measurement technology in the given year that maps ability into scores. Changes in observed scores year-on-year can reflect changes in student specific attributes such as preparedness and composition, or test specific changes in measurement properties or some combination of both. In this paper, I focus on changes in test measurement technology using AI as a common benchmarking framework. Separating changes in measurement technology requires a benchmark external to the test-taking population.

Let θ_i denote the latent math ability of test-taker i, and let $F_t(\theta)$ denote the test-taker ability distribution in year t. The observed SAT Math score $S_{it} \in [200, 800]$ is produced by a year-specific test technology $\mathcal{T}_t$ applied to latent ability:

$$S_{it} = \mathcal{T}_t(\theta_i) + \varepsilon_{it}, \mathbb{E}[\varepsilon_{it}|\theta_i] = 0. \tag{1}$$

where ε_{it} captures idiosyncratic measurement noise arising from factors such as test-day shocks, administration conditions, and other unobserved disturbances. The test technology $\mathcal{T}_t$ aggregates item-level response function, each characterized by difficulty, discrimination, and guessing parameters into a scaled score via year-specific scoring rules (Lord, 1980; Kolen and Brennan, 2014).

The empirical challenge is that item-level human response data for the SAT are proprietary. I address this limitation by using LLMs as the external benchmark. Holding the model set and prompting fixed, I estimate item parameters from LLM response patterns. The identifying assumption is that LLM capability on SAT Math items is stable over the study period (all responses from the 20 LLMs were collected via Together AI's API within 2 days in October 2025), so that cross-year variation in estimated item parameters reflects changes in the items rather than benchmark drift. Using a supplementary human validation exercise, I test this assumption and find students exhibit a difficulty ordering concordant with the LLM-implied ranking.

Formally, let p_{mj} denote the probability that model m answers item j correctly. Under a three-parameter logistic (3PL) IRT model, this probability depends on model ability α_m and item parameters (a_j, b_j, c_j):

$$p_{mj} = c_j + (1 - c_j) \cdot \frac{1}{1 + \exp[-a_j(\alpha_m - b_j)]}. \tag{2}$$

Because all items are evaluated by the same models simultaneously, item parameters are identified on a common latent scale without the need for cross-year linking (Kolen and Brennan, 2014). Year-over-year trends in estimated difficulty parameters b_j therefore capture changes in item characteristics relative to a fixed benchmark, rather than changes in the benchmark itself.

3 Methods

3.1 Item corpus

I compile SAT Math items from official paper based examinations between 2011–2023 from the internet archives. Each item is collected along with administration year, form identifiers, and associated question type which could be multiple-choice or numerical answer. The 2016 redesign reduced multiple-choice options from five to four. I retain the option-count metadata to regularize the pseudo-guessing component of the IRT model. Items with figures that could not be reliably reconstructed due to poor quality from archived PDFs are excluded to avoid introducing unobserved variation in item presentation across years. Items were excluded if the solution required extracting quantitative information from a figure that could not be reliably represented in text. The final corpus contains 1,291 unique items, representing approximately 65% of the expected item pool across all administered forms during 2011–2023, with annual coverage rates ranging from 58% to 73%.

3.2 LLM response collection

I evaluate $M = 20$ open-source large language models (LLMs) as synthetic examinees using a standardized prompting and extraction protocol. For multiple-choice items, prompts include the question and response options. Further, to assess response stability, I randomized option order, reprompt the question multiple times (five) to assess order sensitivity and within-model variability. For numerical response items, prompts request numeric answers and extract the output via regex constraints. All models are run with deterministic decoding, that is, they are prompted with temperature set to 0.

As all items from 2011–2023 are evaluated by the same models under a fixed prompting protocol, cross-year variation in response patterns reflects item-level differences under a constant benchmark rather than changes in model capability. This design parallels recent work for diagnosing consequences of changing signal quality in other screening contexts (Frankel and Kartik, 2019; Cui *et al.*, 2025). A contamination robustness check using paraphrased items confirms that LLM responses reflect problem-solving rather than memorization of training data. For a full SAT form from 2011, 2015, 2019 and 2023, the overall accuracy on original items (54.3%) and paraphrased items (54.0%) is nearly identical, with year-level differences ranging from -1.8 to $+1.9$ percentage points. This pattern is inconsistent with memorization, which would predict systematically higher accuracy on original items (Online Appendix G).

3.3 Pooled Bayesian IRT model

I estimate item parameters using a pooled Bayesian item response model. For multiple-choice items, I employ a three-parameter logistic (3PL) specification with item discrimination a_{jt}, difficulty b_{jt}, and pseudo-guessing c_{jt}. For numerical answer response items, I impose $c_{jt} \equiv 0$ (Fox, 2010). Pooling all items across years in a single likelihood places item parameters on a common latent scale without cross-year linking, consistent with IRT practice when a stable external benchmark is available (Birnbaum, 1968; Lord, 1980; Kolen and Brennan, 2014).

Let $y_{mjtr} \in \{0, 1\}$ denote the scored response of an LLM model m on item j from year t on attempt r. Although decoding is deterministic, response variation arises from option-order shuffling across attempts.Models have been found to exhibit sensitivity to the position of correct answers, a phenomenon in LLM evaluation (Zheng *et al.*, 2023). The parameter α_m therefore captures each model's systematic tendency to answer correctly—analogous to latent ability in human IRT, though it is not a cognitive construct. To capture within-model variability from option-order effects, I allow attempt-specific effective ability:

$$\alpha_{mr} = \alpha_m + \sigma_{\text{rep}} \varepsilon_{mr}, \qquad \varepsilon_{mr} \sim \mathcal{N}(0, 1). \tag{3}$$

Conditional on α_{mr}, the probability of a correct response is given by,

$$P_{jt}(\alpha_{mr}) = c_{jt} + (1 - c_{jt})\Lambda(a_{jt}(\alpha_{mr} - b_{jt})), \qquad \Lambda(x) = \frac{1}{1 + e^{-x}}. \tag{4}$$

The likelihood therefore can be written as,

$$\mathcal{L}(\boldsymbol{\Theta} \mid \mathbf{Y}) = \prod_{m} \prod_{j,t} \prod_{r} P_{jt}(\alpha_{mr})^{\, y_{mjtr}} \left[1 - P_{jt}(\alpha_{mr})\right]^{1 - y_{mjtr}}, \tag{5}$$

where $\boldsymbol{\Theta}$ collects all item and model parameters.

Posterior inference is implemented using Hamiltonian Monte Carlo using Turing.jl and posterior means are provided as point estimates in Table 1 (Neal, 2011; Betancourt, 2017; Ge *et al.*, 2018). Full prior specifications, convergence diagnostics, and robustness checks are provided in the Online Appendix B. Appendix B also reports item characteristic curve validation, confirming that the monotonic ability-accuracy relationship assumed by the 3PL holds in the LLM response data. This supports the use of IRT as a framework for directional assessment of item properties, which is the goal of this analysis.

3.4 Cohort recovery and test information

To connect the estimated measurement technology to observed score distributions, I recover year-specific cohort ability distributions that reconcile LLM-implied item parameters with published College Board population level score moments. I parameterize latent math ability in year t as

$$\theta \sim \mathcal{N}(\mu_t, \sigma_t^2), \tag{6}$$

and recover (μ_t, σ_t) by simulated moment matching. Importantly, these are model-consistent latent parameters on the LLM-defined IRT scale, not structural estimates of human ability. Their role is to reconcile item parameters with observed score moments.

For a candidate (μ, σ), I simulate a large synthetic cohort, draw abilities from $\mathcal{N}(\mu, \sigma^2)$, generate item responses using the estimated item parameters, apply year-specific scoring rules from official raw-to-scaled score conversions, and compute simulated score moments (College Board Research, n.d.). Let M_t^{obs} denote the vector of published score moments for year t and $\widehat{M}_t(\mu, \sigma)$ the simulated moments. The objective uses squared relative errors for scale comparability across moments:

$$M_t^{\text{obs}} = \Big(\bar{s}_t, \text{sd}(s_t), Q_{0.25}(s_t), Q_{0.50}(s_t), Q_{0.75}(s_t)\Big), \tag{7}$$

$$\text{err}(m) = \left(\frac{m^{\text{sim}} - m^{\text{obs}}}{m^{\text{obs}}}\right)^2, \tag{8}$$

and minimizes

$$(\hat{\mu}_t, \hat{\sigma}_t) = \arg\min_{\mu, \sigma > 0} L_t(\mu, \sigma), \tag{9}$$

with

$$L_t(\mu, \sigma) = w_{\text{mean}}\Big(\text{err}(\bar{s}) + \text{err}(Q_{0.50})\Big) + w_{\text{others}}\Big(\text{err}(\text{sd}) + \text{err}(Q_{0.25}) + \text{err}(Q_{0.75})\Big). \tag{10}$$

The chosen weighting scheme mirrors the priorities in the SAT scaling and equating process. The College Board defines score meaning and interchangeability through the central region (mean and median) of the score distribution, where scaled scores are stable across forms and administrations. Score precision in the tails declines due to discrete raw-to-scaled mappings and concordance discontinuities. The objective function therefore follows this logic by upweighting mean and median errors relative to other moments (College Board, 2017).

$$w_k = \begin{cases} 4 & k \in \{\text{mean}, \text{median}\}, \\ 1 & k \in \{\text{SD}, Q_{25}, Q_{75}\}. \end{cases} \tag{11}$$

Since the raw-to-scaled score mapping (all individual level SAT scores are multiples of 10) is discrete and incorporates redesign driven concordance discontinuities, the objective function is not smooth. Gradient-based GMM procedures are therefore difficult to implement reliably, and I minimize the objective using the Nelder–Mead simplex algorithm (Nelder and Mead, 1965). Online Appendix D reports sensitivity to an alternative weighting scheme.

Test information. Given year-specific LLM-implied measurement technologies $\widehat{\mathcal{T}}_t(\cdot)$, I characterize score informativeness using the test information function. For an item j in year t with parameters $\phi_{jt} = (a_{jt}, b_{jt}, c_{jt})$, the probability of a correct response at ability θ under the three-parameter logistic (3PL) model is

$$P_{jt}(\theta) = c_{jt} + (1 - c_{jt})\Lambda(a_{jt}(\theta - b_{jt}))\,, \qquad \Lambda(x) = \frac{1}{1 + e^{-x}}. \tag{12}$$

Item-level Fisher information is as defined in Hambleton and Swaminathan (1985):

$$I_{jt}(\theta \mid \phi_{jt}) = \frac{\left[P'_{jt}(\theta)\right]^2}{P_{jt}(\theta)\left[1 - P_{jt}(\theta)\right]}, \tag{13}$$

and for a realized complete test form $S_t \subseteq \mathcal{J}_t$ where $\mathcal{J}_t$ is the item pool for year t, total information is

$$I_t(\theta \mid S_t) = \sum_{j \in S_t} I_{jt}(\theta \mid \phi_{jt}). \tag{14}$$

Because the specific SAT Math form administered in year t is drawn from the available item pool, I report expected test information under stratified uniform sampling without replacement. Let $\mathcal{J}_t^{\text{mcq}}$ and $\mathcal{J}_t^{\text{ans}}$ denote the multiple-choice and numerical answer item sets in a given year, with $N_{\text{mcq},t}$ and $N_{\text{ans},t}$ available items and $n_{\text{mcq},t}$ and $n_{\text{ans},t}$ items appearing on a typical form. Before 2016, the SAT included 54 items (42 multiple-choice, 12 numerical) and since then it included 58 items (44 multiple-choice, 14 numerical) in one complete test form. Expected test information is

$$\mathbb{E}[I_t(\theta)] = \frac{n_{\text{mcq},t}}{N_{\text{mcq},t}} \sum_{j \in \mathcal{J}_t^{\text{mcq}}} I_{jt}(\theta) + \frac{n_{\text{ans},t}}{N_{\text{ans},t}} \sum_{j \in \mathcal{J}_t^{\text{ans}}} I_{jt}(\theta). \tag{15}$$

The implied standard error of measurement at ability θ is

$$SE_t(\theta) = \frac{1}{\sqrt{\mathbb{E}[I_t(\theta)]}}. \tag{16}$$

Interpretation of recovered cohort parameters. The recovered $(\hat{\mu}_t, \hat{\sigma}_t)$ should not be interpreted as estimates of true changes in human ability over time. Since the item difficulty drifts downward in the LLM benchmark, reconciling these easier items with stable observed score distributions requires shifting the cohort distribution leftward on the latent scale. The large numerical change in $\hat{\mu}_t$ therefore reflects the joint identification of item and cohort parameters rather than a structural claim about test-taker preparedness. Test information and $SE_t(\theta)$ are driven by the item parameters themselves. The cohort parameters serve only to locate policy-relevant ability thresholds such as $\theta^t_{0.90}$.

3.5 Precision diagnostics and misclassification

To translate precision into an economically interpretable object, I map $SE_t(\theta)$ into a stylized misclassification probability under a symmetric threshold rule. This metric captures the probability that measurement error causes an examinee near a cutoff to be classified on the wrong side. It is intended as a descriptive diagnostic linking changes in measurement precision to screening reliability rather than as a behavioral model. To illustrate the mechanical implications, for a given admissions threshold θ^*, under some tolerance band $\delta > 0$, the misclassification probability is

$$P_{\text{misclass},t}(\theta^*, \delta) \approx 2 \times \Phi\left(-\frac{\delta}{SE_t(\theta^*)}\right),$$

where $\Phi(\cdot)$ is the standard normal CDF. From the recovered cohort distributions in Section 3.4, I set θ^* to the 90th percentile of each year's cohort distribution, $\theta^t_{0.90}$, and evaluate misclassification probabilities for $\delta = 0.5$ on the latent ability scale, corresponding to approximately 15–30 SAT points depending on year.[1] In addition, I anchor the ability threshold at a fixed high-ability point $\theta^\star = \theta^{2011}_{0.90}$ to isolate intensive-margin changes in measurement precision from changes in cohort composition.

3.6 Human validation

To assess whether AI-implied difficulty trends generalize to human test-takers, I administer a stratified sample of SAT Math items to university business college freshmen students. I draw 12 items using a fixed seed from

[1] The approximate SAT-point equivalent is $\delta \times (\text{sd}^{\text{obs}}_t / \hat{\sigma}_t)$, where sd^{obs}_t is the published SAT Math standard deviation and $\hat{\sigma}_t$ is the recovered cohort standard deviation on the latent scale. With $\delta = 0.5$, $\text{sd}^{\text{obs}}_t \approx 100$–$115$, and $\hat{\sigma}_t \approx 2.0$–$3.3$ (Table 3), this yields approximately 15–30 points where marginal decisions could be affected.

each of four years (2011, 2015, 2019, and 2023), totaling 48 items. Items were presented in a fixed order that was randomized across item years, so that year-blocks were mixed rather than sequential, administered within a 70-minute session (approximately the same length as the SAT math section and time-per-question). Participants were not informed of item year or the purpose of the exercise.

The validation is designed as a directional check rather than a fully powered item-calibration exercise. The primary test is the Jonckheere-Terpstra test for ordinal data, which evaluates whether human accuracy is stochastically increase across year groups under the null of identical distributions (Jonckheere, 1954). This nonparametric test directly assesses whether the ordering implied by the LLM benchmark of declining difficulty from 2011 to 2023 is concordant with human-experienced difficulty. A supplementary linear probability model with student fixed effects provides interpretable effect sizes, these details are included in the Online Appendix H.

4 Results

This section reports three measurement diagnostics implied by the AI-based IRT benchmark: (i) a downward drift in item difficulty, (ii) the corresponding shift in where the test concentrates information along the latent ability scale, and (iii) the resulting change in upper-tail precision, evaluated both within-year and at a fixed high-ability anchor. Under standard IRT, items are most informative near their difficulty locations, so shifts in the distribution of b translate directly into shifts in the test information function and standard errors of measurement (Birnbaum, 1968; Lord, 1980). These changes are AI implied diagnostics of the measurement properties of the SAT Math section rather than causal estimates of policy effects.

4.1 Item difficulty declines while discrimination remains broadly stable

Table 1 summarizes year-level means of discrimination ($\bar{a}_t$), difficulty ($\bar{b}_t$), and pseudo-guessing ($\bar{c}_t$), along with item counts by format. Mean difficulty exhibits a downward shift after 2018: $\bar{b}_t$ is positive through 2017 and becomes consistently negative from 2018 onward, reaching its lowest mean in 2022. In contrast, discrimination remains in a relatively narrow band across years, indicating that items retain local differentiating power even as their locations shift. Pseudo-guessing increases discontinuously in the post-2016 era, consistent with the change from five-option to four-option multiple-choice items and the corresponding change in random-guess baselines. A contamination robustness check using paraphrased items confirms that LLM responses reflect problem-solving rather than memorization of training data. For a full SAT form from 2011, 2015, 2019 and 2023, the overall accuracy on original items (54.3%) and paraphrased items (54.0%) is nearly identical, with year-level differences ranging from -1.8 to $+1.9$ percentage points. This pattern

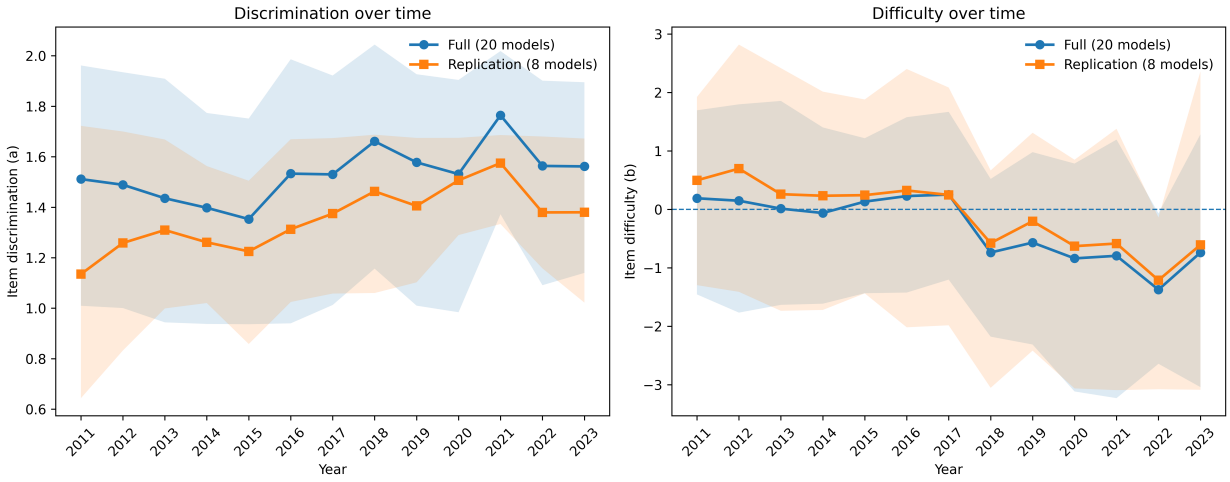

Figure 1: This figure shows the trends in discrimination parameter a and difficulty parameter b over time, comparing the full 20-model AI benchmark to an 8-model subset as a replication check. The trends illustrate the downward shift in difficulty. The shaded portion represents the inter-quartile range. More details regarding the 8-model replication are provided in the Online Appendix.

is inconsistent with memorization, which would predict systematically higher accuracy on original items. (Online Appendix G).

Table 1: Year-level item-parameter summaries from the full AI benchmark (2011–2023).

Year	N items	N MCQ	N numeric	$\bar{a}_t$	$\bar{b}_t$	$\bar{c}_t$
2011	94	75	19	1.511	0.191	0.213
2012	139	114	25	1.489	0.149	0.214
2013	73	60	13	1.436	0.013	0.214
2014	128	104	24	1.398	-0.061	0.207
2015	69	56	13	1.353	0.134	0.210
2016	116	90	26	1.533	0.228	0.261
2017	81	65	16	1.530	0.254	0.259
2018	125	96	29	1.661	-0.737	0.260
2019	126	98	28	1.578	-0.567	0.265
2020	73	61	12	1.531	-0.838	0.255
2021	73	53	20	1.764	-0.793	0.267
2022	120	90	30	1.564	-1.373	0.266
2023	74	56	18	1.562	-0.738	0.261

Notes: $\bar{a}_t$, $\bar{b}_t$, and $\bar{c}_t$ are means of the estimated 3PL item discrimination, difficulty, and pseudo-guessing parameters, respectively, using the pooled AI benchmark. For numeric-response items, c_j is fixed at zero in estimation; $\bar{c}_t$ therefore reflects the MCQ composition in year t. Parameter interpretation follows standard 3PL conventions (Birnbaum, 1968; Lord, 1980).

4.2 Where the SAT is most informative shifts left

The difficulty drift is shown in Table 1 while Figure 2 implies a shift in where the exam concentrates information. Under IRT, information is maximized in neighborhoods of item difficulty, so a leftward shift

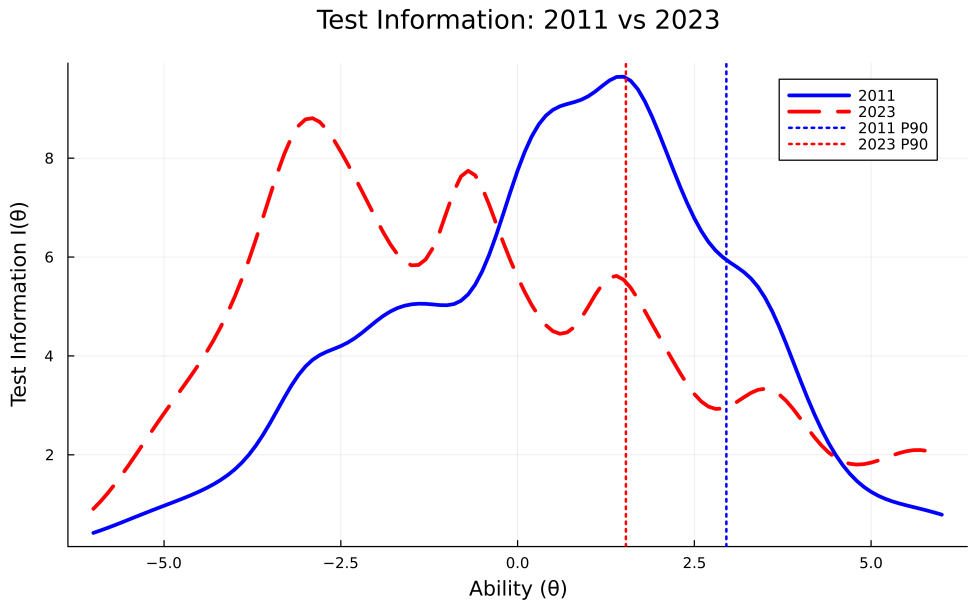

Figure 2: Item-level Fisher information is computed using the standard 3PL information function $I_t(\theta)$ (based on the 3PL probability and its derivative; see Equation 13). To translate item information into a *test*-information curve, I compute the expected test information under stratified uniform sampling from the year-t item pool. Sum $I_t(\theta)$ over all MCQ items and scale by the expected sampling rate $n_{\text{MCQ}}/N_{\text{MCQ}}$, and analogously for numeric-response items, then add the two components as shown in Equation 15. Then evaluate this expected test information on a dense ability grid (e.g., $\theta \in [-6, 6]$ in steps of 0.1) yielding $I_t(\theta)$ for the full range of ability values. **The vertical dashed lines indicate the 90th percentile ability for each year. In 2011, the test information function provides substantial precision at the 90th percentile. By 2023, the leftward shift in item difficulty has reduced information at high ability levels, leaving the 90th percentile in a region where the test is less informative.**

in the b distribution moves the peak and mass of the test information function toward lower ability values (Lord, 1980). Two diagnostics make this implication concrete without relying on figures.

First, within-year upper-tail precision depends on how far the 90th percentile of that year's cohort distribution lies from the region where items cluster. Table 3 reports the standard error of measurement evaluated at $\theta^t_{0.90}$, the 90th percentile of each year's recovered cohort distribution which is computed using Equations 15 and 16. Second, to isolate intensive-margin changes from any year-to-year changes in cohort composition, Table 2 presents the precision at a fixed high-ability anchor $\theta^\star = \theta^{2011}_{0.90}$.

4.3 Upper-tail precision deteriorates at a fixed high-ability benchmark

Table 2 shows a post-2018 deterioration in precision at the fixed high-ability anchor. The standard error $SE_t(\theta^\star)$ is roughly stable through 2013, worsens modestly in 2014–2017, and then rises sharply after 2018. This pattern is the intensive-margin implication of the item-location drift: holding ability constant at a high level relevant for selective screening at elite institutions, later-year item pools deliver less Fisher information,

hence larger measurement error (Lord, 1980). The table also illustrates increasing concentration of information in a small subset of items (Top-10 share), consistent with a smaller effective set of high-discrimination, high-location items in later years. While this metric is influenced by pool size N_t, the magnitude of the observed increase exceeds what variation in N_t alone would predict.

Table 2: Fixed-point measurement precision at a high-ability anchor ($\theta^\star = \theta^{2011}_{0.90}$).

Year	N items	$\theta^\star$	$\mathbb{E}[I_t(\theta^\star)]$	$SE_t(\theta^\star)$	ΔSE	Median $I_{jt}(\theta^\star)$	Top-10 share
2011	94	2.953	5.942	0.410	+0.000	0.023	0.519
2012	139	2.953	6.062	0.406	-0.004	0.027	0.459
2013	73	2.953	6.301	0.398	-0.012	0.027	0.599
2014	128	2.953	5.448	0.428	+0.018	0.026	0.485
2015	69	2.953	4.372	0.478	+0.068	0.023	0.736
2016	116	2.953	5.236	0.437	+0.027	0.024	0.481
2017	81	2.953	4.835	0.455	+0.045	0.034	0.574
2018	125	2.953	3.344	0.547	+0.137	0.007	0.698
2019	126	2.953	4.199	0.488	+0.078	0.012	0.570
2020	73	2.953	2.201	0.674	+0.264	0.006	0.751
2021	73	2.953	3.275	0.553	+0.142	0.002	0.857
2022	120	2.953	1.409	0.843	+0.432	0.004	0.694
2023	74	2.953	2.958	0.581	+0.171	0.005	0.812

Notes: $\theta^\star$ is the 2011 90th-percentile ability on the latent scale. $\mathbb{E}[I_t(\theta^\star)]$ denotes expected test information under stratified uniform sampling without replacement from the year-t item pool, as defined in Section 3.4. $SE_t(\theta^\star) = 1/\sqrt{\mathbb{E}[I_t(\theta^\star)]}$ (Lord, 1980). ΔSE is measured relative to 2011. "Top-10 share" is the share of $\mathbb{E}[I_t(\theta^\star)]$ contributed by the ten most informative items at $\theta^\star$ in year t. This metric is mechanically influenced by pool size N_t: holding the item-information distribution fixed, smaller pools yield higher Top-10 shares. However, years with similar N_t (e.g., 2013 with $N = 73$ vs. 2023 with $N = 74$) show substantially different concentration levels, indicating that the pattern reflects substantive changes in item composition rather than pool-size artifacts alone.

Two features of the item-level information distribution clarify the mechanism behind the rise in $SE_t(\theta^\star)$. First, the median item information at $\theta^\star$ falls markedly after 2018, indicating that the typical item contributes little information at a high-ability benchmark, consistent with leftward drift in item locations and saturation of correct-response probabilities at $\theta^\star$. Second, information becomes increasingly concentrated: the ten most informative items account for a larger share of total information in later years, implying that upper-tail precision depends disproportionately on a thin subset of items. This concentration also implies sensitivity to effective item-pool coverage. When high-information items are under-represented in the observed pool, fixed-point precision can fall sharply even if mean item parameters change only modestly. I discuss this further in Section 5.

4.4 Within-year upper-tail precision and stylized classification error

Table 3 evaluates precision at the 90th percentile of each year's recovered cohort distribution, $\theta^t_{0.90}$, and translates it into a conditional misclassification probability for examinees near the threshold under the sym-

metric threshold rule described in Section 3. First, within-year upper-tail precision fluctuates substantially over time. It is notably worse in 2020 and best in 2022 in these diagnostics. Second, these within-year fluctuations do not contradict the fixed-anchor deterioration in Table 2: within-year $SE_t(\theta^t_{0.90})$ incorporates both the measurement technology and the location of the cohort upper tail. The fixed-anchor results isolate the measurement technology component by itself.

Table 3: Upper-tail standard errors at $\theta^t_{0.90}$ and stylized classification implications.

Year	$SE_t(\theta^t_{0.90})$	False neg.	False pos.	Total error	$\hat{\mu}_t$	$\hat{\sigma}_t$
2011	0.410	0.111	0.111	0.223	0.074	2.033
2012	0.407	0.110	0.110	0.219	0.063	2.097
2013	0.395	0.103	0.103	0.205	-0.010	2.148
2014	0.429	0.122	0.122	0.244	0.074	2.004
2015	0.497	0.157	0.157	0.315	0.014	2.178
2016	0.433	0.124	0.124	0.248	-0.042	2.258
2017	0.396	0.103	0.103	0.207	-0.736	2.217
2018	0.416	0.115	0.115	0.230	-1.616	2.425
2019	0.368	0.087	0.087	0.174	-1.611	2.740
2020	0.547	0.180	0.180	0.361	-2.097	3.032
2021	0.426	0.120	0.120	0.241	-2.129	3.094
2022	0.349	0.076	0.076	0.152	-2.613	2.590
2023	0.426	0.121	0.121	0.241	-2.663	3.273

Notes: $SE_t(\theta^t_{0.90})$ is evaluated at the 90th percentile of the recovered cohort distribution in year t. "Total error" is the conditional misclassification probability for examinees located at distance δ from the threshold, implied by the tolerance-band rule in Section 3; this is a local precision diagnostic rather than a system-wide error rate. By symmetry, false-positive and false-negative are identical. $(\hat{\mu}_t, \hat{\sigma}_t)$ are the cohort calibration parameters used to locate $\theta^t_{0.90}$ and to map item parameters to published score moments. Early-year calibration estimates show minor sensitivity to the optimization window; robustness to alternative weighting schemes is reported in Online Appendix D.

Institutional selectivity and academic match has meaningful consequences for educational attainment and labor market outcomes (Dillon and Smith, 2020; Hoxby and Avery, 2013; Chetty *et al.*, 2020). This literature distinguishes between *undermatch*, wherein students attend less selective institutions than their credentials would support and *overmatch* , wherein students attend more selective institutions than their credentials would support (Hoxby and Avery, 2012; Dillon and Smith, 2017). Table 3 provides under a stylized admissions rule that measurement errors in the screening instruments generate misclassification in both directions. False negatives tantamount to undermatch, while false positives correspond to overmatch. This diagnostic therefore is a measurement-based decomposition of match inefficiency distinct from behavioral channels such as information barriers, application costs or preference heterogeneity.

4.5 Human validation

To test whether the AI-implied difficulty ordering generalizes to human examinees, 26 business college freshmen from the University of Cincinnati each answered 48 SAT Math items (12 drawn from each of four

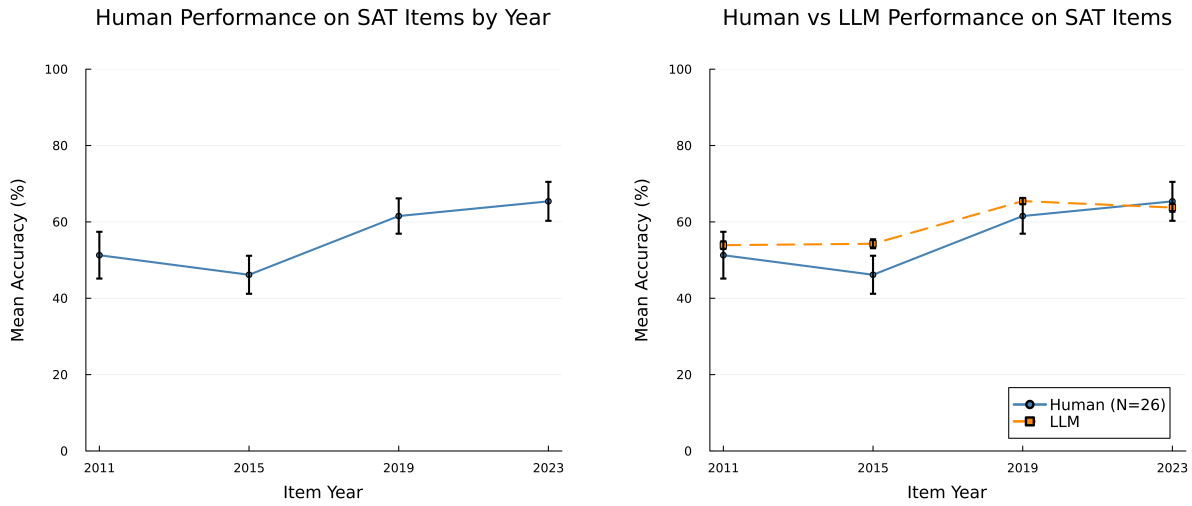

Figure 3: Human validation results: mean accuracy by item year. Error bars represent 95% confidence intervals. The upward trend in accuracy across item years shows the direction to be consistent with the AI-implied difficulty ordering. The University of Cincinnati's median SAT math score for admissions according to the College Scorecard in 2024 was 580.

years: 2011, 2015, 2019, 2023, items were randomized to mask sequential items) within a 70-minute session. Mean accuracy increases from 51.2% on 2011 items to 65.3% on 2023 items. The Jonckheere-Terpstra test for ordered alternatives yields $z = 4.16$ ($p < 0.0001$), confirming that human accuracy is stochastically increasing in item year (Jonckheere, 1954). The JT test is a nonparametric test that does not rely on variance estimation under clustering. A supplementary linear probability model with student fixed effects yields $\hat{\beta}_{2023} = 0.141$ (SE $= 0.035$, $p < 0.001$), indicating students score 14.1 percentage points higher on 2023 items relative to 2011; full regression results appear in the Online Appendix G[2]. These results establish that the difficulty ordering recovered from LLM responses aligns with human-experienced difficulty, given the collected sample.

5 Discussion

5.1 Interpreting test-optional outcomes through a measurement lens

Recent research documents shifts in selective admissions during the expansion of test-optional and test-free policies. Even in periods with a large pool of high-scoring applicants, elite institutions enrolled students with stronger high school grades but weaker standardized test scores (Sacerdote *et al.*, 2025). Other work documents "compression at the top," in which institutions placed less weight on SAT scores even among

[2]This study was exempted by the University of Cincinnati's IRB in accordance with 45 CFR 46.104. The estimate $\hat{\beta}_{2023}$ is from a linear probability model with student fixed effects and standard errors clustered at the student level.

applicants who submitted them (Avery *et al.*, 2025). These patterns are often interpreted as evidence of a deliberate pivot toward holistic assessment, a reweighting toward noncognitive attributes, or responses to equity objectives.

The analysis offers a complementary interpretation grounded in measurement. If the SAT's informativeness changes over time, both in where it is most precise on the latent ability scale and in the magnitude of measurement error at the upper tail, then a stable admissions rule that heavily weights test scores becomes mechanically less defensible. Under this view, "compression at the top" is consistent with an admissions office encountering a signal that is less discriminating in the region it cares about most. Importantly, this framework does not presume that institutions observe the latent objects estimated here. Rather, it provides a structured explanation for why de-emphasizing test scores can be a response to changes in the signal environment, even absent a shift in institutional preferences over applicant attributes.

Item drift and where the SAT is informative. A central empirical finding is that LLM-implied item difficulty declines over 2011–2023, while discrimination remains broadly stable. Under standard IRT mechanics, this combination implies a leftward shift in where items deliver maximal information (Birnbaum, 1968; Lord, 1980). The SAT Math section continues to differentiate examinees locally, but it does so at lower regions of the latent ability distribution.

This "where-information-lives" result is distinct from an average-score narrative. Public debates often focus on whether mean SAT scores rise or fall. Selective admissions decisions are sensitive to local signal precision near relevant cutoffs rather than to unconditional means. The measurement diagnostics therefore emphasize reliability in rank-ordering applicants at the upper tail, the margin at which elite institutions operate. The estimates therefore characterize the measurement properties of the observed in the item pools rather than the full operational forms. However, the main patterns emphasize are diagnostic in nature: at a fixed high-ability anchor at $\theta^\star = \theta^{2011}_{0.90}$ in Table 2, median item information falls and total information becomes increasingly concentrated in a small subset of items. These features are difficult to reconcile with purely random subsampling from a stable underlying pool and highlight that upper-tail precision is susceptible to the item set. Accordingly, the conclusions are evidence consistent with drift under the minimal public availability of the true item-pool and concentration diagnostics quantify this sensitivity. Under these constraints, using LLMs as benchmarking tools, the study provides a falsifiable diagnostics about measurement drift and signal informativeness.

Upper-tail precision and misclassification interpretation. To make the precision results interpretable in admissions language, I evaluate measurement error at two reference points: the upper tail of each year's recovered cohort distribution ($\theta^t_{0.90}$) and a fixed high-ability anchor ($\theta^\star = \theta^{2011}_{0.90}$). Within-year upper-tail precision fluctuates substantially and exhibits a deterioration in 2020. At the fixed anchor, preci-

sion deteriorates after 2018, reflecting an intensive-margin change driven by item-location drift rather than cohort composition.

Mapping these standard errors into a stylized threshold setting yields implied misclassification probabilities that vary widely across years. These calculations are not estimates of realized admissions mismatch, nor do they assume a single cutoff rule. Instead, they provide an intuitive mapping from measurement precision to classification reliability. The point is that modest changes in standard error near admissions-relevant regions can generate large changes in implied error rates, precisely the environment in which an admissions office might reduce reliance on test scores.

The diagnostics in Table 3 also decompose stylized misclassification into false positives and false negatives. Hoxby and Avery (2013) document high-achieving, low-income students are often undermatched at higher rates than affluent peers, due to information barriers. These students rely disproportionately on standardized tests to signal their ability. Dillon and Smith (2020) show that mismatch in either direction reduces graduation rates and earnings, while moderate mismatch has ambiguous effects. Further, Arcidiacono and Lovenheim (2016) review evidence that overmatch may harm students through grade penalties and STEM attrition. Bleemer (2022) however finds the California Prop 209's reduction in overmatch worsened outcomes for affected students, suggesting institutional quality and support systems can dominate match effects.

The measurement driven diagnostics provided in this paper offer a supply-side complement to the demand-side literature on mismatching. When upper tail precision declines, the threshold-based screening generated mismatch in both directions. The welfare consequences of these are asymmetric. If institutions respond to lower precision by downweighting test scores, the alternative signals are more resource-sensitive and AI can game them more easily (Sacerdote *et al.*, 2025; Friedman *et al.*, 2025; Cui *et al.*, 2025; Frankel and Kartik, 2019). This can exacerbate information barriers identified by Hoxby and Avery (2012, 2013), disproportionately harming disadvantaged, high-ability students.

2020 as a measurement stress test. The decline in upper-tail precision in 2020 coincides with documented disruptions in test administration and participation. It is therefore plausible that 2020 represented a year in which SAT Math scores became less reliable as screening inputs, even holding test design fixed.

This paper does not claim that test-optional adoption was caused by the measurement deterioration documented here. Institutions faced a bundle of simultaneous constraints in 2020–2021, including test-center closures, unequal access to testing dates, applicant stress, and equity concerns. Any of these could independently justify policy changes. If institutions perceived the score as less reliable or less comparably available across applicants, then de-emphasizing the SAT is a mechanically consistent response to a noisier signal.

The more interesting policy question concerns persistence: many test-optional policies remained in place

after 2020. A measurement perspective suggests that persistence need not reflect continued decline in precision. It may instead reflect institutional inertia, uncertainty about whether the 2020 disruption was temporary, selection into score submission, and reputational or political costs of reversing course.

Distributional implications and equity tradeoffs. A central concern raised in the test-optional literature is that de-emphasizing standardized scores may reduce enrollment of low-income, high-ability applicants, the group for whom a comparable metric can provide the strongest counterweight to weaker school resources or less polished holistic credentials (Sacerdote *et al.*, 2025; Chetty *et al.*, 2023; Friedman *et al.*, 2025). If the SAT becomes less informative at the upper tail, reliance on the score may weaken mechanically. But if alternative signals are even more resource-sensitive, abandoning the standardized signal can still worsen equity outcomes.

When test scores are down weighted, admissions processes implicitly reallocate weight toward grades, course-taking, recommendations, essays, and extracurricular portfolios. These signals known to vary systematically across schools and socioeconomic contexts (Chetty *et al.*, 2023; Frankel and Kartik, 2019). The measurement perspective reframes equity debates as questions about the relative precision and bias of the full signal portfolio used in admissions decisions.

5.2 Limitations and interpretation

Several limitations must be acknowledged. First, SAT item contents are proprietary, and the publicly available set through the internet archives is not guaranteed to be a random draw from the operational pool. The low fixed-anchor information in 2022 illustrates why coverage matters for high-ability precision. This pattern is mechanically consistent with the estimated leftward shift in item difficulty but may also reflect non-representative item availability. To guard against over-interpreting any single year, emphasis is placed on trends that are stable across adjacent years and across alternative benchmarking subset of models.

Second, item parameters are estimated from LLM response patterns rather than human microdata. They should be interpreted as benchmark-implied diagnostics under a maintained stability assumption, not as definitive psychometric estimates. While partial training-data exposure cannot be ruled out entirely, the consistency of trends across a heterogeneous set of open-source models mitigates concerns that results are driven by a single model or vendor (see Online Appendix A for LLM details). The training contamination robustness check by paraphrasing items also supports the suggestion that data leakage is not significantly present (see Online Appendix G for contamination check details).

Third, the recovered cohort parameters $(\hat{\mu}_t, \hat{\sigma}_t)$ are model-consistent constructions on a latent scale. They reconcile easier item locations with observed score moments but do not separately identify preparation

changes and compositional shifts in the test-taking population. These implied metrics are not to be considered as true estimates of ability changes in student cohorts. Finally, the standard error and misclassification calculations are mechanical implications of the estimated measurement technology. They do not incorporate the full admissions process, in which institutions combine multiple signals and use scores in non-threshold ways.

5.3 Implications and conclusion

Taken together, the results suggest that the test-optional policies can be understood partly as a response to changes in the signal environment. Where the SAT is most informative, and how reliable it is at the upper tail, varies over time and exhibits a disruption in 2020. While this paper does not identify a causal link from measurement precision to admissions policy, the measurement perspective reconciles post-2020 admissions patterns: institutions adjust the weight placed on signals when the marginal information content of those signals changes. As universities begin to reconsider the SAT's role in admissions, with GenAI polishing essays and recommendations, understanding the measurement properties of standardized tests remains crucial because they are the few signals free from gaming and resource biases.

One plausible mechanism for the observed item-location drift is endogenous adaptation by test designers. The College Board pretests items on sample cohorts and may adjust item pool to maintain target score distributions. If underlying math proficiency declines consistent with the NAEP and PISA trends, item pools would shift towards easier items to maintain score stability. The mapping from ability to scores adapts to the population, creating a Lucas critique analog in the measurement environment.

From an economic standpoint, upper tail precision is critical for students from disadvantaged backgrounds who rely on standardized tests to credibly signal high ability. This raises policy questions about the test-optional equilibria, human-capital investment incentives, and access to opportunity. The AI-audit approach illustrates how large language models can generate novel measurement diagnostics in data-scarce environments, informing future work on admissions and education policy.

The policy implication is not that standardized testing should be uncritically retained or discarded, but evaluated based on the quality of the signal it is able to transmit. Admissions should be evaluated as portfolios of signals, with attention to precision, comparability, and distributional consequences for the applicants most reliant on standardized measures.

References

ACT, Inc. (2025). 'Act profile report – national: Graduating class 2025', ACT, Inc.

Arcidiacono, P. and Lovenheim, M. (2016). 'Affirmative action and the quality-fit trade-off', *Journal of Economic Literature*, vol. 54(1), pp. 3–51, doi:10.1257/jel.54.1.3.

Avery, C., Shi, L. and Magouirk, P. (2025). 'Test-optional college admissions: ACT and SAT scores, applications, and enrollment changes', *NBER Working Paper*, (34260).

Betancourt, M. (2017). 'A conceptual introduction to hamiltonian monte carlo', *arXiv preprint*.

Birnbaum, A. (1968). 'Some latent trait models and their use in inferring an examinee's ability', in (F. M. Lord and M. R. Novick, eds.), *Statistical Theories of Mental Test Scores*)pp. 397–479, Reading, MA: Addison-Wesley.

Bleemer, Z. (2022). 'Affirmative action, mismatch, and economic mobility after california's proposition 209', *The Quarterly Journal of Economics*, vol. 137(1), pp. 115–160, doi:10.1093/qje/qjab029.

Chetty, R., Deming, D.J. and Friedman, J.N. (2023). 'Diversifying society's leaders? the determinants and causal effects of admission to highly selective private colleges', National Bureau of Economic Research, forthcoming, *Quarterly Journal of Economics*.

Chetty, R., Friedman, J.N., Saez, E., Turner, N. and Yagan, D. (2020). 'Income segregation and intergenerational mobility across colleges in the united states', *The Quarterly Journal of Economics*, vol. 135(3), pp. 1567–1633, doi:10.1093/qje/qjaa005.

College Board (2017). *SAT Suite of Assessments: Technical Manual*, College Board, New York, NY, characteristics of the SAT; psychometrics, scaling, equating, reliability, and validity.

College Board Research (n.d.). 'Previous concordances: Concordance tables comparing new sat and old sat scores', .

Cui, J., Dias, G. and Ye, J. (2025). 'Signaling in the age of AI: Evidence from cover letters', .

Cunha, F. and Heckman, J.J. (2007). 'The technology of skill formation', *American Economic Review*, vol. 97(2), pp. 31–47.

Cunha, F., Heckman, J.J. and Schennach, S.M. (2010). 'Estimating the technology of cognitive and noncognitive skill formation', *Econometrica*, vol. 78(3), pp. 883–931.

Dillon, E.W. and Smith, J.A. (2017). 'Determinants of the match between student ability and college quality', NBER Working Paper No. 19286.

Dillon, E.W. and Smith, J.A. (2020). 'The consequences of academic match between students and colleges', *Journal of Human Resources*, vol. 55(3), pp. 767–808, doi:10.3368/jhr.55.3.0318-9370R1.

Fox, J.P. (2010). *Bayesian Item Response Modeling: Theory and Applications*, Statistics for Social and Behavioral Sciences, Springer New York, ISBN 978-1-4419-0741-7, 978-1-4419-0742-4, doi:10.1007/978-1-4419-0742-4.

Frankel, A. and Kartik, N. (2019). 'Muddled information', *Journal of Political Economy*, vol. 127(4), pp. 1739–1776.

Friedman, J.N., Sacerdote, B., Staiger, D.O. and Tine, M. (2025). 'Standardized test scores and academic performance at ivy-plus colleges', *AEA Papers and Proceedings*, vol. 115, pp. 676–681.

Ge, H., Xu, K. and Ghahramani, Z. (2018). 'Turing: A language for flexible probabilistic inference', *Journal of Statistical Software*, vol. 84(10), pp. 1–32, doi:10.18637/jss.v084.i10.

Hambleton, R.K. and Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*, Evaluation in Education and Human Services Series, Boston / Dordrecht: Kluwer Academic Publishers / Springer Dordrecht, ISBN 0-89838-065-0, doi:10.1007/978-94-017-1988-9, 332 p., illus.

Holland, P.W. and Wainer, H. (1993). *Differential Item Functioning*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Hoxby, C.M. and Avery, C. (2012). 'The hidden supply of high-achieving, low-income students', NBER Working Paper No. 18586.

Hoxby, C.M. and Avery, C. (2013). 'The missing "one-offs": The hidden supply of high-achieving, low-income students', *Brookings Papers on Economic Activity*, vol. 2013(1), pp. 1–65, doi:10.1353/eca.2013.0000.

Jonckheere, A.R. (1954). 'A distribution-free k-sample test against ordered alternatives', *Biometrika*, vol. 41(1-2), pp. 133–145, doi:10.1093/biomet/41.1-2.133.

Kolen, M.J. and Brennan, R.L. (2014). *Test Equating, Scaling, and Linking: Methods and Practices*, New York, NY: Springer, 3 edn., doi:10.1007/978-1-4939-0317-7.

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*, Hillsdale, NJ: Lawrence Erlbaum Associates.

National Center for Education Statistics (2024). 'Naep report card: Mathematics 2024 – national results at grades 4, 8, and 12', U.S. Department of Education, Institute of Education Sciences, National Center for

Education Statistics, accessed January 2026; includes average scores and trends for Grades 4, 8, and 12 mathematics assessments.

Neal, R.M. (2011). 'Mcmc using hamiltonian dynamics', in (S. Brooks, A. Gelman, G. Jones and X.-L. Meng, eds.), *Handbook of Markov Chain Monte Carlo*)pp. 113–162, chap. 5, Chapman and Hall/CRC.

Nelder, J.A. and Mead, R. (1965). 'A simplex method for function minimization', *The Computer Journal*, vol. 7(4), pp. 308–313, doi:10.1093/comjnl/7.4.308.

Organisation for Economic Co-operation and Development (2023). 'Pisa 2022 results (volume i): The state of learning and equity in education', OECD Publishing, Paris, doi:10.1787/53f23881-en, u.S. mathematics literacy score and trends highlighted in national reports (NCES 2023-115).

Rothstein, J.M. (2004). 'College performance predictions and the SAT', *Journal of Econometrics*, vol. 121(1–2), pp. 297–317.

Sacerdote, B., Staiger, D.O. and Tine, M. (2025). 'How test-optional policies in college admissions disproportionately harm high-achieving applicants from disadvantaged backgrounds', *NBER Working Paper*, (33389).

Spence, M. (1973). 'Job market signaling', *Quarterly Journal of Economics*, vol. 87(3), pp. 355–374.

Wells, C.S., Subkoviak, M.J. and Serlin, R.C. (2002). 'The effect of item parameter drift on examinee ability estimates', *Applied Psychological Measurement*, vol. 26(1), pp. 77–87, doi:10.1177/0146621602026001005.

Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E. and Stoica, I. (2023). 'Judging LLM-as-a-judge with MT-Bench and chatbot arena', in (*Advances in Neural Information Processing Systems*)vol. 36.

Online Appendix

A Benchmark IRT specification and identification details

This appendix records the pooled Bayesian IRT model used to estimate item parameters from LLM response patterns. The core model is a 3PL for multiple-choice items and a 2PL restriction for numeric-response items, consistent with standard psychometric practice (Birnbaum, 1968; Lord, 1980; Kolen and Brennan, 2014).

Likelihood. Let $y_{mjtr} \in \{0, 1\}$ denote whether benchmark model m answers item j from year t correctly on attempt r. Attempt-specific effective ability is

$$\alpha_{mr} = \alpha_m + \sigma_{\text{rep}} \varepsilon_{mr}, \qquad \varepsilon_{mr} \sim \mathcal{N}(0, 1). \tag{17}$$

For multiple-choice items, the correctness probability is

$$P_{jt}(\alpha) = c_{jt} + (1 - c_{jt}) \Lambda(a_{jt}(\alpha - b_{jt})), \qquad \Lambda(x) = \frac{1}{1 + e^{-x}}. \tag{18}$$

For numeric-response items, I impose $c_{jt} \equiv 0$ (2PL). The pooled likelihood over all responses is

$$\mathcal{L}(\mathbf{\Theta} \mid \mathbf{Y}) = \prod_m \prod_{j,t} \prod_r P_{jt}(\alpha_{mr})^{\, y_{mjtr}} \left[1 - P_{jt}(\alpha_{mr})\right]^{1 - y_{mjtr}}. \tag{19}$$

Pooling all items across years in a single likelihood places (a_{jt}, b_{jt}, c_{jt}) on a common latent scale without cross-year linking, justified by a stable benchmark evaluated under fixed prompting (Kolen and Brennan, 2014).

Parameter	**Prior / distribution**	**Notes / hyperparameters**
Model ability (subset) θ_m	$\theta_m \sim \mathcal{N}(0, 2^2)$	$m = 1, \ldots, M_{\text{sub}}$ (subset-only)
Reprompt SD σ_{rep}	$\sigma_{\text{rep}} \sim \text{HalfNormal}(0, 0.5)$	Common reprompt noise across models
Item discrimination (log) $\log a_k$	$\log a_k \sim \mathcal{N}(0, 1^2)$	$k = 1, \ldots, K_{\text{sub}}$; implies $a_k = \exp(\log a_k)$
Item discrimination (level) a_k	$a_k = \exp(\log a_k)$	Log-normal implied prior; median = 1
Item difficulty b_k	$b_k \sim \mathcal{N}(0, 2^2)$	Wide prior spanning broad difficulty range
Pseudo-guessing (MCQ) c_k	$c_k \sim \text{Beta}(\alpha_k, \beta_k)$	For non-numeric items only; centered above random baseline
Pseudo-guessing (numeric) c_k	$c_k \equiv 0$	Numerical response items fixed at zero

Table 4: Priors used in the pooled Bayesian IRT model (subset specification) are weakly informative.

We center pseudo-guessing priors above the mechanical random-guess baseline (e.g., 1/5 pre-2016; 1/4 post-2016) to reflect that the benchmark population (LLMs) can occasionally eliminate options but still

guess, while preserving support on $(0, 1)$ (Gelman *et al.*, 2013, 2008).

Item type	Guessing prior	Mean	Notes
MCQ (5-option era)	$c_k \sim \text{Beta}(\alpha, \beta)$	≈ 0.21	Slightly above 0.20 random baseline
MCQ (4-option era)	$c_k \sim \text{Beta}(\alpha, \beta)$	≈ 0.26	Slightly above 0.25 random baseline
Numeric response	$c_k \equiv 0$	—	Fixed at zero (2PL)

Table 5: Pseudo-guessing parameter priors by item format.

Item types and guessing. The pseudo-guessing parameter is estimated for MCQ items and fixed to zero for numeric-response items. The estimated c_{jt} series shifts upward mechanically after the 2016 redesign (five-option to four-option MCQ), consistent with random-guess baselines (Lord, 1980).

Item characteristic curve validation. To assess whether LLM response patterns conform to the 3PL specification, I compare empirical item response functions to theoretical ICCs for items spanning the difficulty range. Figure 4 displays representative items that show variation in response patterns. Across these 1,236 items, the median correlation between observed and predicted response probabilities is 0.45. The monotonic relationship between model ability and accuracy generally holds, supporting the use of IRT as a measurement framework for directional assessment of item properties. Individual item fits show expected noise given the limited benchmark sample of 20 models.

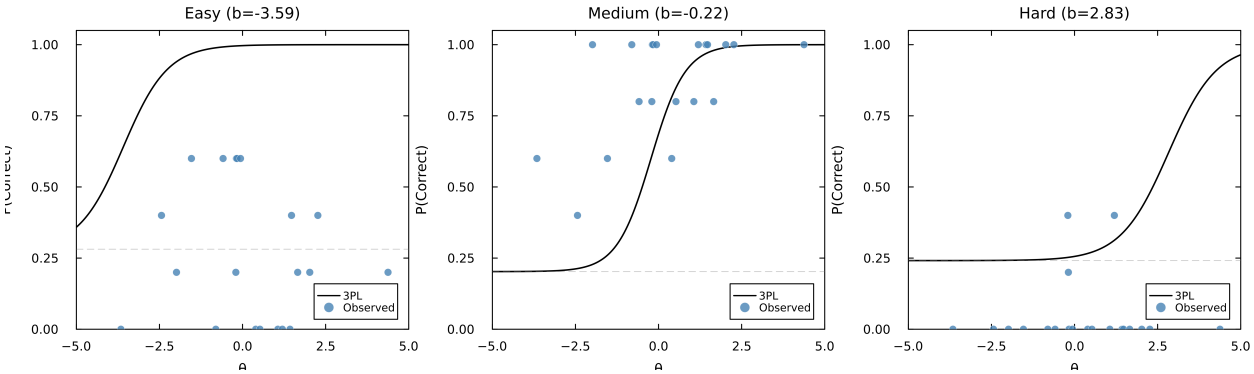

Figure 4: Empirical versus theoretical 3PL item characteristic curves for representative items at different difficulty levels. Points represent observed accuracy rates for each of 20 benchmark LLMs; curves show fitted 3PL predictions. Dashed horizontal lines indicate the guessing parameter.

B Benchmark model ability summaries and MCMC convergence

B.1 Benchmark model abilities

Table 6 reports posterior summaries for benchmark model abilities. Cross-model heterogeneity provides identification leverage for item parameters.

Model ID	Theta Mean	Theta Median	Theta SD	Theta Q025	Theta Q975	Theta ESS	Theta $\hat{R}$	Model Name
20	4.377	4.378	0.091	4.201	4.557	350.609	1.008	openai/gpt-oss-20b
7	2.266	2.264	0.067	2.139	2.399	159.086	1.016	deepcogito/cogito-v2-preview-deepseek-671b
4	2.023	2.021	0.065	1.897	2.148	152.033	1.012	Qwen/Qwen3-Coder-480B-A35B-Instruct-FP8
17	1.659	1.658	0.065	1.536	1.787	158.108	1.013	moonshotai/Kimi-K2-Instruct
18	1.474	1.474	0.063	1.355	1.596	139.516	1.013	moonshotai/Kimi-K2-Instruct-0905
3	1.428	1.427	0.064	1.310	1.562	137.437	1.011	Qwen/Qwen3-235B-A22B-Instruct-2507-tput
5	1.198	1.198	0.061	1.084	1.314	138.573	1.008	Qwen/Qwen3-Next-80B-A3B-Instruct
15	1.063	1.062	0.061	0.949	1.183	126.661	1.012	meta-llama/Meta-Llama-3.1-405B-Instruct-Turbo
9	0.524	0.524	0.059	0.414	0.641	122.146	1.013	deepcogito/cogito-v2-preview-llama-405B
2	0.396	0.397	0.061	0.278	0.515	129.601	1.010	Qwen/Qwen2.5-VL-72B-Instruct
14	-0.062	-0.061	0.062	-0.185	0.057	134.786	1.011	meta-llama/Llama-3.3-70B-Instruct-Turbo
6	-0.167	-0.167	0.060	-0.281	-0.052	118.812	1.012	arcee-ai/coder-large
1	-0.185	-0.186	0.059	-0.293	-0.065	121.065	1.011	Qwen/Qwen2.5-Coder-32B-Instruct
8	-0.202	-0.199	0.062	-0.322	-0.080	131.924	1.010	deepcogito/cogito-v2-preview-llama-109B-MoE
10	-0.586	-0.585	0.063	-0.710	-0.465	146.571	1.008	deepcogito/cogito-v2-preview-llama-70B
19	-0.805	-0.803	0.066	-0.933	-0.679	151.615	1.008	nvidia/NVIDIA-Nemotron-Nano-9B-v2
12	-1.538	-1.539	0.065	-1.665	-1.409	161.598	1.011	lgai/exaone-3-5-32b-instruct
13	-1.987	-1.988	0.069	-2.124	-1.850	194.245	1.007	lgai/exaone-deep-32b
11	-2.439	-2.440	0.074	-2.582	-2.296	216.296	1.006	google/gemma-3n-E4B-it
16	-3.660	-3.657	0.087	-3.833	-3.491	318.398	1.004	mistralai/Mistral-7B-Instruct-v0.3

Table 6: Posterior summaries of benchmark model ability parameters (α_m) with MCMC diagnostics.

B.2 Convergence diagnostics: $\hat{R}$ and ESS

Across 2011–2023 SAT Math items, HMC diagnostics indicate stable posterior recovery. The potential scale reduction statistic $\hat{R}$ compares within-chain and between-chain variation; values near one indicate chains are sampling from the same stationary distribution (Gelman and Rubin, 1992). Effective sample size (ESS) summarizes the Monte Carlo precision of correlated draws, capturing how many independent samples would deliver comparable accuracy (Gelman and Rubin, 1992).

For completeness, for a scalar parameter δ with m chains of length n post-warmup, define the between-chain variance

$$B = \frac{n}{m-1} \sum_{j=1}^{m} \left(\bar{\delta}_j - \bar{\delta}\right)^2, \tag{20}$$

and the average within-chain variance

$$W = \frac{1}{m} \sum_{j=1}^{m} s_j^2, \qquad s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(\delta_{ij} - \bar{\delta}_j\right)^2. \tag{21}$$

The marginal posterior variance estimator is

$$\widehat{\mathrm{Var}}(\delta) = \left(1 - \frac{1}{n}\right) W + \frac{1}{n} B, \tag{22}$$

and

$$\hat{R} = \sqrt{\widehat{\mathrm{Var}}(\delta)/W}. \tag{23}$$

For ESS, let N be the total post-warmup draws aggregated across chains and ρ_t the lag-t autocorrelation.

Then

$$N_{\text{eff}} \approx \frac{N}{1 + 2\sum_{t=1}^{\infty} \rho_t}. \tag{24}$$

Table 7: MCMC convergence diagnostics for item parameters (2011–2023 items).

Parameter	$\hat{R}$ (median)	$\hat{R}$ (p95)	$\hat{R}$ (max)	ESS (min)	ESS (p05)	ESS (median)	ESS (max)
Discrimination a	1.001	1.004	1.012	802	2268	3095	5484
Difficulty b	1.001	1.004	1.014	526	1099	2298	3842
Guessing c	1.001	1.005	1.011	855	2768	3955	6593

Table 8: Convergence diagnostics by question type (2011–2023 items).

Question type	Parameter	$\hat{R}$ (median)	$\hat{R}$ (p95)	$\hat{R}$ (max)	ESS (min)	ESS (p05)	ESS (median)	N items
MCQ	Discrimination a	1.001	1.004	1.012	802	2255	3109	1018
MCQ	Difficulty b	1.001	1.004	1.014	558	1095	2307	1018
MCQ	Guessing c	1.001	1.005	1.011	855	2768	3955	1018
Numeric	Discrimination a	1.001	1.004	1.007	1072	2294	3065	273
Numeric	Difficulty b	1.001	1.003	1.006	526	1125	2262	273

C Item parameter summaries by question type

Table 9 reports posterior summaries of item parameters separately for MCQ and numeric-response items. Discrimination is similar across types, while numeric-response items are modestly easier on average (more negative b). Guessing behaves as expected: present for MCQs and fixed at zero for numeric-response items (Lord, 1980).

Table 9: Posterior summaries of item parameters by question type (2011–2023 items).

		Discrimination a				Difficulty b				Guessing c			
Question type	N	Mean	SD	Q25	Q75	Mean	SD	Q25	Q75	Mean	SD	Q25	Q75
MCQ	1018	1.529	0.718	1.005	1.936	−0.285	2.431	−2.053	1.312	0.242	0.042	0.217	0.274
Numeric	273	1.544	0.701	1.074	1.864	−0.455	2.405	−2.176	1.118	0.000	0.000	0.000	0.000

D Score construction: raw scoring, concordance, and simulated moment matching

To compare scaled scores before and after the 2016 redesign, I replicate period-specific scoring rules and map simulated raw scores to official SAT Math scaled scores using published raw-to-scale conversion tables. Con-

cordance mapping is applied for pre-2016 scaled scores to express all outcomes on the post-2016 concordance scale (Kolen and Brennan, 2014).

Raw scoring. Let K_t denote the number of items in year t. For a simulated examinee i, let C_{it}, W_{it}, and O_{it} denote correct, wrong, and omitted counts with $K_t = C_{it} + W_{it} + O_{it}$. Before 2016, the wrong-answer penalty implies

$$R_{it} = C_{it} - 0.25\, W_{it}, \tag{25}$$

and R_{it} is rounded to the nearest integer before applying the published integer-to-scale table. Since 2016, the raw score is $R_{it} = C_{it}$.

Concordance. For pre-2016 administrations, scaled scores are transformed using official concordance maps to the post-2016 concordance scale (Kolen and Brennan, 2014). Post-2016 scores are already on the concordance scale.

Simulated moment matching. For each year t, define observed concorded score moments

$$M_t^{\text{obs}} = \Big(\bar{s}_t, \text{sd}(s_t), Q_{0.25}(s_t), Q_{0.50}(s_t), Q_{0.75}(s_t)\Big), \tag{26}$$

where s_t denotes the cross-sectional distribution of concorded SAT Math scores and $Q_p(\cdot)$ denotes the pth quantile. Given the year-t item set J_t and estimated parameters $\widehat{\phi}_{jt} = (\hat{a}_{jt}, \hat{b}_{jt}, \hat{c}_{jt})$, and given scoring/concordance rules, for a candidate cohort distribution $\theta \sim \mathcal{N}(\mu, \sigma^2)$ I simulate a synthetic cohort $\{\theta_i\}_{i=1}^N$ (with $N \approx 10^5$) and draw item responses

$$Y_{ijt} \sim \text{Bernoulli}\Big(\hat{c}_{jt} + (1 - \hat{c}_{jt})\Lambda(\hat{a}_{jt}(\theta_i - \hat{b}_{jt}))\Big)\,. \tag{27}$$

For pre-2016 years, the simulation incorporates a probabilistic omission rule: simulated examinees omit items where their perceived probability of correctness falls below 0.20, the breakeven threshold for five-option MCQs under the wrong-answer penalty. Attempted items with $Y_{ijt} = 0$ count as wrong (W_{it}); omitted items contribute zero to the raw score. For post-2016 years, all items are attempted and no penalty applies.

I then compute the simulated concorded score moments $M_t^{\text{sim}}(\mu, \sigma)$.

The objective uses squared relative errors for scale comparability across moments:

$$\text{err}(m) = \left(\frac{m^{\text{sim}} - m^{\text{obs}}}{m^{\text{obs}}}\right)^2, \tag{28}$$

and minimizes

$$(\hat{\mu}_t, \hat{\sigma}_t) = \arg\min_{\mu,\sigma>0} L_t(\mu, \sigma), \tag{29}$$

with

$$L_t(\mu, \sigma) = w_{\text{mean}}\Big(\text{err}(\bar{s}) + \text{err}(Q_{0.50})\Big) + w_{\text{others}}\Big(\text{err}(\text{sd}) + \text{err}(Q_{0.25}) + \text{err}(Q_{0.75})\Big). \tag{30}$$

The baseline specification sets $w_{\text{mean}} = 4$ and $w_{\text{others}} = 1$. Because the raw-to-scale mapping is discrete with redesign-era discontinuities, the objective is non-smooth; optimization therefore uses Nelder–Mead (Nelder and Mead, 1965). For reproducibility, simulations are seeded deterministically by year.

	Calibration: 2× weight on mean			**Calibration: 4× weight on mean**		
Year	$\hat{\mu}_t$	$\hat{\sigma}_t$	Loss	$\hat{\mu}_t$	$\hat{\sigma}_t$	Loss
2011	0.312	2.016	1.01e-06	0.074	2.033	0.00833
2012	0.094	2.138	0.00181	0.063	2.097	0.00306
2013	-0.019	2.147	0.00180	-0.010	2.148	0.00259
2014	0.078	1.992	0.000541	0.074	2.004	0.000541
2015	-0.039	2.232	0.000988	0.014	2.178	0.000893
2016	-0.051	2.259	2.16e-06	-0.042	2.258	2.97e-07
2017	-0.739	2.213	1.38e-05	-0.736	2.217	2.53e-05
2018	-1.616	2.425	1.41e-09	-1.616	2.425	7.26e-10
2019	-1.611	2.740	7.82e-09	-1.611	2.740	1.78e-09
2020	-2.194	3.068	0.000521	-2.097	3.032	0.000435
2021	-2.129	3.094	2.59e-10	-2.129	3.094	1.60e-09
2022	-2.612	2.590	1.02e-08	-2.613	2.590	3.01e-09
2023	-2.663	3.273	0.000288	-2.663	3.273	0.000288

Table 10: Robustness of cohort calibration to loss-function weighting.

Notes: The two calibration runs use different optimization windows: the 4× specification optimizes from 2011 onward (matching the main analysis), while the 2× specification optimizes from 2008 onward with only 2011+ years reported. Loss values are therefore not directly comparable across columns, as they reflect different underlying objective functions. Parameters that appear identical are subject to rounding; full-precision estimates differ slightly.

E Test information and fixed-anchor precision diagnostics

Test information and standard error follow standard IRT definitions (Lord, 1980). Let $\phi_{jt} = (a_{jt}, b_{jt}, c_{jt})$. For a given year-$t$ item pool J_t, item information at ability θ is

$$I_{jt}(\theta \mid \phi_{jt}) = \frac{\big(P'_{jt}(\theta)\big)^2}{P_{jt}(\theta)\,(1 - P_{jt}(\theta))}, \qquad P_{jt}(\theta) = c_{jt} + (1 - c_{jt})\Lambda(a_{jt}(\theta - b_{jt})). \tag{31}$$

To reflect uncertainty over the realized test form drawn from the year-t pool, expected test information

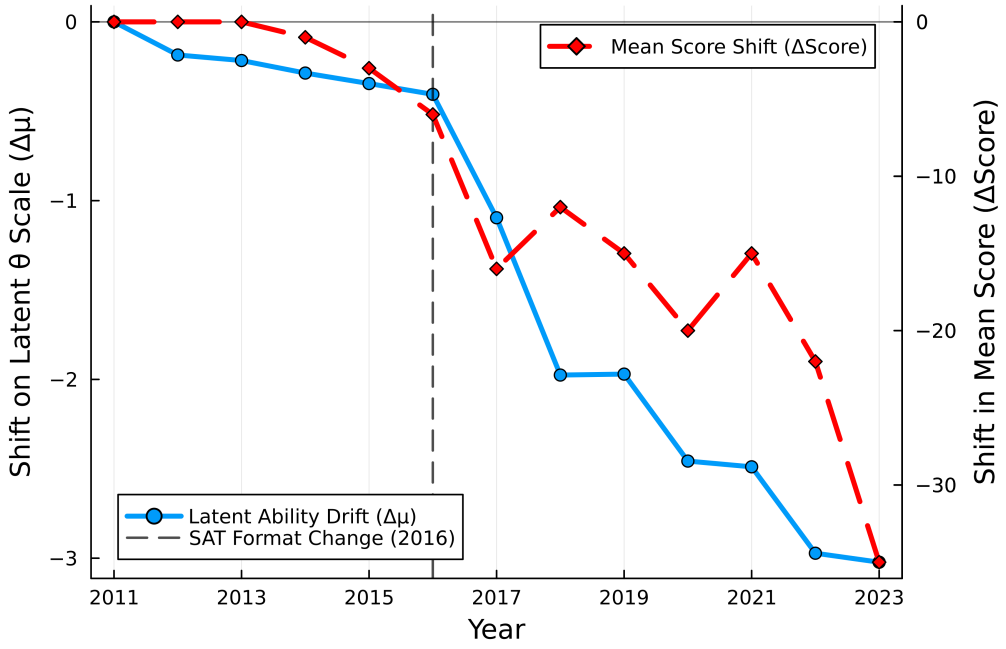

Figure 5: This figure shows two trends overlaid for comparison: (1) the estimated cohort mean ability changes relative to baseline year 2011, $\Delta\mu_t$ (blue line, left axis) and (2) the average SAT Math score changes relative to 2011 (red dashed line, right axis). Both series exhibit similar downward trends over 2011–2023. Using the IRT parameters recovered from LLM responses, the estimated cohort mean ability $\hat{\mu}_t$ captures shifts in the underlying ability distribution needed to reproduce observed score population trends. $\hat{\mu}_t$ are not estimates of human test-taker ability per se, but reconcile observed score trends with stable item parameters under the IRT model.

is computed under stratified uniform sampling without replacement:

$$\mathbb{E}[I_t(\theta)] = \frac{n_{\text{mcq},t}}{N_{\text{mcq},t}} \sum_{j \in J_t^{\text{mcq}}} I_{jt}(\theta) + \frac{n_{\text{ans},t}}{N_{\text{ans},t}} \sum_{j \in J_t^{\text{ans}}} I_{jt}(\theta), \tag{32}$$

with implied standard error

$$SE_t(\theta) = \frac{1}{\sqrt{\mathbb{E}[I_t(\theta)]}}. \tag{33}$$

Table 11 reports fixed-point precision at the high-ability anchor $\theta^\star = \theta_{0.90}^{2011}$ and documents concentration in item-level information at that anchor.

Table 11: Fixed-point measurement precision and item-level information diagnostics at a high-ability anchor.

Year	N items	$\theta^\star$	$\mathbb{E}[I_t(\theta^\star)]$	$SE_t(\theta^\star)$	ΔSE	Median $I_{jt}(\theta^\star)$	Top-10 share
2011	94	2.953	5.942	0.410	+0.000	0.023	0.519
2012	139	2.953	6.062	0.406	-0.004	0.027	0.459
2013	73	2.953	6.301	0.398	-0.012	0.027	0.599
2014	128	2.953	5.448	0.428	+0.018	0.026	0.485
2015	69	2.953	4.372	0.478	+0.068	0.023	0.736
2016	116	2.953	5.236	0.437	+0.027	0.024	0.481
2017	81	2.953	4.835	0.455	+0.045	0.034	0.574
2018	125	2.953	3.344	0.547	+0.137	0.007	0.698
2019	126	2.953	4.199	0.488	+0.078	0.012	0.570
2020	73	2.953	2.201	0.674	+0.264	0.006	0.751
2021	73	2.953	3.275	0.553	+0.142	0.002	0.857
2022	120	2.953	1.409	0.843	+0.432	0.004	0.694
2023	74	2.953	2.958	0.581	+0.171	0.005	0.812

Notes: $\theta^\star$ is fixed at the 2011 90th-percentile ability, $\theta^\star = \theta_{0.90}^{2011}$. $\mathbb{E}[I_t(\theta^\star)]$ is expected test information computed from the year-t item pool using stratified sampling weights (MCQ vs. numeric) implied by SAT section composition; $SE_t(\theta^\star) = 1/\sqrt{\mathbb{E}[I_t(\theta^\star)]}$ (Lord, 1980). ΔSE is relative to 2011. Median $I_{jt}(\theta^\star)$ is the median item information at $\theta^\star$ across items in year t. Top-10 share is the fraction of $\sum_j I_{jt}(\theta^\star)$ contributed by the 10 highest-information items in that year.

F Robustness: eight-model replication subset

As a robustness check, year-level item-parameter summaries are recomputed using an eight-model subset selected from the middle of the benchmark ability distribution shown in Table 6. The difficulty series closely reproduces the full-benchmark pattern, indicating that downward drift in $\bar{b}_t$ is not driven by a small number of extreme models. Discrimination comparisons are mechanically attenuated under the narrower ability range used for the subset, consistent with the identification of slope parameters in IRT (Lord, 1980).

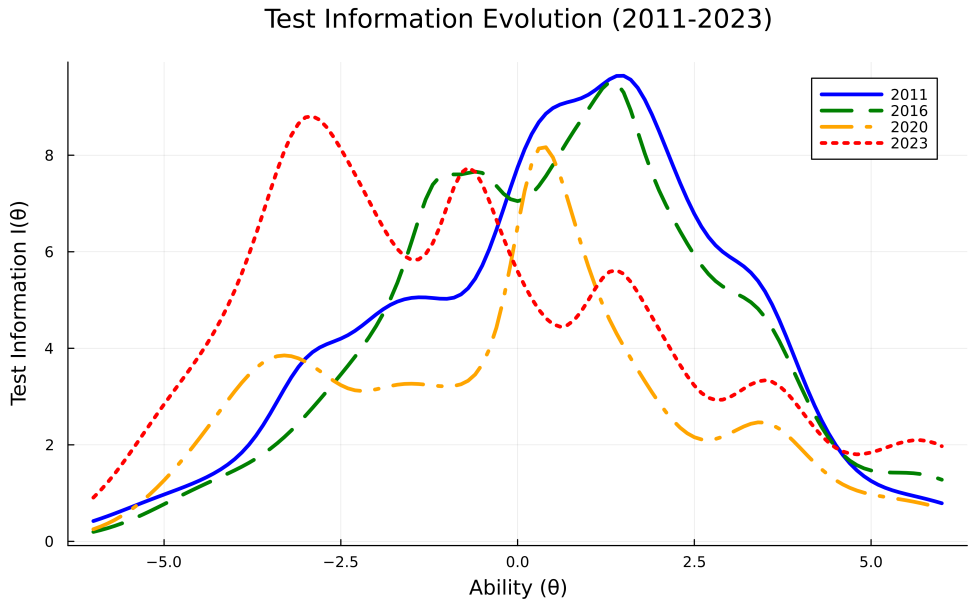

Figure 6: Expected test information curves for select years. Test information is computed under stratified uniform sampling without replacement from the year-t item pool, reflecting the actual SAT Math section composition. Over time, the peak of test information moves leftward, consistent with downward drift in item difficulty.

Table 12: Item-parameter trends: full 20-model benchmark versus 8-model replication subset (2011–2023).

Year	$\bar{a}$ (Full)	$\bar{a}$ (8-model)	$\bar{b}$ (Full)	$\bar{b}$ (8-model)
2011	1.511	1.135	0.191	0.498
2012	1.489	1.258	0.149	0.698
2013	1.436	1.310	0.013	0.261
2014	1.398	1.261	-0.061	0.234
2015	1.353	1.225	0.134	0.244
2016	1.533	1.313	0.228	0.324
2017	1.530	1.375	0.254	0.248
2018	1.661	1.463	-0.737	-0.576
2019	1.578	1.405	-0.567	-0.202
2020	1.531	1.507	-0.838	-0.628
2021	1.764	1.575	-0.793	-0.583
2022	1.564	1.379	-1.373	-1.210
2023	1.562	1.380	-0.738	-0.606

Table 13: Agreement between full 20-model benchmark and 8-model replication (year-level means, 2011–2023).

Parameter	Pearson r	Spearman ρ	Slope (Rep on Full)	Intercept	R^2	N (years)
Discrimination (a)	0.763	0.830	0.863	0.031	0.582	13
Difficulty (b)	0.969	0.907	1.022	0.225	0.940	13

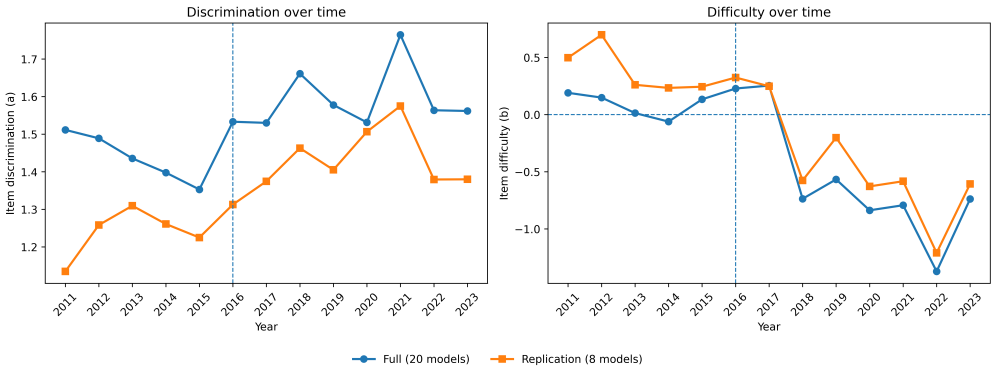

Figure 7: Item-parameter trends for the eight-model replication subset (2011–2023). The left panel shows average discrimination $\bar{a}_t$ by year; the right panel shows average difficulty $\bar{b}_t$ by year. The difficulty trend closely matches that of the full 20-model benchmark, indicating that downward drift in $\bar{b}_t$ is not driven by a small number of extreme models. Discrimination comparisons are attenuated under the narrower ability range used for the subset, consistent with IRT identification of slope parameters.

G Robustness: contamination check

To assess whether LLM responses reflect memorization of training data rather than genuine problem-solving, I conduct a contamination robustness check using paraphrased items. A stratified subsample of 224 items (54 items for years 2011, 2015 and 58 items for years 2019, 2023 to match one test form) is paraphrased using GPT-4o-mini to preserve mathematical content while altering surface-level wording. Six mid-ability models (the same models from the 8 model robustness check in Appendix F excluding 2 deprecated models, the Arcee Coder Large and Qwen2.5 72B Instruct, at the time of running this contamination robustness in late January 2026) from the benchmark are then evaluated on both original and paraphrased versions. Figure 8 shows that accuracy on original items (54.3%) and paraphrased items (54.0%) is nearly identical, with year-level differences ranging from -1.8 to $+1.9$ percentage points. This pattern is inconsistent with memorization, which would predict systematically higher accuracy on original items.

Secondly, common crawl data used for LLM training is filtered, homogenized and run through several quality filters. The item set used in this study were transcribed manually from very low quality scanned PDFs of official test booklets, making it unlikely that exact copies exist in training data. The paraphrasing process and robustness results further mitigate concerns about contamination.

Paraphrasing prompt. Items were paraphrased using GPT-4o-mini and Moonshot's Kimi K2 Instruct 0905 with the following system prompt:

> You rewrite SAT math questions to test memorization vs understanding.
>
> Rules:
>
> 1. Keep the EXACT same mathematical problem—same numbers, same relationships, same answer

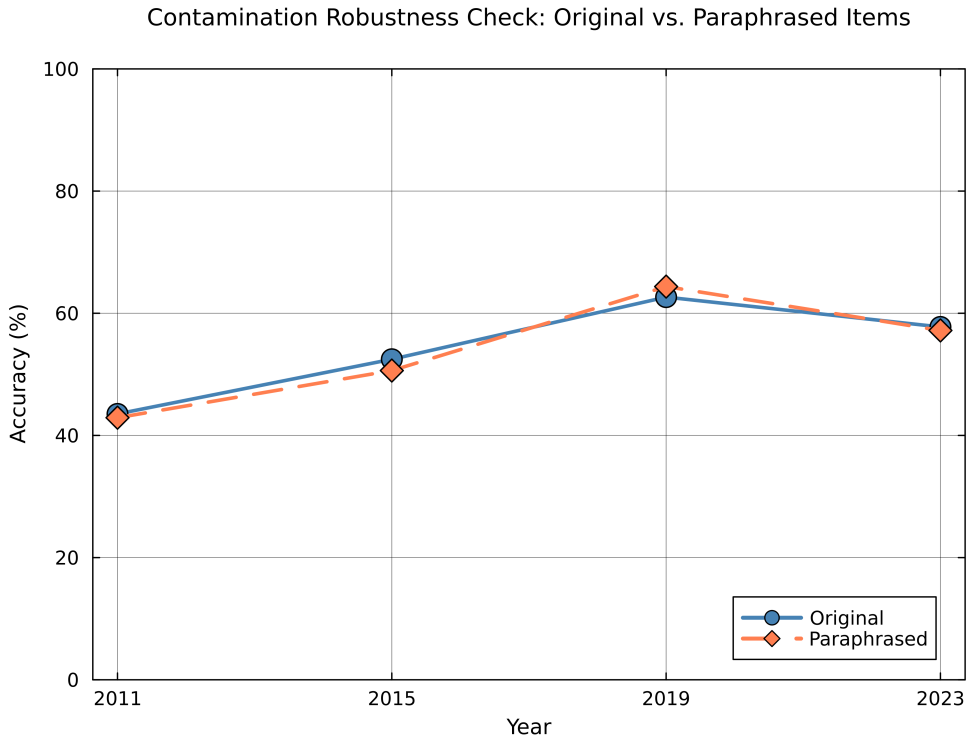

Figure 8: Contamination robustness check: accuracy on original versus paraphrased items. Six mid-ability models from Table 6 were evaluated on 224 items (54 items before 2016 and 58 items for years after 2016, representing one full test form appropriate to the year) in both original and paraphrased form. The near-identical accuracy rates suggest LLM responses reflect problem-solving rather than memorization. The items were sampled stratified by question type (multiple-choice versus numeric response) and year using a fixed seed 42 for reproducibility.

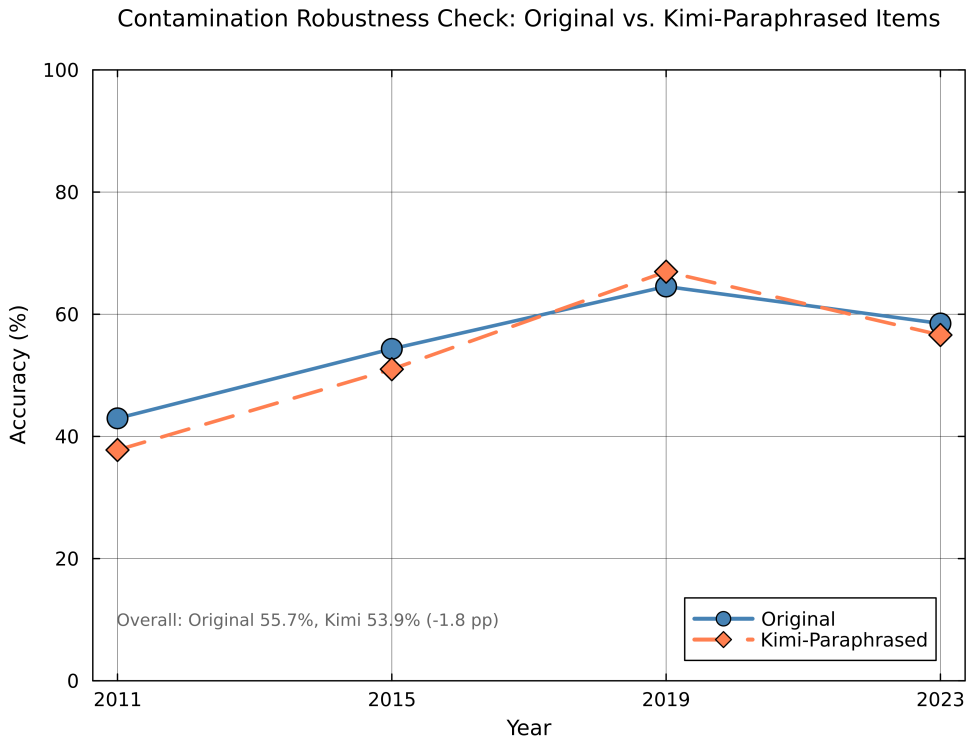

Figure 9: This contamination check uses paraphrased items from Moonshot AI's Kimi K2 Instruct 0905 model. The results are identical to the GPT-4o-mini paraphrased items. While a few items were excluded due to parsing errors resulting in 203 items being tested over 4 years compared to the 224 items in the GPT-4o-mini check in Figure 8. The items were sampled stratified by question type (multiple-choice versus numeric response) and year using a fixed seed 42 for reproducibility.

2. Significantly rephrase the wording—use different sentence structures, synonyms
3. Keep it concise—don't make it longer than the original
4. Preserve all mathematical notation, equations, and numerical values exactly
5. If there's a word problem context (e.g., "John has 5 apples"), you may change names/objects but keep quantities

The user prompt appended each original question with the instruction: "Paraphrase this SAT math question." Generation used temperature 0.7 to encourage lexical diversity while preserving mathematical coherence. Structured outputs using JSON mode ensured consistent formatting across all 224 items.

H Human Validation

H.1 Sample and Administration

The validation sample consists of 26 freshmen students admitted in 2025 from the University of Cincinnati, Lindner College of Business. Each student answered 48 SAT Math items in a single 70-minute session, approximating the time constraints of the actual SAT Math section. Items were drawn using a fixed seed: 12 items from each of four years (2011, 2015, 2019, 2023), stratified by question type (10 multiple-choice and 2 numeric response). All students received items in the same fixed order, which was randomized across item years so that year-blocks were interleaved rather than sequential, mitigating systematic fatigue or learning effects correlated with item year. Participants were not informed of item year or the purpose of the exercise. Students received course credit for participation and no identifying information was collected.

H.2 Primary Analysis: Jonckheere-Terpstra Test

The hypothesis is that human accuracy is stochastically increasing in item year, consistent with the LLM-implied difficulty decline documented in Table 1. I test this using the Jonckheere-Terpstra (JT) test for ordered alternatives, a nonparametric procedure that evaluates whether k groups are stochastically ordered under the alternative hypothesis.

Let Y_{it} denote the accuracy of student i on items from year t. The JT test statistic is

$$J = \sum_{t<t'} U_{tt'}, \tag{34}$$

where $U_{tt'}$ is the Mann-Whitney U statistic comparing groups t and t'. Under the null hypothesis that

accuracy distributions are identical across years, J is approximately normal with

$$E[J] = \frac{N^2 - \sum_t n_t^2}{4}, \quad \text{Var}(J) = \frac{N^2(2N+3) - \sum_t n_t^2(2n_t+3)}{72}, \tag{35}$$

where N is the total number of observations and n_t is the sample size for year t. Here, each student contributes one accuracy observation per year, so $N = 26 \times 4 = 104$ student-year observations with $n_t = 26$ for each year. The standardized test statistic $z = (J - E[J])/\sqrt{\text{Var}(J)}$ is referred to the standard normal distribution.

Table 14 reports mean accuracy by year. Accuracy increases from 51.2% (2011) to 65.3% (2023), with a modest dip in 2015 (46.1%) that is not statistically distinguishable from 2011. The JT test yields $J = 2744.5$, $E[J] = 2028.00$, $z = 4.16$, $p < 0.0001$ (one-sided), rejecting the null in favor of the ordered alternative.

Table 14: Human Accuracy by Item Year

Year	N responses	Mean accuracy	SE	95% CI
2011	312	0.512	0.031	[0.452, 0.574]
2015	312	0.461	0.025	[0.412, 0.511]
2019	312	0.615	0.024	[0.569, 0.661]
2023	312	0.653	0.026	[0.602, 0.705]

Notes: Each cell represents 26 students × 12 items = 312 item-responses. Standard errors are clustered at the student level. The 95% confidence intervals are computed as $\bar{y} \pm 1.96 \times \text{SE}$. The JT test uses student-level accuracy (one observation per student-year, $N = 104$), not the item-level response counts shown here.

H.3 Supplementary Analysis: Linear Probability Model

To provide interpretable effect sizes, I estimate a linear probability model with student fixed effects:

$$y_{ij} = \alpha_i + \beta_{2015} \cdot \mathbf{1}[t_j = 2015] + \beta_{2019} \cdot \mathbf{1}[t_j = 2019] + \beta_{2023} \cdot \mathbf{1}[t_j = 2023] + \varepsilon_{ij}, \tag{36}$$

where $y_{ij} \in \{0, 1\}$ indicates whether student i answered item j correctly, α_i is a student fixed effect, and $t_j \in \{2011, 2015, 2019, 2023\}$ denotes item year with 2011 as the reference category.

Table 15 reports the results. Students score 14.1 percentage points higher on 2023 items relative to 2011 items ($p < 0.001$). The 2015 coefficient is negative but not statistically significant, consistent with the descriptive pattern in Table 14.

Table 15: Linear Probability Model: Correctness on Item Year

	Coefficient	SE (clustered)	t	p-value
Year = 2015	−0.051	0.034	−1.33	0.145
Year = 2019	0.103	0.035	3.08	0.007
Year = 2023	0.141	0.035	3.99	< 0.001
Student FE		Yes		
Observations		1,248		
Students		26		
Items		48		

Notes: Reference category is 2011. Standard errors are clustered at the student level. Each student answered 48 items (12 per year).

H.4 Inference Considerations

Standard errors in Table 15 are clustered at the student level to account for within-student correlation across items. Because all students received items in a common fixed order, within-item correlation across students is also plausible. Two-way clustering following Cameron *et al.* (2011) yields substantially larger standard errors and renders the year coefficients statistically insignificant, reflecting the limited number of clusters in both dimensions (25 students, 48 items). This however may be overly conservative, and not an object of primary interest, since the goal is to validate the directional ordering of item difficulty rather than precise effect-size estimation.

The Jonckheere-Terpstra test, which is nonparametric and does not rely on variance estimation under clustering, provides the primary evidence for directional concordance. The linear probability model complements this by quantifying the magnitude of the accuracy difference, while precise inference is limited by the sample size.

H.5 Interpretation

The validation establishes that the overall easing trend recovered from LLM responses—with early-year items harder and 2023 items clearly easiest—aligns with human-experienced difficulty in the sample collected. In the human data, 2011 and 2015 are statistically similar in difficulty (both harder than later years), while 2019 and 2023 show markedly higher accuracy. The 14 percentage point accuracy gap between 2011 and 2023 items, combined with the significant JT test for ordered alternatives, indicates that the AI benchmark recovers the directional pattern of difficulty change relevant to human examinees, even if the precise year-to-year ordering differs slightly.

I Prompting protocol and templates

All items are evaluated under a fixed prompting protocol with deterministic decoding. For MCQs, option order is randomized as a robustness check to option-position effects; metadata records the shuffle mapping.

I.1 Multiple-choice prompt template

System / instruction header: You are solving an SAT math multiple-choice question.

Question:

{QUESTION_TEXT}

Options:

a) {OPTION_A}

b) {OPTION_B}

c) {OPTION_C}

d) {OPTION_D}

{e) {OPTION_E} *(pre-2016 only)*}

Task: Select the correct answer.

I.2 Numeric-response prompt template

System / instruction header: You are solving an SAT math question that requires a numerical answer.

Question:

{QUESTION_TEXT}

Task: Give the numerical answer. For ranges, use standard mathematical notation like "$1<x<3$" or "$x\geq 5$". For single numbers, provide the exact value.

I.3 Prompt-generation code (for reproducibility)

Code used to assemble prompts and metadata:

```
def generate_multiple_choice_prompt(row, shuffle=False):
"""Generate prompt for multiple choice SAT questions"""
question = row['question']
year = row['year']
original_answer = row['answer']
```

```
# Determine number of options based on year
if year < 2016:
options = [('a', row['option_a']),
('b', row['option_b']),
('c', row['option_c']),
('d', row['option_d']),
('e', row['option_e'])]
option_labels = ['a', 'b', 'c', 'd', 'e']
else:
options = [('a', row['option_a']),
('b', row['option_b']),
('c', row['option_c']),
('d', row['option_d'])]
option_labels = ['a', 'b', 'c', 'd']

original_answer_lower = original_answer.lower()
correct_answer_after_shuffle = original_answer_lower

if shuffle:
# Find correct option content
correct_option_content = None
for label, content in options:
if label == original_answer_lower:
correct_option_content = content
break

# Shuffle option contents only (labels fixed)
option_contents = [content for _, content in options]
random.shuffle(option_contents)
options = [(label, content)
for label, content in zip(option_labels, option_contents)]

# Find new label for correct content
for new_label, content in options:
if content == correct_option_content:
```

```
correct_answer_after_shuffle = new_label
break

prompt = f"""You are solving an SAT math multiple-choice question.

Question:
{question}

Options:
"""
for label, content in options:
prompt += f"{label}) {content}\n"
prompt += "\nSelect the correct answer."

metadata = {
'year': row['year'],
'source_doc': row['source_doc'],
'question_num': row['question_num'],
'question_type': row['question_type'],
'section_num': row['section_num'],
'correct_answer': correct_answer_after_shuffle,
'original_answer': original_answer,
'shuffled': shuffle,
'manual_review': False
}
return prompt, metadata

def generate_numerical_prompt(row):
"""Generate prompt for numerical answer SAT questions"""
question = row['question']
prompt = f"""You are solving an SAT math question that requires a numerical answer.

Question:
{question}
```

```
    Give the numerical answer.
    For ranges, use standard mathematical notation like '1 < x < 3' or 'x > 5'.
    For single numbers, provide the exact value."""

    metadata = {
    'year': row['year'],
    'source_doc': row['source_doc'],
    'question_num': row['question_num'],
    'question_type': row['question_type'],
    'section_num': row['section_num'],
    'correct_answer': row['answer'],
    'shuffled': False,
    'manual_review': True
    }
    return prompt, metadata
```

Online Appendix References

Birnbaum, A. (1968). 'Some latent trait models and their use in inferring an examinee's ability', in (F. M. Lord and M. R. Novick, eds.), *Statistical Theories of Mental Test Scores*)pp. 397–479, Reading, MA: Addison-Wesley.

Cameron, A.C., Gelbach, J.B. and Miller, D.L. (2011). 'Robust inference with multiway clustering', *Journal of Business & Economic Statistics*, vol. 29(2), pp. 238–249.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B. (2013). *Bayesian Data Analysis*, Boca Raton, FL: Chapman and Hall/CRC, 3 edn., ISBN 9781439840955, doi:10.1201/b16018.

Gelman, A., Jakulin, A., Pittau, M.G. and Su, Y.S. (2008). 'A weakly informative default prior distribution for logistic and other regression models', *The Annals of Applied Statistics*, vol. 2(4), pp. 1360–1383, doi:10.1214/08-AOAS191.

Gelman, A. and Rubin, D.B. (1992). 'Inference from iterative simulation using multiple sequences', *Statistical Science*, vol. 7(4), pp. 457–472, doi:10.1214/ss/1177011136.

Kolen, M.J. and Brennan, R.L. (2014). *Test Equating, Scaling, and Linking: Methods and Practices*, New York, NY: Springer, 3 edn., doi:10.1007/978-1-4939-0317-7.

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Nelder, J.A. and Mead, R. (1965). 'A simplex method for function minimization', *The Computer Journal*, vol. 7(4), pp. 308–313, doi:10.1093/comjnl/7.4.308.