Innocent Until Proven Improbable

Detecting a structural break in a million UNAM admission scores

Samantha H. · CS109 Challenge Project · August 2026

1. Background and Motivation

In 2026 UNAM, one of the largest universities in Mexico, administered its undergraduate admission exam online for the first time, and the mean score rose from 56 to 71 [1]. The exam consists of 120 multiple-choice questions with four options each. With roughly 147,000 students taking the exam annually and an admission rate of about 10% this 15-point shift sparked immediate public debate regarding whether the results were valid.

For this project, my goal is to mathematically quantify this anomaly. Since it is a fact that the distribution moved, I wanted to measure whether it moved further than distributions normally do between typical years. Using a dataset of results from 2021 to 2026 [2], my aim is to measure how far 2026 departs from the historical baseline. Because I am working with score totals rather than item-level responses, this analysis cannot identify the specific cause or flag individual students; instead, it focuses purely on detecting a structural break in the population distribution.

2. Data Preprocessing and Baseline Construction

The data set contains 1,027,716 records from 2021 to 2026. To filter valid exam attempts, I applied two filtering rules. First, I removed rows with missing scores and scores of exactly zero. Because the exam consists of 120 questions with four options each with no penalty for wrong answers, the probability of scoring a zero of pure guessing is 0.75^{120} , making a true zero statistically impossible and likely indicates an unsubmitted exam. Second, to maintain consistency, I restricted the analysis to the 107 programmes present across all six years retaining $\geq 99.45\%$ of the applicants from every year (Appendix A).

I used data from 2021 to 2025 to model what a typical year looks like. To quantify the shift between distributions, I chose Total Variation Distance (TVD), which represents the fraction of students that would need to be moved to make two histograms identical. When comparing each pair of the historical years (2021-2025), the TVD ranged from 0.0202 to 0.0440, with an average distance of 0.0304 (Figure A1).

I pooled the five historical years to build a robust baseline. To prevent demographic shifts from skewing the results, I averaged each programme's historical distribution but combined them using 2026 applicant counts. This rules out changes in programme demand as a factor. Finally, to prevent data leakage, I used strict code assertions to guarantee no 2026 scores informed the baseline, models, or estimates.

3. How an Honest Score Is Made

I modeled the baseline because a model expresses the shift in terms of actual questions known rather than pure statistical distance, and it separates genuine knowledge from random guessing.

I iteratively built three models sharing the same Beta-Binomial structure: a student draws a skill level $p \in [0, 1]$ from a Beta(α, β) distribution and answers 120 questions, each independently correct with probability $q(p)$. The models only change in how the skill p translates into the final success probability. The score is modeled using $X \sim \text{Binomial}(120, q(p))$.

I fitted the three models using Maximum Likelihood Estimation (MLE) optimized by exhaustive grid search. For confidence intervals, I bootstrapped 200 replicates by resampling students with replacement, rebuilding the baseline, and refitting the models.

Model 1 assumes students only get right exactly what they know $q(p) = p$.

While this reproduced the expected population variance, the fit was poor. The TVD to the baseline was 0.1688 (5.6 times higher than the variance between two ordinary years). It also placed students below a score of 30. Since the expected value of pure guessing is 30, this model clearly failed by ignoring chance.

Model 2 adjusted the probability to account for guessing $q(p) = p + (1 - p) \cdot \frac{1}{4}$.
This model fixes the left tail, now a student knowing nothing expects a score of 30, matching the data. The TVD dropped to 0.0535, and skewness improved from 0.062 to 0.601. However, the baseline data's true skewness is 0.794. The model still fails because the $\frac{1}{4}$ guessing rate was assumed rather than derived from the data.

Model 3 frees the guessing probability to be estimated by the model: $q(p) = p + (1 - p) \cdot c$.
Here, the MLE estimated c = 0.2748 (3.64 options on a four-option paper). This model found this without being told how many options existed for each question. The TVD fell to 0.0325 and skewness rose to 0.721 making it the closest fit to the baseline (Figure A2).

As a final check, I removed the parametric mechanism entirely using Non-Parametric MLE (NPMLE), leaving the skill distribution free to take any shape. The model naturally recovered the guessing floor: 99.96% of the fitted mass fell above exactly $\frac{1}{4}$ (Figure A3). At a distance of 0.0036, Model 3 closes 82.5% of the theoretical gap that a perfect model could achieve on finite data.

4. Analyzing the 2026 Shift

Everything to this point was built without reference to 2026. The 2026 mean score is 71.18, against a range of 54.90 to 56.36 in the historical range. Standard deviation expanded to 25.18, against the historical ranges of 18.56 to 19.74. The skewness dropepd to −0.013; effectively disappearing compared to the historical range of 0.764 to 0.838. The 2026 distribution isn't just shifted; it took on a different shape.

The TVD from 2026 to the baseline is 0.2934. The TVDs to the individual historical years range from 0.2863 to 0.3157, all falling completely outside normal variance. This is roughly ten times the distance normally seen between two ordinary years. This TVD indicates that at least 40,347 students (29.34% of the 2026 cohort) do not follow the baseline distribution.

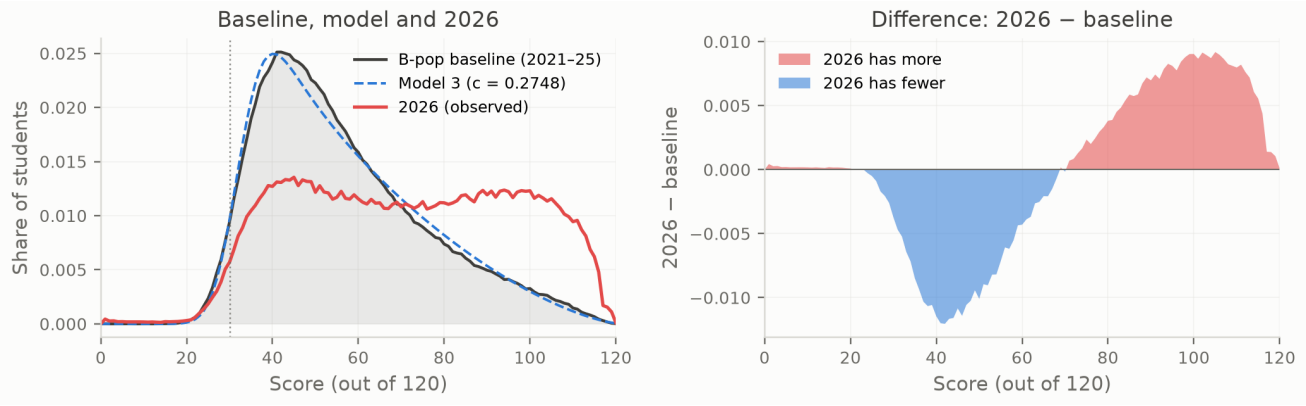

Figure 1. Model 3 tracks the honest baseline closely; 2026 does not. The right panel shows where the mass went..

Tracking the mass displacement shows exactly where the distribution broke. The first middle part of the scale (scores 21 to 68) lost 40,323 students, while the second part (scores 69 to 120) gained 39,810. The bottom tail also grew: scores 1 to 20 gained 513 students, and 124 students scored exactly zero, compared to zero or one in every prior year. An easier exam or a stronger applicant pool would naturally produce fewer low scores. Neither explains the form 2026 results produced. This rules out innocent explanations where the cohort simply performed better, pointing instead to anomalies in the online format itself.

5. Testing Alternative Hypotheses

I tested two explanations that do not require assuming systemic issues: either the entire cohort was simply better, or an abnormally strong subgroup joined the applicant pool.

Hypothesis 1 - the cohort was better

To test this, I refitted Model 3 to the 2026 distribution to see if there were any parameter combination that could reproduce its shape. I ran three versions: ability free (guessing rate fixed to the honest baseline), everything free, and everything free with the guessing rate limited to $c \geq \frac{1}{4}$.

For this hypothesis to be true, the average applicant would need to know 51.57 questions, this represents a 61% jump from the historical baseline of 32.08. Even if we accept that jump, the model fails to fit the 2026 shape. The best TVD it reachesis 0.0722, twice as bad as its fit on an ordinary year (0.0325). When I left the parameters free, the MLE estimated a guessing rate of 0.1589, implying students performed worse than random guessing, When c was limited to be $\frac{1}{4}$ or more the model predicted a skewness of 0.129, missing the actual one of -0.013. No valid parameter reproduces the 2026 curve (Figure A4).

Hypothesis 2 - a stronger subgroup joined (Mixture Model)

Next, I modeled 2026 as a mixture of ordinary students and a new, unknown group, solving for what that unknown group's distribution would have to look like. (Appendix D)

Since a population cannot have a negative number of students at any score, this constraint requires the new group to make up at least 48% of the cohort (about 66,000) students with a mean score of 87.5 and only 16% of this theoretical group would score between 30 and 70, a range that historically captures 76% of the applicants (Figure A5). A real population of capable students still spreads out naturally. A subgroup that is almost entirely absent from the middle of the scale is not realistic. If I try to adjust the math to make this group look more like a normal distribution, its size jumps to 65% of the cohort, which makes the scenario less realistic.

The innocent explanations do not fail because a model rejects them. They fail on arithmetic.

6. What This Does Not Show

This analysis shows a structural break, not its cause. An easier set of questions, the online format, or irregular access would all break the pattern in the exact same way, and score totals alone just can't separate them.

It also doesn't identify anyone. Without item-level responses or timing data, a score of 110 in 2026 looks exactly like a 110 in 2023. The 40,347 figure is a floor, not a true estimate. A stricter bound of about 66,000 also mathematically holds, but it rests on a single density ratio. I chose to report the lower floor because it averages across all 121 possible scores, making it much more robust to noise.

A score does carry information, for example, among the 1,298 students who scored 110, about one in nine is consistent with a historical year; the other eight are not. But nothing in the data tells us which is which. If I classified students simply by whichever side is mathematically more likely, it would label 42% of the cohort as anomalous. However, that rule treats a wrong accusation and a missed detection as equally costly, which is just not true in an admissions context.

Ultimately, this framework detects and quantifies an anomaly. It does not catch anyone.

References

[1] El País (México), *Examen de control UNAM: quiénes deberán repetir la prueba de ingreso, fechas y sedes de aplicación*, 5 August 2026. https://elpais.com/mexico/2026-08-05/examen-de-control-unam-quienes-deberan-repetir-la-prueba-de-ingreso-fechas-y-sedes-de-aplicacion.html

[2] A. G. (agn3si), *ResultadosUNAM—data: UNAM undergraduate admission results, 2021–2026*, CC BY 4.0. https://github.com/agn3si/ResultadosUNAM---data

LLM use: Claude was used to implement the analysis code, to discuss methodology, and to review drafts of this write-up, including suggesting phrasings that I then rewrote. The argument, the structure and the final text are my own.

A. Data and exclusions

year	registered	no score	score of 0	analysed
2021	167,432	26,986	1	140,445
2022	177,780	24,188	0	153,592
2023	178,404	25,544	1	152,859
2024	162,463	23,178	0	139,285
2025	174,543	25,110	1	149,432
2026	167,094	28,688	124	138,282
2021–25	860,622	125,006	3	735,613

The 107-programme panel retains at least 99.45% of every year's applicants. The zero-score rule was adopted after 2026's low scores had been inspected; it removes 3 students from the five honest years combined and 124 from 2026, and changes the honest baseline by a TVD of 0.000011.

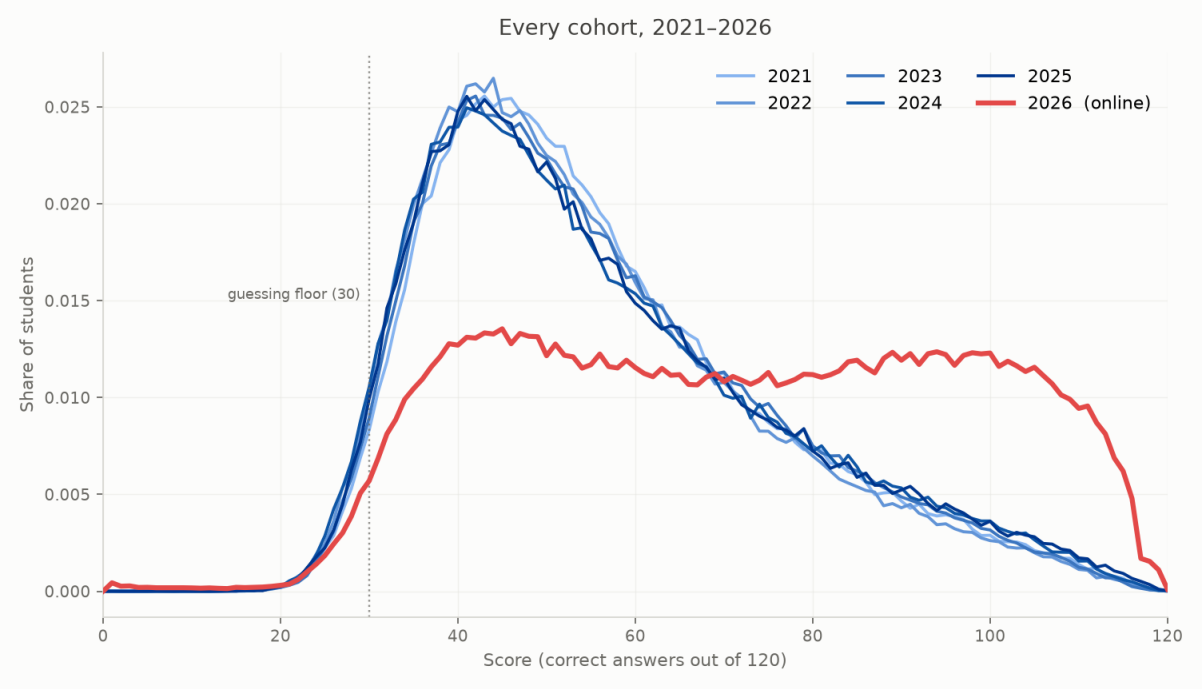

Figure A1. The five in-person years are nearly indistinguishable from one another; 2026 is in red.

B. Model formulas

The shared structure. A student draws a skill p from a Beta(α, β) and answers 120 questions, each correct with probability $q(p)$:

$$P(X = k) = \int_0^1 \binom{120}{k} q(p)^k \big(1 - q(p)\big)^{120-k} \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)}\, dp$$

Model 1. With $q(p) = p$ the integral has an exact solution and no numerical integration is needed:

$$P(X = k) = \binom{120}{k} \frac{B(k + \alpha,\ 120 - k + \beta)}{B(\alpha, \beta)}$$

Models 2 and 3. A student knows i of the 120 questions and guesses the remaining $120 - i$, each correct with probability c. To score k they must guess $k - i$ of those right:

$$P(X = k) = \sum_{i=0}^{k} \underbrace{\text{BetaBinom}(i; 120, \alpha, \beta)}_{\text{knows } i} \cdot \underbrace{\binom{120 - i}{k - i} c^{k-i} (1 - c)^{120-k}}_{\text{guesses } k-i \text{ of the rest right}}$$

with $c = \frac{1}{4}$ fixed in Model 2 and estimated in Model 3. Because c does not depend on p, this is an exact finite sum rather than an integral.

C. Fitted models

	$q(p)$	params	α	β	c	TVD	skew	BIC
1 Beta-Binomial	p	2	4.4215	4.9356	—	0.1688	0.062	6,419,776
2 fixed $\frac{1}{4}$	$p + (1-p)\frac{1}{4}$	2	1.1570	2.7960	$\frac{1}{4}$ fixed	0.0535	0.601	6,278,609
3 fitted c	$p + (1-p)c$	3	0.8991	2.4644	0.2748	0.0325	0.721	6,271,067
observed							0.794	

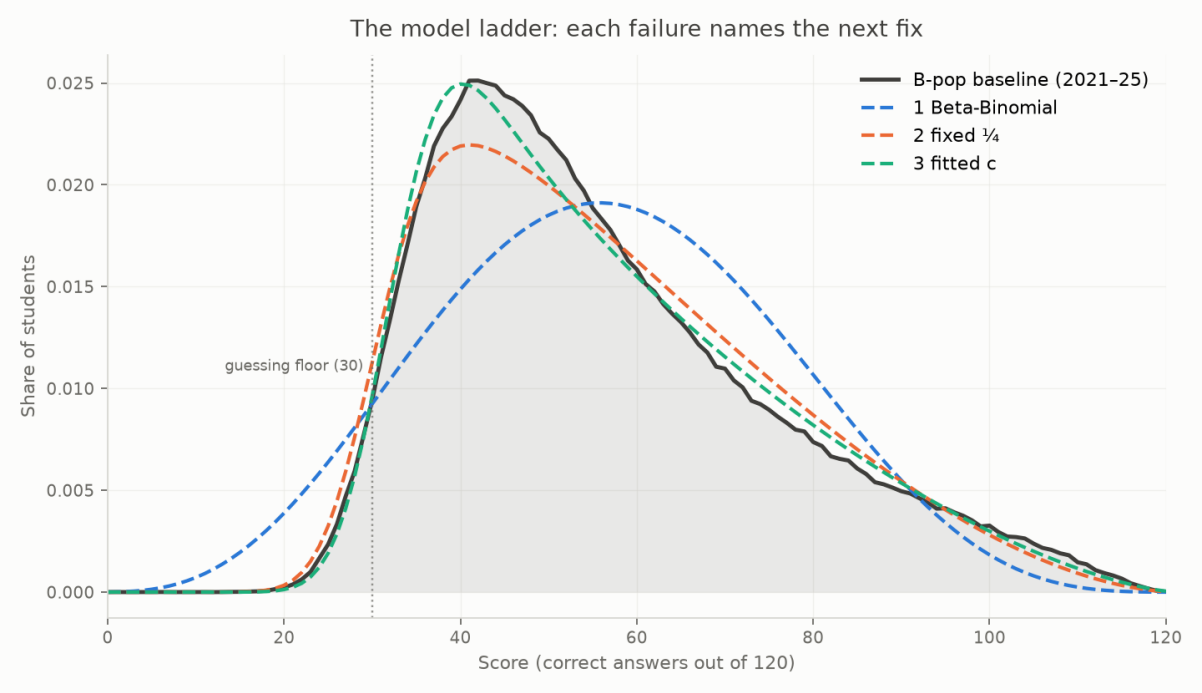

Figure A2. Only Model 1 places students below the guessing floor of 30, where the data has almost none. Each model's failure names what the next one changes.

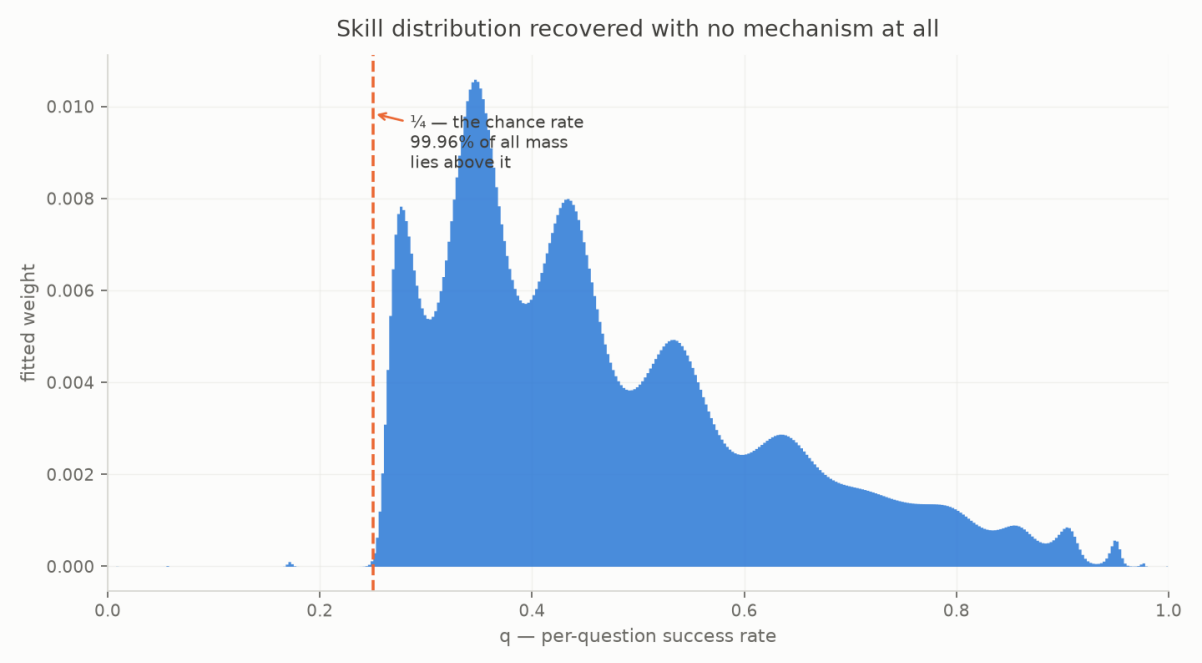

Figure A3. With the mechanism removed and the skill distribution left free to take any shape, 99.96% of the fitted mass falls above exactly $\frac{1}{4}$. Nothing in the fitting told the model how many options the exam has.

D. Comparing 2026 to the baseline

Displacement. Write the difference between the two distributions as $\delta(k) = f_{2026}(k) - f_{\text{honest}}(k)$. Both sum to one, so the differences cancel exactly:

$$\sum_{k=0}^{120} \delta(k) = 0, \qquad \sum_{k:\,\delta(k)>0} \delta(k) \;=\; -\sum_{k:\,\delta(k)<0} \delta(k) \;=\; \text{TVD}$$

Mass that leaves one region reappears in another in equal amount, and that common amount is the TVD. Multiplying by the 137,524 students who sat in 2026 turns it into a headcount.

A bound that assumes nothing about the other group. Suppose 2026 is a mixture of students distributed as the baseline and some other group:

$$f_{2026}(k) = (1-\pi)\, f_{\text{honest}}(k) + \pi\, g(k)$$

where g is any distribution and π is the share of the cohort belonging to it. Substituting into the definition of TVD,

$$\text{TVD}\big(f_{2026},\, f_{\text{honest}}\big) = \tfrac{1}{2}\sum_{k} \big|\pi\, g(k) - \pi\, f_{\text{honest}}(k)\big| = \pi \cdot \text{TVD}\big(g,\, f_{\text{honest}}\big) \;\le\; \pi$$

since a total variation distance never exceeds one. So $\pi \ge \text{TVD} = 0.2934$: at least 29.34% of the cohort is not distributed as the baseline, whatever the other group looks like.

Explanation 1: refitting to 2026. Model 3 was refitted to f_{2026} by the same grid search on the same objective used for the honest years, under three sets of constraints:

- A α and β free, c held at the honest value 0.2748;
- B α, β and c all free;
- B′ α, β and c free with $c \ge \frac{1}{4}$, since guessing at random is always available and no student can do worse than chance on an item they do not know.

Explanation 2: solving for the other group. Rearranging the mixture gives what that group would have to be:

$$g(k) = \frac{f_{2026}(k) - (1-\pi)\, f_{\text{honest}}(k)}{\pi}$$

No group contains a negative number of students, so $g(k) \geq 0$ at every score. That requires

$$1 - \pi \;\leq\; \min_{k} \frac{f_{2026}(k)}{f_{\text{honest}}(k)} \;=\; 0.5202, \qquad \text{so} \qquad \pi \;\geq\; 0.4798$$

This bound is stricter than the TVD one but is not used as the headline: it depends on a single density ratio, while the TVD bound averages over all 121 scores and is therefore far less sensitive to sampling noise.

E. Uncertainty

Bootstrap intervals, 200 replicates. Each replicate resamples students inside every programme–year cell, rebuilds the baseline from scratch, and refits the model.

model	α	β	TVD
1	4.4215 [4.4053, 4.4385]	4.9356 [4.9148, 4.9587]	0.1688 [0.1678, 0.1699]
3	0.8991 [0.8887, 0.9091]	2.4644 [2.4502, 2.4774]	0.0325 [0.0316, 0.0335]

For Model 3, $c = 0.2748$ with a 95% interval of [0.2742, 0.2753], and $E[K] = 32.08$ with a 95% interval of [31.98, 32.16].

F. Fitting procedure

All models were fitted by exhaustive grid search over a log-spaced window $[0.05, 60]$ on both α and β, 120 points per axis, re-centred and narrowed six times. A grid enumerates rather than searches, so it cannot settle on a local optimum, and it is deterministic. The optimum was checked not to lie on a window boundary, and agreed with Nelder–Mead to zero relative error. Model 3 was fitted by profile likelihood: α and β were fitted at each value of c on a grid from 0.10 to 0.45, then refined near the peak.

G. The two innocent explanations

	A: ability free	B: everything free	B′: free, $c \geq \frac{1}{4}$
c	0.2748 (held)	0.1598	0.2500 (on the bound)
effective options	3.64	6.26	4.00
$E[K]$ known	51.57	62.03	54.15
increase on 32.08	+61%	+93%	+69%
best TVD	0.0722	0.0878	0.0573
best skew	0.202	−0.071	0.129
2026's skew		−0.013	

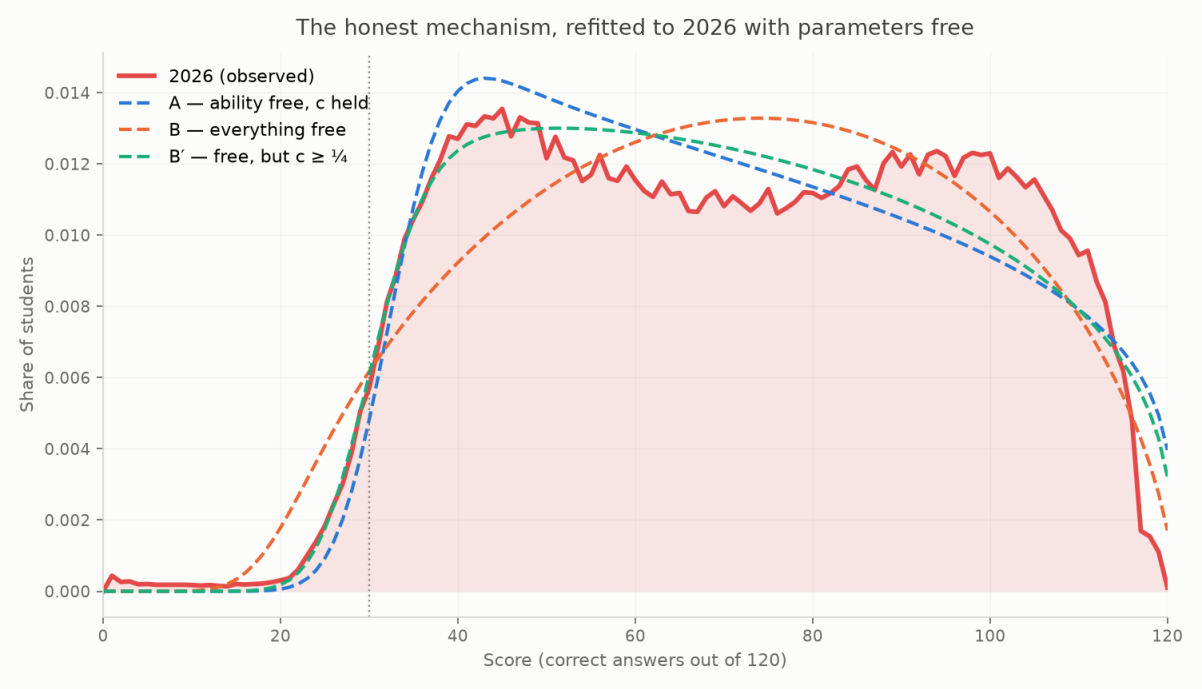

Figure A4. Every refit of the honest mechanism produces a hump. 2026 is a plateau, and no parameter values reach that shape.

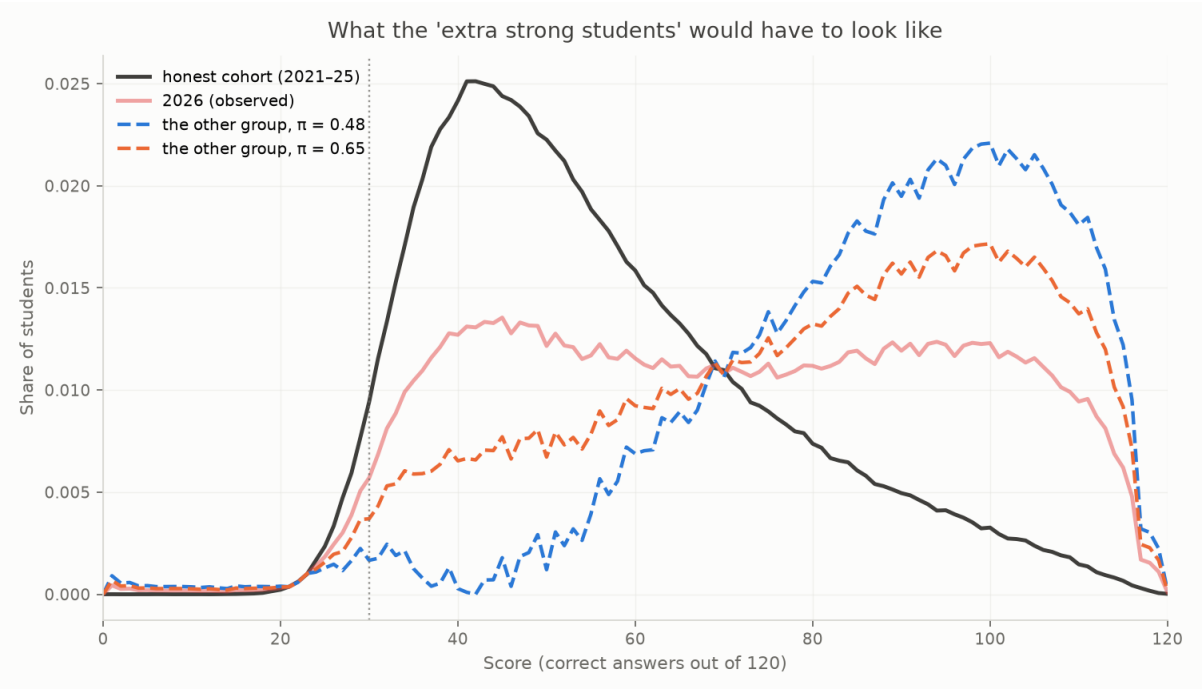

Figure A5. For the second explanation to hold, the other group would need almost no students scoring around 40, where ordinary students are most common.

H. Calibration

reference point	TVD
sampling noise — a model that is exactly true	0.0043
NPMLE ceiling — best any model of this shape can do	0.0036
a real year vs the five-year baseline	0.0194
honest year ↔ honest year (the yardstick)	0.0304
Model 3	0.0325
Model 2	0.0535
Model 1	0.1688
2026 vs the baseline	0.2934