Imperfect Metacognition in Standardized Tests

Francesco Bilotta*

September 21, 2026

First version: January 18, 2025.
Click here for the most recent version.

Abstract

This paper argues that standardized tests measure not only what students know, but also whether they know what they know and what they do not know: their metacognition. I develop a framework that recovers metacognition from selection choices. Since answering is costly, students with good metacognition should omit the questions they are least likely to answer correctly, improving their accuracy. Applying the framework to a large field experiment, I find that students largely depart from this benchmark, because many students sort questions nearly at random and a sizable minority sort them adversely. Yet, sorting varies substantially in the population, with systematic gender differences. I quantify these patterns using a structural model of metacognitive ability, and decompose scores into knowledge and metacognition components. The findings suggest that interpreting scores as signals of knowledge alone entails sizable information loss, and that metacognitive differences may help explain why gender gaps in test-taking widen with students' ability.

Keywords: Metacognition, Overconfidence, Standardized Tests, Higher-Order Skills, Gender Gaps.

*Department of Economics, Bocconi University; francesco.bilotta2@phd.unibocconi.it. Previous titles: Introspection and Decision-Making; Imperfect Metacognition and Task Selection; Imperfect Metacognition in Multiple Choice Tests. I am especially grateful to my advisors Nicola Pavoni, Massimo Morelli and Luca Braghieri for their constant guidance and encouragement, and to Nagore Iriberri and Pedro Rey-Biel, who generously shared their data. I thank Jerome Adda, Jose Apesteguia, Martino Banchio, Erich Battistin, Roland Bénabou, Raphael Boulat, Daniele Caliari, Chris Carnehl, Alessandro Castagnetti, Sylvain Chassang, Carlo Cusumano, Stefano Della Vigna, Xavier d'Haultfoeuille, Sarah Eichmeyer, Astrid Esparza Sánchez, Francesco Fabbri, Martin Fankhauser, Pamela Giustinelli, Katrin Gödker, Nenad Kos, Yves Le Yaouanq, Alessandro Lizzeri, Giacomo Manferdini, Paola Manzini, Marco Mariotti, Pietro Ortoleva, Marco Ottaviani, Filippo Palomba, Wolfgang Pesendorfer, Collin Raymond, Tobias Salz, and Victor Sancibrian for helpful conversations, as well as participants at Northwestern-Kellogg School in Economic Theory, Bocconi-CCA-Cornell Workshop in Political Economy, Princeton Micro Student Lunch, Boston College PhD Conference, HEC Economics PhD Conference, Berlin Behavioral Econ PhD Conference, Transatlantic Theory Workshop, AYEW, ifo Junior Workshop in Economics of Education, Verona Early Career Workshop, CEPR Frontiers in Measurement and Survey Methods, BRIC, CEBI Behavioral Labor PhD Workshop, CREST-LMU-WZB Workshop on Cognitive Economics. The support of Premio Rotary Dott. Gabriele Corbelli is gratefully acknowledged. All errors remain my own.

In modern educational systems, standardized test scores are a major carrier of information (Wainer, 2011). Institutions use them as evidence to decide on college admission, scholarship assignment, student tracking, and teacher promotion. Researchers use them as data to measure the level and evolution of student achievement, as well as the effects of policy interventions. Because standardized test scores matter so much, it matters to understand what information they contain.[1]

The standard interpretation is that test scores measure students' knowledge on the tested subject and, more broadly, their cognitive skill. A natural question is whether that interpretation is complete, or whether scores also capture additional dimensions of human capital, such as higher-order skills (Deming and Silliman, 2024).

The question arises because many tests do not *only* ask students to solve questions. Several design choices—such as penalties for errors, time pressure, and examinee-choice formats—create an explicit or implicit opportunity cost of answering. When this cost is high enough, taking a test becomes a selection problem, too. Students must not only solve questions, but also choose which questions to attempt and which to avoid. The usual rationale for these design choices is that they encourage students to reveal what they know: by limiting uneducated guessing, costs of answering should make scores more informative (Thurstone, 1919). But what if selecting questions well requires a skill of its own?

In this paper, I argue that tests measure not only what students know, but also whether students know what they know and what they do not know—that is, their metacognition (Fleming, 2024). I develop a measurement framework that recovers metacognition from test-taking choices. The core idea is simple: if students have good metacognition, they should select questions efficiently, omitting those they are least likely to answer correctly. Applying the framework to a large field experiment on multiple-choice testing from Iriberri and Rey-Biel (2021), I show that they largely do not. Students sort questions only slightly better than random, on average, and a large minority sort adversely. Selection quality varies substantially across students, and this heterogeneity is organized by observable characteristics, especially gender. To quantify these patterns, I estimate a structural model of metacognitive ability. Using the estimates, I decompose scores into knowledge and metacognition components and derive two broad policy implications. First, interpreting scores as signals of knowledge alone entails sizable information loss; second, metacognitive differences help explain why gender gaps in test-taking widen with student ability, a widely documented puzzle (e.g. Arenas and Calsamiglia, 2025; Coffman and Klinowski, 2020; Conde-Ruiz et al., 2020; Díez-Rituerto et al., 2024; Iriberri and Rey-Biel, 2021).

[1]For instance, Anelli (2020); Ebenstein et al. (2016); Haggag et al. (2024); Hoekstra (2009); Zimmerman (2019) study how test scores impact access to selective universities and its long-run consequences for earnings and social mobility; Abdulkadiroğlu et al. (2014); Bui et al. (2014); Duflo et al. (2011); Figlio and Ozek (2024) study access to gifted programs as well as student tracking; test-score-based measures are also widely used to evaluate teacher effectiveness and inform personnel decisions (Chetty et al., 2014a,b; Macartney et al., 2018; Rockoff et al., 2012).

The analysis starts from an operational definition of metacognition. In psychology and neuroscience, metacognition is usually assessed by the relationship between objective performance and subjective confidence across tasks (Fleming and Lau, 2014). Good metacognition means that confidence is higher when the subject is likely to be correct and lower when the subject is likely to be wrong.[2] Direct confidence reports, however, are often unavailable in economic data, and even when they are, their interpretation is debated.[3] This creates a measurement problem: recovering the relationship between knowledge and confidence from choices alone.

I address this challenge proposing a framework, in the spirit of the literature on selection and Roy models (Heckman, 1979; Heckman and Vytlacil, 2005). A test is a population of matches between students and questions. Each match has two latent indices: knowledge, the probability that the student would answer correctly if forced to attempt the question, and confidence, the student's subjective assessment of that probability. The student answers when confidence exceeds a threshold determined by the opportunity cost of answering and her personal features, including risk attitudes. The analyst observes, for each match, whether the student answers the questions, and success or failure only when the question is answered. Knowledge therefore determines potential performance, confidence determines selection, and metacognition determines the statistical relationship between the two.

Within this framework, I first define a benchmark that summarizes the metacognitive implications of standard belief formation models. Many canonical assumptions—including correct beliefs, overconfidence (e.g., Moore and Healy, 2008), the hard–easy effect (e.g., Erev et al., 1994; Griffin and Tversky, 1992), and probability weighting (e.g., Gonzalez and Wu, 1999; Prelec, 1998)—share a common implication: higher objective probabilities correspond to higher subjective ones. Students may therefore be biased in their average confidence level, but they should still rank their knowledge correctly across questions. I generalize this property, allowing for noise in the knowledge-to-confidence mapping, and define the *Rank-Preserving Metacognition (RPM) benchmark*.

Rank Preserving Metacognition has a simple implication for selection. If confidence ranks knowledge correctly, selecting on confidence should equivalently select on knowledge. Students may attempt too many or too few questions relative to an otherwise identical student with correct beliefs, but the questions they attempt should still be those on which they are more likely to succeed. Hence, selection may be biased on the quantity margin, but remains efficient on quality.

However, because knowledge is latent and success or failure is observed only for attempted questions, this sorting restriction cannot be tested directly. I therefore derive a test that

[2]This notion is distinct from confidence bias, the average displacement between confidence and performance; it is roughly independent of raw intelligence (Mazancieux et al., 2020; Rouault et al., 2018), and instead related to accurate self-assessment (Rouault et al., 2019) and effective effort allocation (Ackerman and Thompson, 2017).

[3]See Bernheim et al. (2026) and Enke and Graeber (2023, 2026) for two alternative takes on this issue.

uses exogenous variation in the cost of answering as an instrument for the propensity to omit: if a higher cost induces more omission, RPM implies that the share of attempted questions answered correctly—henceforth, hit-rate—should rise.

I apply this test to the field experiment of Iriberri and Rey-Biel (2021), which varies the cost of answering within student in the *Concurso de Primavera de Matematicas*, an annual mathematics competition in Madrid. I analyze 9,807 student observations across 12 test sittings, corresponding to three years (2016–2018) and four age levels (covering students aged 10 to 17). Students take a 25-question multiple-choice test. In the test, correct answers yield five points and wrong answers zero throughout, whereas omissions yield zero in the first 13 questions but one point in the final 12, raising the opportunity cost of answering. Aside from this incentive change, the organizers designed the two blocks to be comparable in content and difficulty.

My main empirical result is that higher penalties induce a large and robust increase in omission, but only a small and fragile gain in accuracy—against the predictions of Rank-Preserving Metacognition. The result articulates into three related pieces of evidence.

First, in the overall sample, the omission-rate rises by 34.7 percentage points, from 4.9 to 39.6 percent, whereas the hit-rate rises by only 2.6 points, from 46.4 to 49.0 percent, with a confidence interval that includes zero.[4] These point estimates imply that the marginal questions screened out under penalties still have a 41.9 percent probability of being answered correctly. Attempting such a question yields 2.09 expected points, compared with one point from omission; omission therefore sacrifices 1.09 points, or about 52 percent of the expected payoff from answering. Rationalizing this pattern through preferences alone would require very strong risk aversion.

Second, the individual-level data show that the aggregate quantity–quality wedge reflects widespread behavior rather than a few extreme students. Because each student answers only 25 questions, raw treatment effects are noisy, so I estimate their joint distribution using a hierarchical Bayesian model that separates finite-test noise from underlying heterogeneity across students, questions, and sittings (McElreath, 2018; Meager, 2019). The model predicts that 37.4 percent of students omit more under penalties without improving their hit-rate. Moreover, after partial pooling treatment effects are still substantially heterogeneous on both margins, with posterior standard deviations of 22.4 percentage points for omission and 11.6 points for hit-rates, but are only weakly related, with a correlation of 0.148. Thus, students differ considerably in how strongly they respond to penalties, but stronger omission is only weakly associated with better selection.

Third, the quantity–quality wedge is systematically related to gender and ability in a way that further strains the RPM benchmark. Conditional on other student characteristics,

[4]The statistical fragility of the hit-rate effect reflects heterogeneity across test forms. Whereas omission rises significantly in every sitting—and by more than 30 points in nine of the twelve—the hit-rate effect is significantly *negative* in five of the twelve sittings, and below ten points in ten.

female students increase omissions by 5.1 percentage points more than male students, and the gap grows with mathematics achievement: it is about 3.6–3.7 points in the bottom grade quintiles and roughly 6.0–6.6 points in the upper ones. However, females do not have a significantly larger hit-rate response than male students at *any* quintile. This pattern is puzzling because RPM would predict that, fixed ability, higher omission would translate into better selection; and because whereas differences in willingness to answer may account for greater female omission *on average*, the gap measure of risk aversion in the dataset of Iriberri and Rey-Biel (2021) does not widen with ability—failing to explain the omission *gradient*.

To interpret these patterns, I specialize the nonparametric framework to a binary model that separates confidence levels from the quality of the knowledge–confidence mapping. Each student–question match is in either a high- or low-knowledge state, and confidence can recognize or misclassify that state. Positive metacognition is the probability of high confidence when knowledge is high—knowing what one knows—while negative metacognition is the probability of low confidence when knowledge is low—knowing what one does not know.[5] Their difference determines *confidence bias*, which governs the quantity margin—how often the student feels confident and therefore how much she answers—as in standard models of confidence. The centered sum of metacognitive parameters determines *metacognitive sensitivity*, which governs how quantity translates into quality—whether selective answering improves performance—providing the additional degree of freedom to explain the puzzles.

I take this decomposition to the data by reparameterizing the hierarchical Bayesian model used to recover individual treatment effects. The results show that confidence ranks knowledge in the right direction on average, but only weakly: mean metacognitive sensitivity is 0.124 on a $[-1, 1]$ scale, and its population distribution also extends below zero.[6] Average confidence bias is positive, at 0.222, indicating that students are better at recognizing what they know than what they do not know. Importantly for the reduced-form patterns, both bias and sensitivity are dispersed across students and are largely uncorrelated. Students can therefore differ greatly in how much they answer while differing, largely independently, in how well confidence sorts knowledge. This separation accounts for the dispersion of omission and hit-rate responses in the experiment, as well as their weak relationship.

The structural estimates also reveal that, while the two genders have similar metacognitive sensitivity, they differ sharply in confidence bias: the latter is 0.134 for female students and 0.268 for male students, a gap of 13.4 percentage points.[7] This difference arises because female students are 9.7 percentage points less likely to recognize high-knowledge states but 3.7 points more likely to recognize low-knowledge states. They are therefore better at

[5] Conceptually, these notions echo positive and negative introspection in epistemic logic and game theory (Aumann, 1976; Kripke, 1963).

[6] In the scale, -1,0, and 1 correspond, respectively, to adverse, random and perfect sorting.

[7] Male students also have 6.1 percentage points higher sensitivity, but the posterior distributions for the two genders overlap substantially.

knowing what they do not know, but worse at knowing what they know. This asymmetry accounts for both parts of the reduced-form gender pattern: lower bias combined with similar sensitivity generates a stronger female omission response without a corresponding accuracy gain, while, as knowledge rises, failures to recognize high-knowledge states become more frequent, explaining why the omission gap widens with ability.

Beyond their behavioral interest, these metacognitive differences matter for what scores measure. In the model, scores under penalties decompose into a knowledge component and a possibly negative metacognitive premium. Scores therefore reflect not only what students know, but also whether they recognize and act on what they know. Counterfactuals combining this decomposition with the structural estimates show that metacognition accounts for 26.7 percent of cross-student variation in expected scores. Even after allowing a flexible nonlinear mapping from scores to knowledge, 13.5 percent of latent-knowledge variation remains unrecovered, and 9.4 percent of comparable student pairs are ranked in opposite order by expected scores and latent knowledge. Interpreting penalized test scores as measures of knowledge alone therefore entails a nontrivial loss of information.

More importantly, the decomposition shows that metacognition also changes which students scores reward. The rule used in the experiment is *formula scoring*, under which omission receives the same expected payoff as random guessing.[8] Recognizing a low-knowledge state and omitting it therefore carries no direct expected-score return, whereas recognizing a high-knowledge state and answering does. This creates a complementarity between knowledge and confidence bias: recognizing and attempting high-knowledge states becomes more valuable as knowledge rises. Because male students are more likely to do so, the model predicts a widening gender score gap with ability. This is what the data show: the observed male–female gap rises from 0.165 points per question in the bottom mathematics-grade quintile to 0.324 in the top.[9] Structurally, differences in knowledge account for about 60 percent of the expected gender gap, while differences in recognizing and attempting high-knowledge states account for the remaining 40 percent; moreover, a substantial share of the gap remains within latent-ability bins rather than arising only from differences in ability composition.

I complement the analysis with three exercises that probe the mechanism from different angles and assess its external validity. First, I develop a within-regime test of metacognitive sorting that does not rely on variation in penalties, but instead uses an external proxy for question difficulty. Under RPM, attempted questions should be easier than omitted ones. In the data, an attempted question is easier with probability 0.634, while 23.1 percent of students sort no better than random; moreover, this independent measure of selection

[8]Formula scoring has a long history in psychometrics (Holzinger, 1924; Thurstone, 1919) and remains the most widely used calibration for penalties.

[9]A widening gender gap with ability has also been documented elsewhere in the standardized-testing literature (e.g. Arenas and Calsamiglia, 2025; Coffman and Klinowski, 2020; Conde-Ruiz et al., 2020; Díez-Rituerto et al., 2025; Iriberri and Rey-Biel, 2021).

quality strongly predicts the sorting gain identified from the experiment. Second, I examine an experiment with direct confidence reports from Bordalo et al. (2019), which allows me to assess the gender asymmetry without inferring confidence from choices. Male and female participants have nearly identical *Type-2 AUC*—a direct measure of metacognitive sensitivity (Fleming and Lau, 2014)—but male confidence has more false positives, while female more false negatives, reproducing the distinction between similar sensitivity and different bias in the structural estimates. Third, I assess external validity of my findings in a high-stakes setting using a matched comparison around the removal of penalties in the Chilean college-entry examination studied by Coffman and Klinowski (2020): the same gender-by-ability quantity–quality wedge appears over most of the achievement distribution. Together, these exercises show that the main patterns survive an alternative measure of sorting, appear directly in reported confidence, and extend beyond the experimental setting.

The remainder of the paper is organized as follows. I next discuss the related literature. Section 1 presents the framework, and Section 2 derives the RPM test. Section 3 presents the field-experimental evidence. Section 4 develops and estimates the structural decomposition, and Section 5 studies its implications for scores and gender gaps. Section 6 provides evidence for internal and external validity. Section 7 concludes.

Related Literature By studying the economic implications of imperfect metacognition in the context of standardized testing, this paper ties in several strands of literature.

I first contribute to work on what educational assessments measure (Jacob and Rothstein, 2016). Existing research shows that scalar scores can discard information contained in observed performance: test sections and individual items can differ in predictive content and generate different student rankings (Bettinger et al., 2013; Bruhn et al., 2025; Nielsen, 2019). I identify an upstream source of information loss: selection determines which item outcomes are observed at all. Because omissions depend on whether students can identify the questions on which they are likely to succeed, the missing outcomes are potentially non-ignorable. With costly answering, a test therefore measures not only knowledge, but also the skill with which students select where to reveal it.

This perspective also reframes the literature on gender gaps in omissions and scores. Penalties for wrong answers consistently induce women to omit more than men, with consequences for measured performance (Baldiga, 2014; Coffman and Klinowski, 2020; Iriberri and Rey-Biel, 2021).[10] Existing explanations primarily concern the propensity to answer, through risk attitudes, competitive preferences, or confidence levels. I instead

[10]Related work on gender differences in multiple-choice testing studies differential scoring and willingness to guess (Akyol et al., 2016; Conde-Ruiz et al., 2020; Díez-Rituerto et al., 2024; Funk et al., 2023; Karle et al., 2022; Pekkarinen, 2015), as well as answering incentives (Riener and Wagner, 2017). Other features of the testing environment have also been shown to shape gender differences in performance, including the stakes (Azmat et al., 2016), competitive pressure (Díez-Rituerto et al., 2025; Iriberri and Rey-Biel, 2019; Niederle and Vesterlund, 2010), time constraints (Galasso and Profeta, 2024; Krishna et al., 2025), and question context and framing (Espinosa and Gardeazabal, 2020; Ruiz-Tagle, 2024).

distinguish the *amount* of selection—how much students omit—from its *quality*—whether they omit the questions on which they are least likely to succeed. Larger omission responses therefore need not produce larger accuracy gains, and scoring rules can reward self-knowledge as well as knowledge. The distinction also provides a mechanism for gender score gaps to widen with ability: if men and women differ in which knowledge states they recognize and act on, scoring rules can amplify those differences among higher-knowledge students.

The same distinction separates my paper from most economic work on confidence (see Moore and Healy, 2008, for a review). Economics has largely studied confidence through its level or bias—whether individuals are too confident or not confident enough—and how such distortions affect decisions.[11] For selection across tasks, however, confidence levels and confidence rankings play different roles: levels affect how much individuals select, whereas allocation quality depends on whether confidence ranks task-specific success probabilities correctly. This is the notion of metacognitive sensitivity studied in psychology (Fleming and Lau, 2014). Closest in economics is Enke et al. (2023), who show that the effect of self-selection on group decision bias depends on the *task-level* relationship between confidence and performance across the population. I study this relationship *within individuals*, separate it from confidence bias, and show that it governs task selection and therefore enters measured performance itself.

Finally, the paper connects metacognition to work on higher-order skills and task allocation. Recent research shows that performance depends not only on productivity within tasks, but also on choosing which tasks, inputs, or actions to pursue (Caplin et al., 2023; Deming, 2021, 2022; Deming and Silliman, 2024). I study metacognition as one source of allocative skill: when people are uncertain about task-specific productivity, effective allocation requires identifying where they are more and less likely to succeed. Closely related is Chan et al. (2022), who show that radiologists differ not only in diagnostic thresholds, but also in how accurately they rank patients by disease risk.[12] Here, however, students rank their own question-specific knowledge states rather than external cases drawn from a common risk distribution. Because the same individual generates the latent outcome and their own decision determines whether it is observed, selective answering requires a different measurement framework and source of identification.

[11] Applications include entry (Camerer and Lovallo, 1999; Niederle and Vesterlund, 2007), trading (Barber and Odean, 2001; Odean, 1998), managerial decisions (Malmendier and Tate, 2005, 2008), job search (Spinnewijn, 2015), and political behavior (Ortoleva and Snowberg, 2015).

[12] Related work separates productive or procedural skill from allocative or decision-making skill in health care (Chandra and Staiger, 2020; Currie and MacLeod, 2017).

1 Framework

I introduce a simple framework that allows me to characterize student metacognition based on their question selection choices.

Population Structure I model a standardized test as a population of *matches* between students and questions. The collection of student-question matches forms a measurable space over which all random variables of interest are defined; I suppress match-specific indexing to simplify notation. Each student has type $S = (S^o, S^u) \in \mathcal{S} \equiv \mathcal{S}^o \times \mathcal{S}^u$, where S^o collects observed characteristics and S^u summarizes unobserved heterogeneity. Each question, analogously, has type $Q = (Q^o, Q^u) \in \mathcal{Q} \equiv \mathcal{Q}^o \times \mathcal{Q}^u$. The relevant match-level type is therefore a pair $(S, Q) \in \mathcal{S} \times \mathcal{Q}$. Throughout, I assume limited unobserved heterogeneity: the sets $\mathcal{S}^u$ and $\mathcal{Q}^u$ are finite.

Selection and Outcomes I study both selection and outcomes at the match level. For each student-question match, let $Y^* \in \{0, 1\}$ denote the *latent outcome*: $Y^* = 1$ if the student would answer the question successfully were they forced to attempt it, and $Y^* = 0$ if they would fail. The student chooses whether to answer the question or take an outside option. Let $D \in \{0, 1\}$ denote this decision, where $D = 1$ means that the student answers. The analyst observes the latent outcome only when the student answers. The *observed outcome* is therefore[13]

$$Y = DY^*. \tag{1}$$

Knowledge and Confidence There are two latent variables,

$$(K, C) \in [0, 1]^2,$$

where K is the *objective knowledge index* and C is the *subjective confidence index* for a student on a question. Their roles are pinned down by the following structural assumption.

Assumption 1 (Structure)**.** It holds

(i) $Y^* \mid (K, C, S, Q) \sim \text{Bernoulli}(K)$;

(ii) $D = \mathbf{1}\{C > \tau(S)\}$ for some $\tau(S) \in [0, 1]$.

Assumption 1 separates performance from selection. The latent knowledge index K governs success conditional on attempt. It can be interpreted as the success frequency the student would obtain if forced to answer the same question *ex novo* many times. In the

[13]Under my notation, $Y = 0$ may arise either because $D = 0$ or because $Y^* = 0$. Since D is observed, these two cases are distinguishable in the data. This avoids introducing a separate outcome value for missingness.

data, however, performance is observed only selectively, because the student first decides whether to answer. That decision is governed by the latent confidence index C, which I interpret as the student's subjective forecast of success.[14] The threshold $\tau(S)$ summarizes the student's willingness to answer and therefore captures preference heterogeneity across student-types; for instance, more risk-averse students will have a higher τ.[15]

Metacognition The main object of interest is the joint distribution of (K, C). As a consequence of Assumption 1, this distribution depends only on the match-type (s, q). The marginal distribution of K captures how much students of type s know about questions of type q: distributions concentrated on higher knowledge levels imply a higher frequency of success. By contrast, the family of conditional distributions $(C \mid K = k)_{k \in \text{supp}(K)}$ captures how confidence varies with knowledge. It therefore measures how well students of type s know whether they know or not questions of type q: their *metacognition*. Note that the framework allows both knowledge and metacognition to vary with observed and unobserved heterogeneity on both sides of the match. On the student side, this heterogeneity may be driven by cognitive and metacognitive skill. On the question side, it may instead be driven by question difficulty and framing.

Remark 1. Although the framework is static, types can absorb some dynamic effects in reduced form. For example, if question-type includes question position, and position lowers performance and/or metacognition, then the model can capture *fatigue* (e.g. Ackerman and Kanfer, 2009), potentially with heterogeneous effects across student-types depending on *perseverance* (e.g. Duckworth et al., 2007). Instead, a substantive requirement is that the student considers all questions for answer. In practice, the assumption may be violated if students do not reach the end of the test. In my setting, this issue does not seem to be at play; see Figure A1. More generally, the same issue is common to item response theory models (e.g. De Ayala, 2013), and it is typically handled by excluding from the data those matches for which there is evidence that the student did not reach the question at all (for instance, trimming final stretches of omitted answers Mislevy and Wu, 1988, 1996).[16]

[14] In this sense, knowledge in the framework is *implicit* (Polanyi, 1958), whereas confidence captures conscious access to it (see Augusto, 2010, for a review).

[15] This is an interpretation rather than a logical necessity. More agnostically, the model assumes that, conditional on student-type, question-specific variation in the answer decision operates only through the latent index C, while the cutoff $\tau(S)$ is a stable student-level parameter. In other words, the question type Q must be excluded from τ: an item-dependent threshold may generate correlation between knowledge and selection decision *independently* from how well confidence tracks knowledge, preventing identification.

[16] A fully dynamic model of test-taking could accommodate genuinely sequential effects, such as discouragement after a run of low-confidence questions. But that gain would likely come at the cost of replacing transparent identification results with less transparent sample-level ones. This trade-off already arises in a static model once preferences are non-separable; see Appendix Section A and Appendix Section B.3. I leave a dynamic model to future research.

Microfoundation Appendix Section A provides a microfoundation for Assumption 1. I model test-taking as a multitasking problem. Threshold rules arise from *monotone* preferences: conditional on the number of questions answered, a student prefers more expected successes to fewer. I also provide a Bayesian information-processing foundation for the joint distribution of knowledge and confidence. Students observe cognitive signals generated by Blackwell experiments whose informativeness varies across questions. Unlike standard models of cognitive noise (Enke, 2024; Woodford, 2020), however, students do not perfectly know the informativeness of their own cognitive process. Instead, they receive imperfect *meta-information* about it. Objective knowledge is determined by the informativeness of the realized cognitive experiment, whereas subjective confidence is determined by the posterior-predictive experiment implied by this meta-information. Metacognitive errors therefore arise from a second layer of uncertainty, and, rather than attenuating responses *within* task (cf., Enke and Graeber, 2023; Frydman and Jin, 2022; Gabaix, 2019; Khaw et al., 2021), they distort how students allocate decisions *across* tasks.[17]

2 A Test for Rank-Preserving Metacognition

Having introduced the framework, this Section asks what choice data can reveal about the quality of students' metacognition. I begin by defining *Rank-Preserving Metacognition*, a benchmark that captures the implications of standard models of belief formation while allowing noise in the mapping from knowledge to confidence. I then derive a behavioral test of this benchmark that leverages instrumental variation in the opportunity cost of answering. Section 6.1 develops a complementary test based on the same sorting logic, but exploiting variation in item difficulty while keeping costs fixed. Appendix Section B provides additional results on testing metacognition.

2.1 The Rank-Preserving Metacognition Benchmark

To define a benchmark for metacognitive quality, I examine what standard models of belief formation imply about the relationship between objective knowledge and subjective confidence. Under correct beliefs, the two coincide: $C = K$; metacognition is then perfect. A large literature on probabilistic bias, however, shows that this assumption fails empirically, providing extensive evidence of distortions in confidence *levels*. In particular, the vast literature on underconfidence and overconfidence establishes that subjective probabilities may be shifted downward or upward relative to objective ones (Moore and Healy, 2008). The two distortions may also coexist, as in the inverse-S-shaped patterns documented in the literature on probability-weighting (Gonzalez and Wu, 1999; Prelec, 1998) and in the

[17]The microfoundation is also related to a literature that studies Bayesian updating when agents are uncertain about external signal strength; (e.g. Augenblick et al., 2025, 2024); see also Aina and Schneider (2025) for recent experimental evidence in this setting.

hard–easy effect (Erev et al., 1994; Griffin and Tversky, 1992). In all these cases, however, higher subjective probabilities still correspond to higher objective ones: while these patterns distort confidence *levels*, they still preserve *rankings*.

This observation motivates the benchmark I adopt. Taken literally, the models above fit the present framework if

$$C = \psi(K) \quad \text{for some weakly increasing map} \quad \psi : [0,1] \to [0,1]. \tag{2}$$

This requirement may be too strong for experimental purposes because it makes confidence a deterministic function of knowledge, which is at odds with the empirical evidence (e.g. Enke and Graeber, 2023; Fleming, 2024). I therefore use a stochastic relaxation of Equation (2).

Definition 1. Fix a student-type s, and let $\overline{F}_s(c \mid k) := \mathbb{P}(C > c \mid S = s, K = k)$ denote the conditional survival function of confidence. I say that type s has *Rank-Preserving Metacognition (RPM)* if, for every $c_2 > c_1$ and $k_2 > k_1$,

$$\overline{F}_s(c_2 \mid k_2)\overline{F}_s(c_1 \mid k_1) \geq \overline{F}_s(c_2 \mid k_1)\overline{F}_s(c_1 \mid k_2). \tag{3}$$

I say that RPM holds *strictly* if the inequality is strict whenever all four terms are positive.

Equivalently, Definition 1 requires that, within student-type s, the stochastic map $K \mapsto C$ is (strictly) increasing in the *hazard-rate order* (Shaked and Shanthikumar, 2007).[18] The interpretation is simple. Fix two confidence thresholds $c_2 > c_1$. Then Equation (3) says that, among matches that clear the weaker threshold c_1, those with higher knowledge are more likely to also clear the stricter threshold c_2. In this sense, the deterministic ranking property in Equation (2) becomes a stochastic one: higher-knowledge matches are *more likely* to survive a tightening of the confidence threshold. The next remark makes the connection explicit by showing how RPM arises naturally from benchmark models once one adds a well-behaved noise term.

Remark 2. Start from any model as in Equation (2). Fix a strictly increasing link function $G : \mathbb{R} \to [0,1]$, and define $\tilde{\psi} := G^{-1} \circ \psi$. For each $\sigma > 0$, consider

$$C^\sigma = G\Big(\tilde{\psi}(K) + \sigma\varepsilon\Big), \tag{4}$$

where ε is a noise term independent of K. At $\sigma = 0$, this construction reproduces the deterministic benchmark because $C^0 = G(\tilde{\psi}(K)) = \psi(K)$. For every $\sigma > 0$, it instead generates a stochastic relationship between confidence and knowledge. Writing $\overline{F}_\varepsilon$ for the

[18] Equation (3) is a restriction on the mixture distribution of confidence conditional on knowledge across question-types. It neither implies nor is implied by the analogous condition for $\mathbb{P}(C > c \mid S = s, Q = q, K = k)$, because the hazard-rate order does not commute with mixture.

survival function of ε, we have

$$\mathbb{P}(C^\sigma > c \mid K = k) = \overline{F}_\varepsilon\left(\frac{G^{-1}(c) - \tilde{\psi}(k)}{\sigma}\right).$$

Hence, if ε has increasing hazard rate, then Equation (3) holds. This condition is satisfied by the most common additive-noise specifications, such as Gaussian and logistic noise, and more generally by the log-concave family (e.g. Bagnoli and Bergstrom, 2005).

2.2 A Sorting Lemma

Rank Preserving Metacognition has a simple behavioral implication: when selection becomes stricter, the retained questions should be those on which the student is more likely to succeed. I begin by formalizing this sorting logic. For intuition, consider the deterministic case. Combining Equation (2) with the threshold decision rule in Assumption 1.ii implies that the student's decision can be written as a threshold rule in knowledge levels,[19]

$$D = \mathbf{1}\{K > \tilde{\tau}(S)\}, \quad \text{where} \quad \tilde{\tau} = \psi^{-1} \circ \tau.$$

Under the benchmark, students may therefore be *quantity-biased*: they may attempt too many or too few questions relative to their confidence threshold. But they remain *quality-efficient*: they attempt questions they are objectively more likely to answer correctly than those they omit. The following *sorting lemma* formalizes this intuition and provides the key tool for the tests below. Let $A \in \{0, 1\}$ be a nondegenerate binary selection variable, and let $T \in \mathbb{R}$ be a scalar score.

Lemma 1 (Sorting)**.** *If $\mathbb{P}(A = 1 \mid T = t)$ is weakly increasing on the support of T, the distribution of T among the selected first-order stochastically dominates its distribution among the non-selected,*

$$(T \mid A = 1) \succeq_{FOSD} (T \mid A = 0).$$

Moreover, if $\mathbb{P}(A = 1 \mid T = t)$ is strictly increasing and $\mathbb{V}(T) > 0$, strict dominance holds.[20]

The most natural application of Lemma 1 sets the selector equal to the student's answer decision, $A = D$, and the score equal to latent knowledge, $T = K$. Under RPM, the distribution of knowledge among attempted questions should then dominate that among omitted questions. This implication captures the central sorting intuition, but it is not directly testable: each question is either attempted or omitted, and K is latent, so the analyst cannot compare the two knowledge distributions directly. To obtain observable restrictions, one must therefore replace either the latent score—as in the within-regime test developed in Section 6.1—or the selector, as I show next.

[19] Notation: $\psi^{-1}(t) := \inf\{k \mid \psi(k) > t\}$ denotes the left quasi-inverse of ψ.

[20] Notation: here and in the following $\mathbb{V}(\cdot)$ denotes the variance operator.

2.3 Comparison Between Regimes

The test uses variation in the opportunity cost of answering to impute counterfactual performance. Suppose each match is assigned to a *regime* $Z \in \{0, 1\}$ that may shift the student's decision threshold. Specifically, extend Assumption 1.ii to

$$D(z) = \mathbf{1}\{C > \tau(S, z)\} \qquad \text{for some } \tau(S, z) \in [0, 1].$$

By convention, $Z = 1$ denotes the higher-cost regime. This may capture a direct change in incentives, such as a higher relative penalty for errors, as in my application, but also other changes that raise the effective cost of answering, such as tighter time pressure.

I impose two assumptions that relate the regime to the other variables in the framework. The first is a conventional exclusion restriction.

Assumption 2 (Regime Exclusion)**.** $(K, C) \perp Z \mid (S, Q)$.

Thus, conditional on student and question-type, the regime may shift the threshold but cannot change the joint distribution of knowledge and confidence. In this sense, Assumption 2 requires the metacognitive profile of a given student-type on a given question-type to be invariant to incentives. This separation arises naturally in the microfoundation of Appendix Section A.

The second assumption ensures that, within a student-type, the composition of question-types is comparable across regimes. It holds, for example, when question content is balanced across regimes, as in my application.

Assumption 3 (Balance)**.** $Q \perp Z \mid S$.

Assumption 2 and Assumption 3 allow me to formulate a between-regime test based on two observable statistics. For each student-type s, I compare across regimes the share of omitted questions, which I call the *omission-rate*, and the share of correct answers among attempted questions, which I call the *hit-rate*. For convenience, I restrict attention to the subpopulation of student-types that attempt at least some questions in both regimes, so that hit-rates are well defined.[21] For each such type, define

$$O_{s,z} := \mathbb{P}(D = 0 \mid S = s, Z = z),$$
$$H_{s,z} := \mathbb{P}(Y = 1 \mid D = 1, S = s, Z = z),$$

and let $\Delta(\cdot)_s$ denote the difference between $z = 1$ and $z = 0$.

[21] In my application, virtually all students pass this restriction. More generally, student-types that never attempt any question in a regime carry no information about their profile in that regime.

Proposition 1 (Between-regime test)**.** *Suppose Assumption 1–Assumption 3 hold. If students of type s have Rank-Preserving Metacognition, then*

$$\Delta O_s > 0 \Longrightarrow \Delta H_s \geq 0,$$

with strict inequality, if RPM holds strictly and $\mathbb{V}(K \mid D(0) = 1, S = s) > 0$.

In words, if a student-type omits more questions when moving from a low-cost to a high-cost regime, then the share of correct answers among attempted questions should not fall. Under strict RPM, it should actually rise strictly, provided there is nondegenerate variation in latent knowledge among the questions the student would attempt in the low-cost regime.

Regime variation therefore offers one way around the observability problem described in Section 2.2. Under Assumption 2 and Assumption 3, moving from the low-cost to the high-cost regime changes the selection threshold, but not the distribution of (K, C) within student-type s. When the high-cost regime induces more omissions, it tightens selection within the set of matches that would be attempted in the low-cost regime. Among these matches, some remain attempted and others are newly omitted. Define the selector A to equal one when a low-cost attempted match remains attempted in the high-cost regime. RPM implies that $\mathbb{P}(A = 1 \mid K = k)$ is increasing in k. The sorting lemma then implies that the retained matches are positively selected relative to the newly omitted matches. Hence average latent knowledge among attempted matches cannot fall. Since the hit-rate identifies average latent knowledge among attempted matches, the hit-rate cannot fall either.

Operationally, the test requires within-student variation in answering costs, as in the experiment, and it can be interpreted as a statement about individual treatment effects on omission and hit-rates. In practice, individual treatment effects may be noisy, unless one observes long tests. For this reason, it is useful to record an aggregate implication of Proposition 1, which allows me to leverage the size of the student sample. For that purpose, I impose one more assumption.

Assumption 4 (Uniformity)**.** Either $\tau(s, 1) \geq \tau(s, 0)$ for every student-type s, or $\tau(s, 1) \leq \tau(s, 0)$ for every student-type s.

This assumption rules out cases in which the regime shift moves selection in opposite directions across student-types. Instead, the shift must move all students either toward more omission or toward less omission. Let $w_s := \mathbb{P}(S = s)$ denote the population share of student-type s.

Corollary 1 (Pooled between-regime test)**.** *Suppose Assumption 1–Assumption 4 hold and*

every student-type satisfies Rank-Preserving Metacognition. Then

$$\sum_s w_s \Delta O_s > 0 \Longrightarrow \sum_s w_s \Delta H_s \geq 0,$$

with strict inequality, if RPM holds strictly and $\mathbb{V}(K \mid D(0) = 1, S = s) > 0$ *for every* s.

I conclude this Section discussing some additional features and results.

Connection with LATE The between-regime test can also be recast in terms of a statistic whose cardinal value is directly interpretable and which clarifies a connection with the LATE framework (Imbens and Angrist, 1994; Kline and Walters, 2019; Vytlacil, 2002). Within each student-type s, define *always-taker matches* as those attempted under both regimes, $\mathcal{A}_s := \{D(0) = 1,\ D(1) = 1,\ S = s\}$, and *complier matches* as those attempted in the low-cost regime but omitted in the high-cost regime, $\mathcal{C}_s := \{D(0) = 1,\ D(1) = 0,\ S = s\}$. Then define the *sorting gain*

$$\Gamma_s := \mathbb{E}[Y^* \mid \mathcal{A}_s] - \mathbb{E}[Y^* \mid \mathcal{C}_s], \tag{5}$$

which measures how much average latent knowledge improves when the student moves from the marginal questions screened out by the tighter regime to the questions that remain selected. In this sense, Γ_s quantifies the strength of metacognitive sorting induced by the increase in answering costs. A simple algebraic manipulation yields that, whenever $1 > O_{s,1} > O_{s,0}$,

$$\Gamma_s = \frac{1 - O_{s,0}}{O_{s,1} - O_{s,0}}(H_{s,1} - H_{s,0}).$$

Thus, Proposition 1 can be restated as a restriction on Γ_s: under RPM, the tighter regime must generate nonnegative sorting gains.

Identification Result Appendix Section B shows an identification result that asks how much of the stochastic confidence map can be recovered, when the analyst observes more than two selection regimes. Replacing the binary regime comparison with a continuum of answer thresholds allows one to trace a *hit-omit curve*, which records how hit-rates and omission-rates co-move as selection becomes stricter. Proposition A11 shows that increasingly strong geometric properties of this curve are equivalent to increasingly strong monotonicity properties of the stochastic map $K \mapsto C$. The test in Proposition 1 follows as a binary version of this result, obtained by comparing two points on the curve. Thus, provided sufficiently rich threshold variation, a behavioral restriction akin to the one derived above can actually *identify* the underlying monotonicity of confidence in knowledge.

Quantitative Bounds The Appendix also asks what RPM implies for the *magnitude* of sorting patterns. To do so, I introduce a quantitative strengthening of RPM, called

κ-metacognition that requires the cross-second derivative of the conditional log-survival function of confidence to be bounded below by κ, nesting RPM for $\kappa = 0$. Using exponential-tilting methods (Ellis, 2012), Proposition A9 derives lower bounds on the between-regime sorting statistics, expressed either as hit-rate changes or sorting gains. These bounds increase both in κ, which measures how tightly confidence tracks knowledge, and in a dispersion term capturing how much a student's knowledge varies across questions. The result gives the magnitude of the observed sorting gain a quantitative interpretation as an index of metacognitive quality, holding fixed the relevant dispersion in latent knowledge.

Sample-Level Tests Finally, the Appendix relaxes the fixed-threshold decision rule in Assumption 1.ii. As the microfoundation in Appendix Section A makes clear, a fixed threshold requires a separability assumption in addition to monotone preferences over successful answers. More generally, students may choose portfolios that are monotone in confidence while the threshold itself depends on the entire realized confidence vector. In this case, the sorting logic can be leveraged to deliver sample-level implications. Using concentration inequalities (Boucheron et al., 2013), Proposition A12 shows that, for long tests and large student samples, omission without positive selection becomes asymptotically unlikely under *deterministic* Rank-Preserving Metacognition.

3 Evidence from a Large Field Experiment

I bring the Rank-Preserving Metacognition benchmark to the data using a large-scale field experiment in which students face within-individual variation in penalties for errors. The empirical evidence does not support the benchmark predictions, at least in their strict form. Students omit many more questions when answering becomes more costly, but these additional omissions do not generate a statistically significant gain in accuracy, on average. For a sizable minority of students, moreover, they generate an accuracy loss, indicating adverse selection. At the same time, selection patterns vary substantially across students, and this heterogeneity is organized by observables in ways that further strain the RPM benchmark. Female students omit more questions than male students of comparable ability and risk attitude, and this omission gap widens with ability; yet the additional omission does not translate into higher sorting gains.

3.1 The Experiment

Iriberri and Rey-Biel (2021) study test-taking behavior in a large-scale field experiment conducted in collaboration with the organizers of the *Concurso de Primavera de Matemáticas*, a regional mathematics competition held annually in Madrid.

The key design feature is within-student variation in the scoring rule. Each student

receives a 25-question multiple-choice test with five answer options per question, divided into two blocks. In Questions 1–13, a correct answer earns 5 points, while a wrong answer or an omission earns 0. In Questions 14–25, correct and wrong answers are scored as before, but an omission earns 1 point. The second block therefore raises the payoff to omission, or equivalently the opportunity cost of answering. Aside from this incentive change, the organizers designed the two blocks to be comparable, especially in content and difficulty.

I observe the 2016–2018 editions of the competition. Each year contains four age levels, spanning ages 10–17. Since each year-by-level combination receives a different test-form, this yields 12 distinct *experimental sittings* which nest 11,095 students and 300 questions in total. I retain students with at least one attempted question in each block, so that both hit-rates are defined. The resulting sample contains 9,807 student observations.[22] Some students participate in more than one year, yielding 8,474 different "physical participants". The main analysis sample hence contains 245,175 student-item matches, among which 192,444 correspond to attempted questions.

The dataset also records several student characteristics including gender, school mathematics grade, school type, and previous participation in the competition. In 2017 and 2018, a post-test survey additionally elicits a forecast of own performance and willingness to answer under the omission-reward rule, which I use respectively as a proxy for confidence levels and risk attitudes. Table A1 reports descriptive statistics for the student covariates.

On the question side, I augment the dataset annotating features of the items, based on their stem text, answer options and their proposed solution text.[23] A first set of features is measured directly, including length, the density of mathematical-expressions and images, and the label of the correct answer. Other characteristics require classification. Following the emerging literature that uses large language models to extract semantic features from text data (Ash et al., 2025), I run a structured annotation protocol with ChatGPT-5.6 Sol on the item bank. Three independent LLM runs annotate mathematical domain, whether the problem is abstract or contextualized, whether it requires extra-curricular knowledge, the type of reasoning involved in the proposed solution, as well as properties of the answer menu. I then assign final labels by two-of-three consensus whenever possible and resolve remaining disagreements.[24] Appendix Section E reports the details, including the protocols' ontology and prompt, as well as inter-LLM-run agreement statistics. Table A2 reports descriptive statistics on item covariates.

[22]Moreover, students who skip an entire block arguably exhibit inattention or, more generally, behavior that is qualitatively different from that captured by my framework.

[23]This text-data is available in PDF format at https://www.concursoprimavera.es/.

[24]Only 8/300 items require at least one such adjudication.

3.2 Average Treatment Effects

I begin by bringing to the data the comparison of the average omission- and hit- rates between regimes (Corollary 1).[25]

Discussion of the Identification Assumption Identification requires the two blocks to generate comparable distributions of knowledge and confidence, conditional on student type, while differing in the incentive to omit (Assumptions 2 and 3). This requirement is subject to two natural concerns. The first is systematic differences in question composition. The design explicitly addresses this issue, since the organizers maintained the two blocks comparable in content and difficulty. To support this, Table A2 compares the 24 annotated item characteristics across blocks using a permutation test that reshuffles questions' regime assignment within sitting. For 22 of the 24 observed item covariates, the difference is not statistically significant. The only exceptions are symbolic answer options and questions containing extra-curricular material, both slightly more common in the second block.[26] To further counter the concern, I therefore include robustness checks that control for the full set of question characteristics.

The second concern is that later questions may be affected by fatigue or disengagement rather than the scoring rule. Since organizational constraints prevented randomizing the position of the penalty block, the design does not directly address this issue.[27] Still, several features of the experimental implementation mitigate worries about time effects. Students receive the full test at once, may answer questions in any order, and are encouraged to read the test first and begin with questions they find easiest. Presentation order therefore need not coincide with completion order.[28]

Accordingly, Figure A1 shows a sharp increase in omission at the scoring-rule boundary with little comparable drift within blocks; and among students who answer every question—and hence whose performance is not confounded by selection—Figure A2 likewise shows no systematic deterioration in hit-rates with question position. Finally, in the 2013–2015 editions, when the scoring rule did not change across blocks, Iriberri and Rey-Biel (2021, Table A2) finds no significant block differences in omissions, hit-rates, or scores. Taken together, these checks support the interpretation that the scoring rule, rather than potential confounding factors, predominantly drives the change in selection behavior.

[25] Questions 1–13 define the lower-cost regime, $Z = 0$, and Questions 14–25 as the higher-cost regime, $Z = 1$.

[26] Symbolic answer options increase from 2.1 to 7.5 percent of questions, while beyond-curriculum questions increase from 3.5 to 9.4 percent.

[27] Position-specific effects that evolve comparably within blocks can be instead absorbed into question type, as discussed in Remark 1.

[28] Moreover, students were unaware of being studied, limiting experimenter demand and anticipation effects that can threaten within-individual designs (List, 2025).

Estimation I estimate regime effects at the student-question level. Let B_{ij} denote the outcome of student i on question j, and let Z_{ij} indicate that the question belongs to the higher-cost block. I estimate

$$B_{ij} = \alpha_i + \beta Z_{ij} + \varepsilon_{ij}, \tag{6}$$

by weighted least squares, where α_i are student fixed effects. For omission, $B_{ij} = 1 - D_{ij}$, and each question in block z receives weight $1/J_z$, with $J_0 = 13$ and $J_1 = 12$. For the hit-rate, I restrict the sample to attempted questions, set $B_{ij} = Y_{ij}$, and weight each observation by $1/A_{iz}$, where A_{iz} is the number of questions student i attempts in block z. Hence, in both regressions, each student receives total weight one within each block. As a consequence, $\widehat{\beta}$ is the average within-student change in either omission- or hit-rate, which I interpret as the treatment effect of penalties on these quantities.

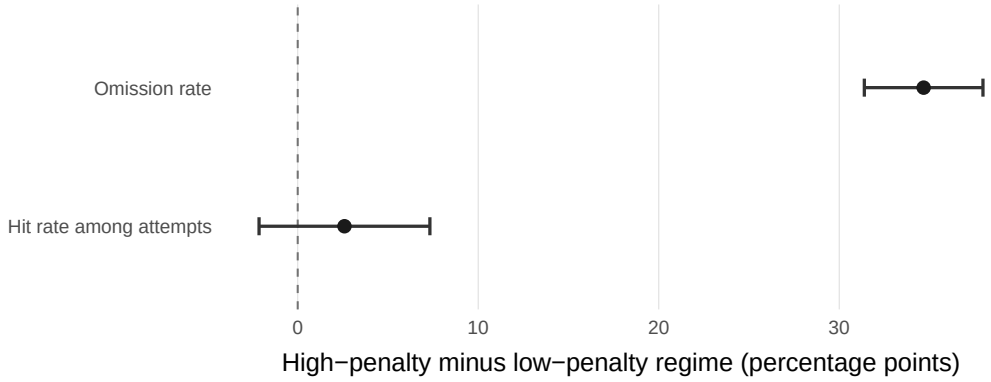
Note that the same point estimates for the β's could be obtained by first collapsing the data to student-by-block rates and then comparing averages between sittings. I retain the student-question representation for two reasons. First, it permits inference along both dimensions of the data. Throughout the main analysis, standard errors are clustered two ways, by physical participant and by item. The participant cluster allows arbitrary dependence across questions answered by the same person, including repeated competition appearances; the item cluster allows arbitrary dependence across students facing the same question. This second dimension is important because the decisions and performance of all students facing a given item may be affected by its features. Second, the question-level representation lets me add the observed item characteristics directly to Equation (6), which I use below to assess sensitivity to differences in item composition.

Results Figure 1 reports the main estimates. When answering becomes relatively more costly, students become much more selective; the omission-rate rises by 34.7 percentage points, from 4.9 to 39.6 percent, with a 95 percent confidence interval of $[31.4, 38.0]$ percentage points. By contrast, the hit-rate among attempted questions rises by only 2.6 percentage points, from about 46.4 to 49.0 percent, with a 95 percent confidence interval of $[-2.1, 7.3]$ points. Thus the incentive change induces students to screen out more than one third of all questions, while producing only a small and statistically uncertain improvement in the success rate on those they retain.

Cross-sitting evidence sharpens the contrast between the quantity and quality margin of selection, and helps to explain why the average hit-rate response is imprecise. Figure A3 reports the same within-student estimates separately for the 12 sittings.[29] The omission response is positive and economically large in *every* sitting: it never falls below 14 percentage points and exceeds 30 points in 9 of the 12 sittings. The hit-rate response, by contrast, is much less stable. Its point estimate ranges from about -12 to $+20$ percentage points,

[29] In this figure inference is at participant level, because confidence intervals are intended to display realized heterogeneity across sittings rather than inference over new test forms.

Figure 1: Pooled Between-Regime Omission and Hit-Rate Effects



Notes: Points report weighted student-fixed-effects estimates in percentage points. Horizontal intervals are 95 percent confidence intervals clustered two ways by physical participant and form-specific item.

is negative in 5 sittings, and remains below 10 points in 10. Thus, the pooled hit-rate estimate does not reflect uniformly positive but noisy responses across tests. Penalties consistently increase omissions, but the hit-rate response varies substantially across tests and is often close to zero or negative.

Finally, a back-of-the-envelope calculation for the sorting-gain statistic makes the economic magnitude of this pattern transparent, and further underscores the difficulty of explaining the pattern through a standard model. Using the pooled estimates,

$$\hat{\Gamma} = \frac{1 - \hat{O}_0}{\widehat{\Delta O}} \widehat{\Delta H} \simeq 0.071,$$

meaning that the questions retained under the penalty regime are therefore only 7.1 percentage points more likely to be answered correctly than the marginal questions screened out by the stronger omission incentive. Since the hit-rate among retained attempts is about 0.490, the implied success probability on marginally omitted questions is still about $0.490 - 0.071 \simeq 0.419$. Answering such a question would therefore yield, on average, $5 \times 0.419 \simeq 2.09$ points, compared with one point for sure from omission. Students thus give up about $1.09/2.09 \simeq 52\%$ of the expected value of answering the marginal omitted question, and they do so on more than one third of the test. This suggests that a preference-based explanation would require excessive risk aversion.

Robustness Checks I conduct two sets of robustness checks on the main finding. First, I account for observable differences in item composition. Starting from the baseline specification, I progressively add controls for directly measurable item features and then for the LLM-annotated semantic characteristics. The omission effect remains between 34.5 and 35.0 percentage points throughout. The hit-rate effect rises only modestly, from 2.59 to 3.56 percentage points in the richest specification, and remains statistically indistinguishable from zero; see Figure A4. Observable differences in question composition therefore do not

explain the contrast between the quantity and quality responses.

Second, I assess the sensitivity of inference following MacKinnon et al. (2023). To reduce sensitivity to finite-sample distortions from heterogeneous or high-leverage clusters, I replace the conventional two-way CV1 estimator with the two-way CV3 maximum-standard-error rule. The standard error changes only from 1.67 to 1.74 percentage points for omission and from 2.40 to 2.61 for the hit-rate. Consistent with this result, leave-one-item-out and leverage diagnostics show that neither estimate is driven by a small number of influential questions; see Figure A5 and Table A3. I then treat the entire sitting as the clustering unit, allowing dependence across different questions on the same realized test form. Because this leaves only 12 clusters, I report both sitting-level CV3 inference and a WCR-S wild-cluster bootstrap with Webb weights. The former yields 95 percent confidence intervals of $[28.42, 40.93]$ percentage points for omission and $[-4.46, 9.65]$ for the hit-rate; the latter gives $p < 0.001$ and $p = 0.445$, respectively. Hence, neither finite-cluster corrections, influential items, nor form-level dependence changes the conclusion.

Overall, average treatment effects are difficult to reconcile with the Rank-Preserving Metacognition benchmark. Stronger incentives generate a large increase in omission, but only a small and unstable improvement in the quality of retained answers. Naturally, this average pattern can arise in very different ways: for instance, sorting may be weak for most students, or positive for the majority of students but offset by strongly adverse sorting from a minority of outliers. Distinguishing these possibilities requires moving from average to individual treatment effects.

3.3 Individual Treatment Effects

The object of interest is the joint distribution between the student-level changes in omission- and hit-rates, respectively ΔO_i and ΔH_i. Under Proposition 1, a student who becomes more selective should not become less accurate on the questions she retains. The benchmark therefore predicts no mass in the southeast quadrant of the space of omission- and hit-rate differentials.

$$\mathbb{P}\left(\Delta O_i > 0,\ \Delta H_i \leq 0\right) = 0. \tag{7}$$

The raw data already point in the opposite direction. In the working sample, 87.75 percent of students omit at least one additional question under the penalty regime. Among these students, 45.41 percent have a non-positive hit-rate effect. These raw individual changes are noisy, however, because each student faces only 25 questions. Their empirical dispersion therefore conflates genuine heterogeneity in regime responses with finite-test noise. Moreover, students are observed in different test sittings, whose underlying response distributions may themselves differ.

Bayesian Model To address these challenges, I adopt a Bayesian approach and estimate a two-part generalized linear mixed model for omission and correctness conditional on attempt, crossed with item effects (Gelman et al., 1995; McElreath, 2018).[30] The model is easiest to understand by first fixing a sitting. Within a sitting, each student-question outcome reflects four sources of heterogeneity: observed and unobserved student characteristics, and observed and unobserved question characteristics. Let i index students, j questions, and $b(i)$ the sitting containing student i. For convenience, center the regime indicator as $r_j = Z_j - \frac{1}{2} \in \{-1/2, 1/2\}$. The observation model is

$$D_{ij} \sim \text{Bernoulli}(1 - o_{ij}), \qquad Y_{ij} \mid D_{ij} = 1 \sim \text{Bernoulli}(h_{ij}),$$

where, for $q \in \{O, H\}$, $p_{ij}^O := o_{ij}$ and $p_{ij}^H := h_{ij}$, and

$$\text{logit}(p_{ij}^q) = \underbrace{\bar{a}_{b(i)}^q + X_i'\beta_q + s_i^q}_{\text{student baseline}} + r_j \underbrace{\left(\bar{t}_{b(i)}^q + X_i'\delta_q + v_i^q\right)}_{\text{student regime response}} + \underbrace{W_j'\gamma_q + u_j^q}_{\text{question baseline}}. \tag{8}$$

Here, o_{ij} is the probability that student i omits question j, while h_{ij} is the probability that they answer it correctly conditional on attempt. The vectors X_i and W_j contain the student and question covariates introduced above. For $q \in \{O, H\}$, $X_i'\beta_q$ captures observable student heterogeneity in baseline behavior, while $X_i'\delta_q$ captures observable heterogeneity in the regime response. The residual terms s_i^q and v_i^q allow both components to vary further across students. Questions enter analogously through $W_j'\gamma_q$ and residual item effects

$$u_j^q = u_j^{q,\star} - \frac{1}{J_{bz}} \sum_{\ell \in (b,z)} u_\ell^{q,\star}, \qquad j \in (b, z),$$

obtained by centering the raw effects $u_j^{q,\star}$ within each sitting-by-regime cell. Hence, $\sum_{j \in (b,z)} u_j^q = 0$, preventing residual item effects from absorbing the regime contrast.[31]

Student and item effects are therefore crossed: each outcome depends jointly on the student and on the particular question she faces. I model their residual effects jointly across the two outcome equations,

$$(s_i^O, v_i^O, s_i^H, v_i^H)' \sim \text{MVNormal}(0, \Sigma_S), \qquad (u_j^{O,\star}, u_j^{H,\star})' \sim \text{MVNormal}(0, \Sigma_Q).$$

The first distribution describes the joint population of student varying intercepts and varying regime slopes. Specifically, Σ_S governs both the dispersion and covariance of these

[30] A valuable alternative would be a plug-in Empirical Bayes procedure (e.g. Kline and Walters, 2021; Walters, 2024). I choose instead a fully Bayesian model since it allows obtaining joint posterior inference for the average effects and posterior predictions for a new sitting.

[31] Note that, whereas I allow both baseline behavior and the response to the scoring regime to vary across students, I allow only the first channel across questions. This choice reflects the asymmetry of the experimental design: each student is observed under both scoring regimes; by contrast each question appears either under the high-cost or the low-cost one.

student effects, while partial pooling prevents noisy student-specific estimates from being interpreted mechanically as genuine heterogeneity. The second distribution performs the analogous role for residual question heterogeneity.

To estimate student heterogeneity correctly, however, the model must also allow the center of the student distribution to differ across sittings, since each sitting receives a different test. Collect the four sitting-specific means as

$$m_b \equiv (\bar{a}_b^O, \bar{t}_b^O, \bar{a}_b^H, \bar{t}_b^H)'.$$

Following the hierarchical approach used to aggregate heterogeneous treatment effects across experiments (e.g., Meager, 2019, 2022), I assume

$$m_b = MG_b + \text{diag}(\tau)z_b, \qquad z_b \sim N_4(0, I).$$

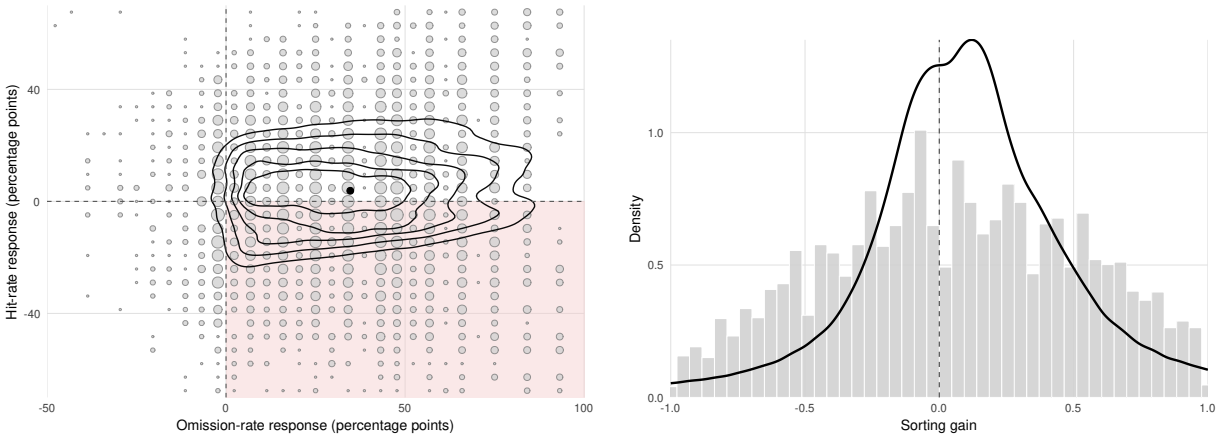
The covariates G_b capture systematic differences by year and level, while τ captures residual heterogeneity across otherwise comparable sittings. This upper layer allows the center of the student-effect distribution to shift across tests, preventing between-sitting differences from being attributed to student-level heterogeneity.[32]

All the model estimates below are reported on the probability (rather than the logit) scale, for interpretability. Appendix Section D.1 provides details on the transformation as well as complete prior specification, implementation, and diagnostics.

Individual Heterogeneity Figure 2 separates underlying heterogeneity from finite-test noise in individual treatment effects, addressing the measurement problem above. Panel A compares the raw joint distribution of student omission- and hit-rate changes (grey circles) with the distribution implied by the hierarchical model (black contours). Partial pooling removes much of the extreme raw dispersion, but substantial heterogeneity remains. The posterior standard deviation is 22.37 percentage points for the omission response, with a 95 percent credible interval of $[21.98, 22.76]$, and 11.56 points for the hit-rate response, with interval $[10.27, 12.83]$. The two responses are only weakly related: their posterior correlation is 0.148, with interval $[0.069, 0.231]$. Table A4 reports the complete posterior summaries.

For the RPM benchmark, the key fact is that substantial mass remains in the southeast quadrant ruled out by Equation (7). The posterior mean for this quantity is 37.4 percent, with a 95 percent credible interval of $[32.4, 42.6]$ percent. Thus, even after accounting for finite-test noise, the model predicts that more than one third of students omit more without

[32] I use independent residuals across the four components of m_b, rather than estimating an unrestricted 4×4 covariance matrix. With only 12 sittings, the six additional correlation parameters would be weakly identified. Dependence between omission and hit-rate responses at the student and item level is instead left unrestricted through Σ_S and Σ_Q.



(a) Individual regime effects

(b) Sorting gains

Figure 2: Individual Regime Effects and Sorting Gains

Notes: In Panel A, grey circles report raw student-level changes, grouped into a 33×29 grid; circle size increases with the number of students in each cell. Black contours enclose 25, 45, 65, 82, and 93 percent of the posterior-mixture distribution of individual effects, and the black point reports the posterior mean average effect. The shaded region corresponds to $\Delta O_i > 0$ and $\Delta H_i \leq 0$. In Panel B, the grey histogram reports raw sorting gains and the black curve their posterior-mixture density, in both cases for $\Delta O_i > 0.01$ and normalized over the displayed interval $[-1, 1]$. See Appendix Section D.1 for construction details.

improving their hit-rate. Panel B quantifies the same pattern in cardinal terms, using the sorting gain statistics Γ_i defined in Equation (5). The grey histogram reports the statistic calculated from raw student rates, while the black density reports its posterior-average distribution.[33] The posterior median sorting gain is 0.101, with a 95 percent credible interval of $[0.055, 0.148]$: the typical student therefore sorts in the right direction, but only weakly. At the same time, 38.4 percent of defined sorting gains are nonpositive, with interval $[33.3, 43.8]$ percent.

Parent Distribution As a byproduct of the analysis, the upper level of the model's hierarchy provides a Bayesian estimate for the average treatment effects, which supports the same conclusion as the reduced-form regression analysis. The parent mean response is 34.90 percentage points for omission, with a 95 percent credible interval of $[31.62, 38.00]$, but only 3.49 points for the hit-rate, with interval $[-3.40, 10.24]$. Substantial heterogeneity remains across sittings, as captured by the parent standard deviations (10.61 points $[8.32, 13.57]$ for omit-rates and 13.86 points $[8.85, 21.21]$ for hit-rates). Moreover, even if partial pooling reduces the dispersion of the sitting-specific means, it does not eliminate the asymmetry between the two margins (Figure A6). Table A5 reports the full posterior summaries.

Overall, the results suggest that, even after finite-test noise is removed, sorting is weak for the typical student, and adverse sorting is widespread. At the same time, the substantial dispersion in individual effects shows that the ability to select questions differs meaningfully

[33] Because the statistic becomes unstable as the omission response approaches zero, I define it for $\Delta O_i > 0.01$, a condition satisfied by 96.95 percent of students on average across posterior draws.

across students. I next ask whether observable student characteristics predict distinctive sorting patterns.

3.4 Observable Heterogeneity

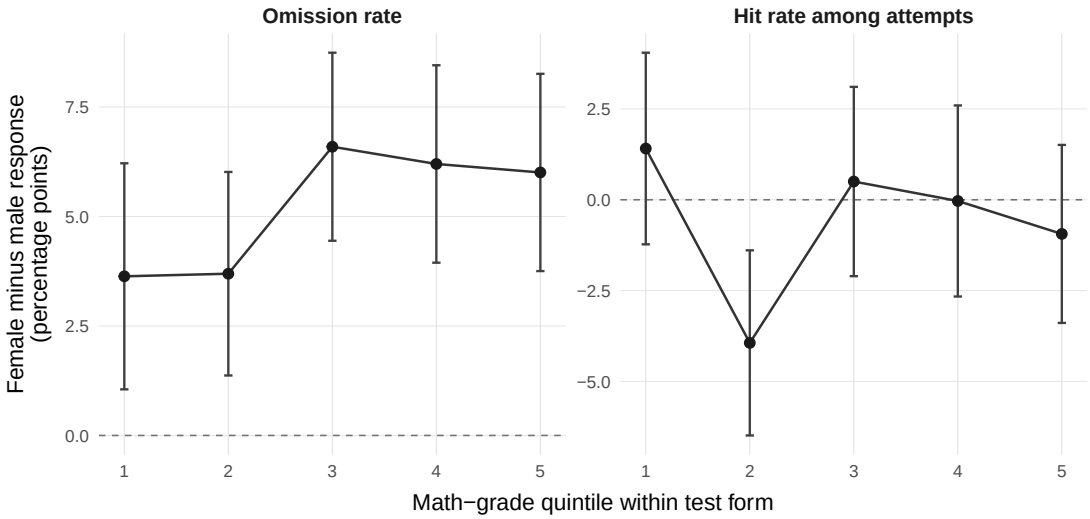
To ask whether observables organize the heterogeneity documented above, I augment Equation (6) by interacting the scoring regime with gender, standardized school mathematics grade, previous participation, and an indicator for attending a non-public school.

Table A6 reports the estimates. The heterogeneity is concentrated especially on the omission margin. Repeat participants and students in non-public schools increase omissions by 3.32 and 3.76 percentage points less, respectively, while their hit-rate responses do not differ significantly from the reference group. Mathematics achievement instead predicts somewhat larger responses on both margins: a one-standard-deviation increase in school mathematics grade is associated with a 0.91-point larger omission response and a 0.97-point larger hit-rate response. Gender stands out. Conditional on the other characteristics, female students increase omissions by 5.14 percentage points more than male students, consistent with the large literature documenting gender differences in willingness to guess in tests (e.g. Akyol et al., 2016; Baldiga, 2014; Coffman and Klinowski, 2020; Díez-Rituerto et al., 2025; Iriberri and Rey-Biel, 2021; Krishna et al., 2025), but their hit-rate response is statistically indistinguishable from that of male students. This difference is especially striking because female students have slightly higher school mathematics grades on average, by about 0.14 standard deviations.

I therefore examine how the gender gaps in omission and hit-rate responses vary with ability. Figure 3 reports female-minus-male differences in regime responses across quintiles of school mathematics grade, constructed within sitting. The left panel shows a pronounced gradient in omissions: the female excess omission response is about 3.6–3.7 percentage points in the bottom two quintiles and rises to roughly 6.0–6.6 points in the upper three. Overall, the female–male gap in the omission response is 2.35 percentage points larger among above-mean-grade students than among below-mean ones. The right panel shows no corresponding gradient in selection quality. Female-minus-male hit-rate responses fluctuate around zero throughout the achievement distribution and are negative in the second quintile; the corresponding above- versus below-mean difference is statistically indistinguishable from zero (see Table A7). These patterns create two related puzzles.

Can Preferences Explain the Omission Gradient? First, why does the gender gap in omissions widen with ability? A candidate explanation is preference heterogeneity: students who are less willing to answer may respond more strongly to penalties. The survey measures in Iriberri and Rey-Biel (2021) suggest that differences in willingness to answer may account for gender differences in the level of omission, but not for why the gap widens with ability.

Figure 3: Gender Differences in Regime Responses by School Math Grade Quintile



Notes: Points report female-minus-male differences in the within-student regime response, in percentage points, separately by school-math grade quintile. Quintiles are constructed within sitting using average ranks, with tied grades kept together. Estimates use the weighted student-question specification with student fixed effects and sitting-specific regime slopes. Intervals are 95 percent confidence intervals clustered two ways by physical participant and form-specific item. The differences describe heterogeneity in regime responses and should not be interpreted as causal effects of gender or achievement.

Female students report significantly lower willingness to answer than male students, yet this gender gap does not widen with mathematics achievement: the female-by-math interaction is 0.024 and statistically indistinguishable from zero, with a sign that, if anything, goes in the opposite direction. Moreover, in the 2017–2018 survey sample, conditioning directly on willingness to answer leaves a female-by-ability interaction in the omission response of 2.86 percentage points, which remains significant; see Table A8.

More Omission, but No Better Selection Second, a separate puzzle arises without relying on any measure of preferences. Holding observed ability fixed, if the larger female omission response reflected only a higher answering threshold, the additional selection should translate into a larger improvement in the quality of retained answers. It does not, as shown in Figure 3; moreover, Figure A7 shows that the same quantity–quality wedge appears when ability is measured using performance on the test itself rather than school grades.

Whereas these comparisons are not themselves formal cross-group tests of Rank Preserving Metacognition, they are difficult to reconcile with a simple account in which females and males differ in preferences but rank questions in the same way, further motivating a model to understand how students depart from the benchmark.

4 Structural Assessment of Metacognitive Errors

The evidence above suggests that the Rank-Preserving Metacognition benchmark is too restrictive to capture students' selection behavior. Students respond strongly to omission incentives, yet the additional omissions often produce little improvement in hit-rates and sometimes generate adverse selection. To interpret this pattern, I specialize the framework from Section 1 to a parsimonious parametrization, which allows false positives and false negatives in the mapping from knowledge to confidence. The parametrization permits direct violations of the monotonicity restriction in Definition 1, and it yields simple expressions that provide an interpretable decomposition of the reduced form objects.

4.1 A Binary Decomposition

Fix a match type (s, q). I restrict latent knowledge to two values, $K \in \{\underline{k}, \overline{k}\}$, with $0 \leq \underline{k} < \overline{k} \leq 1$. The parameter $\underline{k}$ is the success probability in the low-knowledge state, while $\overline{k}$ is the success probability in the high-knowledge state. In a multiple-choice setting, the former can be interpreted as a state in which the student is effectively guessing, and the latter as one in which they know the answer. Let

$$\pi_{s,q} := \mathbb{P}(K = \overline{k} \mid S = s, Q = q)$$

denote the probability that student type s is in the high-knowledge state when facing a question of type q.

I similarly restrict confidence to two values, $C \in \{\underline{c}, \overline{c}\}$, with $0 \leq \underline{c} < \overline{c} \leq 1$. The stochastic map from knowledge to confidence is then summarized by

$$\psi^{+}_{s,q} := \mathbb{P}(C = \overline{c} \mid K = \overline{k}, s, q), \qquad \psi^{-}_{s,q} := \mathbb{P}(C = \underline{c} \mid K = \underline{k}, s, q).$$

The parameter $\psi^{+}_{s,q}$ captures *positive metacognition*: the ability to recognize high-knowledge states, or to know what one knows. The parameter $\psi^{-}_{s,q}$ captures *negative metacognition*: the ability to recognize low-knowledge states, or to know what one does not know. Metacognitive errors can therefore arise in two directions. Weak positive metacognition generates false negatives: students may omit questions they could answer successfully, an illusion of ignorance. Weak negative metacognition generates false positives: students may attempt questions they are unlikely to answer successfully, an illusion of knowledge.

Define *metacognitive sensitivity* as the centered sum of the two parameters

$$\psi^{\Sigma}_{s,q} := \psi^{+}_{s,q} + \psi^{-}_{s,q} - 1$$

The next result shows this sufficient statistic governs the link between the quantity and quality margins of selection.

Proposition 2 (Binary selection frontier)**.** *In the binary model, for each match type* (s, q)*, either* $O_{s,q,1} = 0$ *and* $H_{s,q,1} = H_{s,q,0}$*, or* $O_{s,q,1} > 0$ *and*

$$\underbrace{(H_{s,q,1} - H_{s,q,0})}_{quality}\underbrace{(1 - O_{s,q,1})}_{quantity} = (\overline{k} - \underline{k})\pi_{s,q}(1 - \pi_{s,q})\psi^{\Sigma}_{s,q}, \tag{9}$$

The result follows by direct algebra; its interest lies in interpretation. The left-hand side of Equation (9) defines the—possibly negative—surplus from selection: the gain in average quality from selective answering, multiplied by the share of questions answered. Equivalently, it is the covariance between the decision to answer and latent success, $\text{Cov}(D, Y^*)$. The right-hand side decomposes this surplus into two factors. The first, $(\overline{k} - \underline{k})\pi_{s,q}(1 - \pi_{s,q})$, captures the scope for sorting, which is larger when knowledge varies more across matches. The second is metacognitive sensitivity, $\psi^{\Sigma}_{s,q}$, which determines whether the student decision converts this knowledge variation into gains or losses from selection.

Proposition 2 therefore defines a "technological frontier" for the selection problem. When $\psi^{\Sigma}_{s,q} > 0$, confidence ranks knowledge positively: selective answering disproportionately removes low-knowledge matches and raises the hit-rate. When $\psi^{\Sigma}_{s,q} = 0$, selection is effectively random, so omission changes quantity but not quality. When $\psi^{\Sigma}_{s,q} < 0$, confidence ranks knowledge adversely: the student disproportionately omits good matches and retains bad ones, so answering more partly *insures* them against their own selection mistakes.

In this picture, a different combination of the same primitives governs the quantity margin. Define *metacognitive bias* as

$$\psi^{\Delta}_{s,q} := \psi^{+}_{s,q} - \psi^{-}_{s,q}.$$

Holding sensitivity fixed, higher $\psi^{\Delta}_{s,q}$ makes high-confidence states more frequent and therefore increases answering. Bias and sensitivity thus separate two margins that standard models of confidence levels tend to conflate: bias governs how much a student answers, while sensitivity governs whether selective answering improves performance. A student may therefore omit many questions because confidence is low on average, yet gain little from doing so if confidence discriminates knowledge only weakly.

This distinction helps to explain the puzzling reduced-form evidence. Weak and sometimes adverse sorting suggests that sensitivity may be low or negative for a substantial share of students. The same hypothesis also clarifies the gender-by-ability pattern. If female students have lower metacognitive bias, they will omit more questions, but unless they also have sufficiently positive sensitivity, this additional omission may fail to translate into higher hit-rates. The next subsection makes these intuitions precise by taking the model to the data.

4.2 Structural Estimation

I estimate the binary decomposition using a hierarchical Bayesian model that parallels the reduced-form model in Section 3.3. The unit of observation remains the student–question match, and the hierarchy again separates heterogeneity across sittings, students, and questions from measurement error at each level. Here, however, I parameterize omission and hit-rate responses through the three latent objects introduced above: knowledge $\pi_{i,j}$, positive metacognition $\psi^{+}_{i,j}$, and negative metacognition $\psi^{-}_{i,j}$.

Conditional on a choice of $(\underline{k}, \overline{k})$, the three reduced-form rates map one-to-one into these objects, except in degenerate boundary cases. The hit-rate in the nonselective regime identifies the mixture of the two knowledge states and therefore their relative frequency, π. In the selective regime, the attempt rate and the difference in hit-rates jointly identify ψ^{+} and ψ^{-}. For estimation, I set $\underline{k} = 0.20$, corresponding to random guessing among five alternatives, and $\overline{k} = 1$, corresponding to knowing the correct answer.

The hierarchy mirrors the reduced-form specification. For $r \in \{\pi, +, -\}$, let $\theta^{r}_{i,j}$ denote $\pi_{i,j}$, $\psi^{+}_{i,j}$, and $\psi^{-}_{i,j}$, respectively. I specify

$$\operatorname{logit}(\theta^{r}_{i,j}) = m^{r}_{b(i)} + X_i'\beta_r + s^{r}_{i} + W_j'\gamma_r + u^{r}_{j},$$

where $m^{r}_{b(i)}$ allows the center of the student distribution to vary across sittings, X_i and W_j collect observed student and question characteristics, and s^{r}_{i} and u^{r}_{j} capture residual student and item heterogeneity, as in Section 3.3. Sitting means are partially pooled around year and level components,

$$m^{r}_{b} = \mu_r + \lambda^{r}_{\text{year}(b)} + \ell^{r}_{\text{level}(b)} + \tau_r \zeta^{r}_{b}, \qquad \zeta^{r}_{b} \sim N(0,1).$$

I model the three student residuals jointly, $(s^{\pi}_{i}, s^{+}_{i}, s^{-}_{i})' \sim N(0, \Sigma_S)$, allowing latent knowledge and the two forms of metacognition to be correlated across students. Residual item effects vary separately across the three primitives. As in the reduced-form model, I center them within sitting and regime so that they capture within-block question heterogeneity rather than average differences between the two parts of a test. I place weakly regularizing priors on the hierarchical coefficients and variance components and an LKJ prior on the student correlation matrix. Section D.2 reports the complete prior specification and computational details.

Given the model's simplicity, I restrict the structural estimation sample to observations for which its interpretation is most credible. First, I require no Part 1 omissions, so that the low-cost block provides a nonselective measure of knowledge. I retain items with a hit rate of at least 0.15 in this Part-1-completer sample; for Part 2 items, the hit rate is calculated among students who attempt the item. This restriction excludes questions whose observed performance is difficult to reconcile with the $\underline{k} = 0.20$ guessing benchmark.

Finally, I require at least one omission among the retained Part 2 items, ensuring that the selective block contains evidence of an answer–omit margin rather than all-attempting behavior. The resulting sample contains 6,789 student-sittings, 282 questions, and 159,240 student–question observations.

4.3 Structural Results

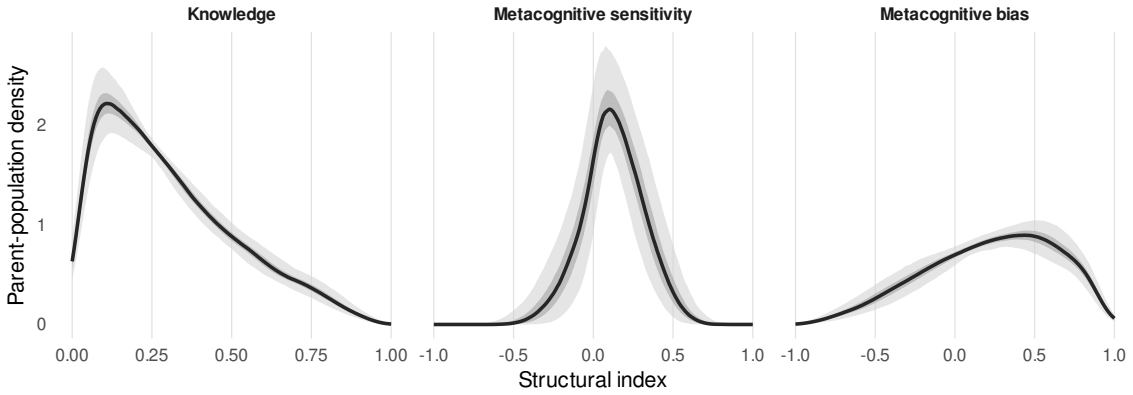
I now turn to the results of the structural decomposition. The model closely reproduces the response margins used for estimation. At the level of student-covariates profile, the root mean squared errors are 0.9 percentage points for Part 1 correctness, 1.2 points for Part 2 omission, and 1.6 points for Part 2 correctness conditional on attempt. The corresponding item-covariates profile errors are 1.4, 0.2, and 1.7 percentage points. Appendix Figure A8 reports the underlying fit and additional diagnostics.

Population Heterogeneity Figure 4 reports the parent-population distributions of knowledge, metacognitive sensitivity, and metacognitive bias. These distributions provide the structural counterpart to the individual treatment effects distribution estimated in Figure 2. Solid curves report the estimated population distributions, while the shaded bands capture posterior uncertainty about them.

The population mean of knowledge is 0.302, with a 95 percent credible interval of $[0.271, 0.339]$, implying that students are in the high-knowledge state in roughly 30 percent of student–question matches. Knowledge also varies substantially across students, with a population standard deviation of 0.209. Metacognitive sensitivity is positive but modest on average: its mean is 0.124, $[0.025, 0.227]$, relative to the benchmark $\psi^{\Sigma} = 0$ under random sorting. Confidence therefore ranks knowledge correctly on average, but only weakly. Sensitivity also varies across students, with a population standard deviation of 0.189, $[0.146, 0.232]$, and a distribution that extends below zero, implying adverse sorting for part of the population. Heterogeneity in metacognitive bias is larger. Its mean is 0.222, $[0.130, 0.313]$, indicating average overconfidence and hence excessive answering, while its population standard deviation is 0.401, $[0.364, 0.434]$, more than twice that of sensitivity. Students thus differ much more in their tendency to answer than in how well confidence ranks their knowledge. Bias and sensitivity also vary largely independently: their population correlation is 0.078, $[-0.168, 0.273]$.

Together, these two facts provide the structural counterpart to the joint reduced-form distribution in Figure 2: omission responses are more dispersed than hit-rate responses, and the two are only weakly related. The decomposition attributes these patterns to greater heterogeneity in bias than in sensitivity and to little correlation between the two. Knowledge, by contrast, is positively associated with sensitivity, with correlation 0.447, $[0.258, 0.624]$, and more moderately with bias, 0.223, $[0.108, 0.328]$. Higher-knowledge

Figure 4: Population Distributions of Structural Objects



Notes: The figure reports parent-population distributions of knowledge, metacognitive sensitivity, and metacognitive bias. Solid lines report posterior median densities; dark and light bands report pointwise 50 and 95 percent posterior credible intervals. The distributions are evaluated on the observed year-by-level support, fixing residual sitting heterogeneity at its parent mean, integrating over the parent distribution of residual student heterogeneity, and using the average-item convention.

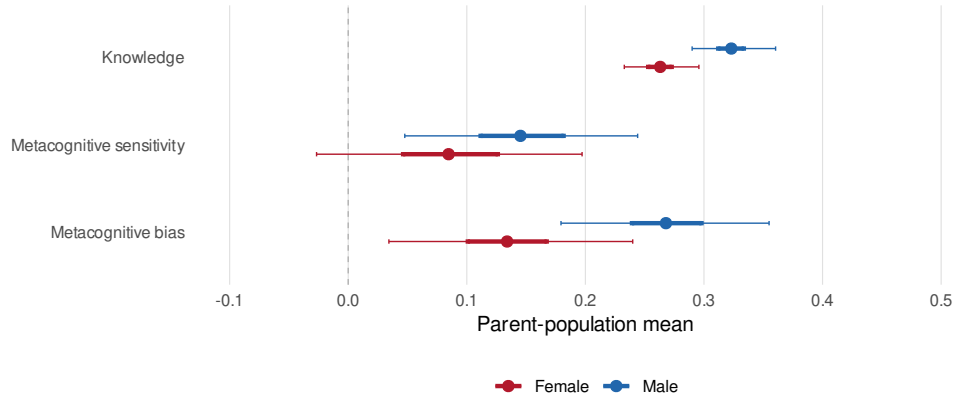
students therefore tend to sort their knowledge more effectively and, to a lesser extent, to display greater confidence.

Gender Heterogeneity Figure 5 returns to the gender-by-ability pattern documented above. Female students have slightly lower knowledge: (0.263 versus 0.323, a -6.0 pp difference with credible interval $[-7.1, -4.8]$), and slightly lower metacognitive sensitivity (0.085 versus 0.145; -6.1 pp $[-9.8, -2.6]$). The much larger difference is in metacognitive bias: 0.134 for female students versus 0.268 for male students, a gap of -13.4 pp $[-16.5, -10.2]$. Female students are therefore substantially less overconfident and answer less, but this greater omission does not sort knowledge more effectively.

This combination accounts for the reduced-form gender-by-ability puzzle documented in Section 3.4. Lower bias generates the stronger quantity response to the penalty, while close to null sensitivity explains why the additional omissions do not translate into a corresponding improvement in the quality of retained answers. Indeed, relative to male students, female students have 9.7 percentage points lower recognition of knowledge, ψ^{+}, but 3.7 points higher recognition of ignorance, ψ^{-}. They are therefore better at screening out some low-knowledge questions, but also more likely to omit questions they could answer correctly. As knowledge rises, moreover, these false negatives become increasingly frequent, explaining the ability gradient, too.

Other Observable Heterogeneity Figure A9 shows that sensitivity and bias also vary differently with other observed student characteristics. A one-standard-deviation increase in school mathematics grade is associated with 6.5 percentage points higher latent knowledge and 2.2 points higher sensitivity, while leaving bias essentially unchanged. Prior participation is associated with substantially higher knowledge, by 11.0 points, but its

Figure 5: Gender Heterogeneity in Structural Objects



Notes: The figure reports parent-population means of knowledge, metacognitive sensitivity, and metacognitive bias by gender. Points report posterior means; thick and thin intervals report 50 and 95 percent posterior credible intervals, respectively. The corresponding female-minus-male differences are −0.060, −0.061, and −0.134.

metacognitive association operates mainly through bias: bias is 7.9 points higher, compared with a 2.2-point difference in sensitivity that is imprecisely estimated. Students in the upper grade within an age level and students in non-public schools likewise have higher knowledge and higher bias, but little systematic difference in sensitivity. These patterns further suggest that the two metacognitive margins capture different forms of heterogeneity as characteristics that strongly predict one are not associated with the other.

Validation Checks Two checks provide independent validation of the student-level structural estimates. First, among participants observed in multiple years, the estimated metacognitive profiles are persistent across test administrations (Table A9). Adjacent-year Spearman correlations are about 0.35 for sensitivity and 0.43 for bias, with similar correlations over a two-year horizon. Second, bias predicts the survey confidence measure, which is not used anywhere in estimation of the structural model (Table A10). Conditional on realized test performance and sitting fixed effects, a one-standard-deviation increase in bias is associated with about one additional forecasted correct answer; the association remains strong after controlling for observed student characteristics. Sensitivity is less predictive, and its association becomes small and statistically indistinguishable from zero after covariate adjustment. These checks suggest that the model's parameters capture meaningful behavioral differences rather than test-specific response noise.

Overall, the structural estimates organize the reduced-form evidence around two distinct metacognitive margins: bias governs how much students answer, while sensitivity governs how informative that selection is about knowledge. This decomposition also makes it possible to ask how each margin enters observed test scores and to evaluate counterfactuals relevant for policy, which I turn to next.

5 Policy Implications

The preceding sections show that test-taking choices reveal imperfect metacognition and that students differ substantially along this margin. This section asks why these differences may matter for policy, focusing on their consequences for test scores. I first show that penalties for errors reward not only knowledge, but also metacognitive sensitivity and bias. I then show that this contaminates the mapping from knowledge into scores, generating both information loss in decoding scores and ranking reversals. Finally, I show that the same mechanism contributes substantially to the gender score gap and helps explain why it widens with ability.

5.1 Test Score Decomposition

I start by showing how different scoring rules reward latent knowledge and the two dimensions of metacognition. Fix a regime $z \in \{0, 1\}$. Let a correct answer receive $h_z > 0$ points, an incorrect answer zero points (normalization), and an omission $0 \leq o_z < h_z$ points. Let $W_{s,z}$ denote the expected score per question for type s. For compactness, define

$$g_z(\overline{k}) := h_z\overline{k} - o_z, \qquad g_z(\underline{k}) := o_z - h_z\underline{k}.$$

The first term is the gain from answering rather than omitting in the high-knowledge state; the second is the gain from omitting rather than answering in the low-knowledge state.

Proposition 3 (Score decomposition)**.** *In the binary model, fix a student type s. Either $O_{s,z} = 0$, in which case $W_{s,z} = h_z\mathbb{E}_{\pi_s}[k]$, or $O_{s,z} > 0$, in which case*

$$W_{s,z} = K_z(\pi_s) + \frac{1}{2}\left[A_z(\pi_s)\psi_s^{\Sigma} + B_z(\pi_s)\psi_s^{\Delta}\right], \tag{10}$$

where

$$K_z(\pi_s) := \frac{h_z\mathbb{E}_{\pi_s}[k] + o_z}{2}, \qquad A_z(\pi_s) := \mathbb{E}_{\pi_s}[g_z(k)], \qquad B_z(\pi_s) := h_z\mathbb{E}_{\pi_s}[k] - o_z.$$

The decomposition separates three potential sources of score variation. When every question is answered, expected scores correspond to the average knowledge level, up to scaling. Under selective answering, knowledge still directly increases scores via $K_z(\pi_s)$, but also sensitivity and bias enter, through a quadratic form. The coefficient $A_z(\pi_s)$ determines the return to sensitivity, which equals the expected value of making the best *state-contingent* choice, and therefore grows in the spread of the knowledge distribution. By contrast, $B_z(\pi_s)$ determines the return to bias, which equals expected payoff gain from answering *independently* from the knowledge state, and therefore grows in the average knowledge level.

A scoring rule can therefore reward a general tendency to answer even when answering discriminates only weakly between high- and low-knowledge states. The scoring rule used in the experiment provides a case in point. It is an instance of the widespread *formula scoring rule* (Holzinger, 1924; Thurstone, 1919), under which omission receives the same expected payoff as random guessing. In the experiment, a correct answer receives $h_1 = 5$ points and an omission $o_1 = 1$. When $\overline{k} > \underline{k} = 0.20$ (in particular, $\overline{k} = 1$)

$$\frac{\partial W_{s,1}}{\partial \psi_s^-} = (1 - \pi_s)(o_1 - h_1\underline{k}) = 0, \qquad \frac{\partial W_{s,1}}{\partial \psi_s^+} = \pi_s(h_1\overline{k} - o_1) > 0.$$

Therefore, recognizing a low-knowledge state and omitting it yields no expected-score gain, whereas recognizing a high-knowledge state and answering does. Formula scoring thus rewards knowledge and recognition of knowledge, but not recognition of ignorance. Equivalently, conditional on knowledge, sensitivity and bias receive the same marginal score return. Substituting into Equation (10) gives

$$W_{s,1} = 1 + 2\pi_s \left(1 + \psi_s^{\Sigma} + \psi_s^{\Delta}\right). \tag{11}$$

The score therefore rewards sensitivity—arguably, a desirable feature—but it also rewards overconfidence—arguably, an undesirable one.[34]

This asymmetry has two related consequences. First, scores need not measure knowledge alone: students with the same latent knowledge can obtain different expected scores because they differ in metacognition, and the same score can arise from different combinations of the two. Second, systematic group differences in metacognition can translate into score gaps even at comparable levels of knowledge, with potential consequences for group discrimination. I examine these implications in turn.

5.2 Biased Decoding of Test Scores

Test scores are typically interpreted as noisy measures of knowledge. The decomposition above shows that, under penalties for errors, this interpretation is incomplete: even *predicted* scores can differ across students with the same knowledge because of differences in metacognition. A knowledge-only reading of scores therefore attributes to knowledge variation that partly comes from how students' knowledge maps into the decision to answer a question.

I quantify this distortion using the structural model. For each posterior draw, I construct the predicted Part 2 score over the retained item set. This yields the following student-level

[34]The score-neutrality of negative metacognition is specific to $\underline{k} = 0.20$, but the asymmetry is more general. Indeed, if $\underline{k} > 0.20$, recognizing ignorance is actually *penalized* because answering a low-knowledge question has higher expected payoff than omitting it. Values $\underline{k} < 0.20$ would instead reward recognition of ignorance, but are harder to conceive of, as they correspond to worse than random performance. Thus, among economically sensible calibrations, $\underline{k} = 0.20$ is the most favorable to recognition of ignorance.

factorization. Let J_i be the set of questions answered by student i,

$$W_i = 1 + 4 \underbrace{\left(\frac{1}{|J_i|} \sum_j \pi_{ij} \right)}_{K_i} \underbrace{\left(\frac{\sum_j \pi_{ij} \psi_{ij}^+}{\sum_j \pi_{ij}} \right)}_{M_i},$$

where K_i is average latent knowledge and M_i is the factor influenced by metacognition. I use this factorization to measure the resulting loss of information in three ways, summarized in Figure 6.

First, I ask how much expected-score variation comes from metacognition rather than knowledge. Because the two components enter multiplicatively, I use a symmetric Shapley decomposition, computing the marginal contribution of each component in both possible orderings and averaging them. Figure 6 shows that metacognition accounts for 26.7 percent of cross-student expected-score variance, with a 95 percent credible interval of $[19.5, 33.5]$. Most variation still reflects knowledge, but roughly one quarter comes from differences in how effectively students convert knowledge into attempted answers.

Second, I ask how much knowledge can be recovered from expected scores. Define the *knowledge decoding loss* as

$$L_K = \frac{\mathbb{E}\Big[\{K_i - \mathbb{E}[K_i \mid W_i]\}^2\Big]}{\mathbb{V}(K_i)}.$$

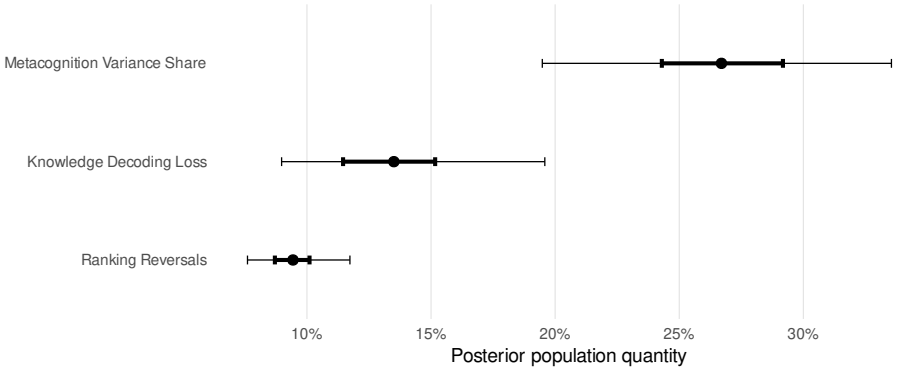
The quantity measures the share of latent-knowledge variation that remains unexplained after decoding knowledge from scores via a standard predictor, i.e. the conditional expectation. I estimate the conditional mean flexibly using equal-frequency score bins, repeating the calculation within each posterior draw. The posterior mean of L_K is 13.5 percent, with a 95 percent credible interval of $[9.0, 19.6]$. Thus, even allowing for a flexible nonlinear mapping from expected score to latent knowledge, a meaningful share of knowledge variation cannot be recovered from the score.

Third, I ask how much this loss of information changes student rankings. Among pairs of students facing the same sitting and test environment, I record a reversal whenever their predicted scores rank oppositely to their latent knowledge,

$$(K_i - K_{i'})(W_i - W_{i'}) < 0.$$

The posterior mean reversal probability is 9.4 percent, with a 95 percent credible interval of $[7.6, 11.7]$. Roughly one in ten comparable student pairs is therefore ranked differently by expected formula score and latent knowledge. Note that these reversals arise in expected scores only, and can only be amplified from finite-test sampling noise.

Figure 6: Metacognition and the Decoding of Test Scores



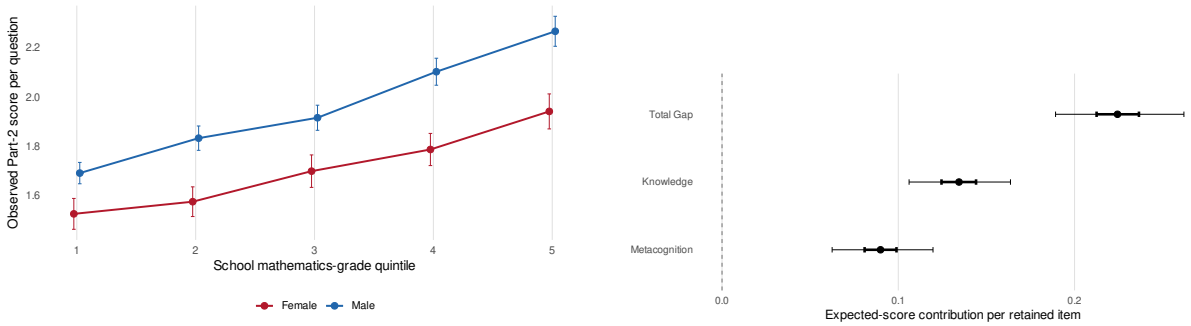
Notes: The figure reports three posterior population quantities. Metacognition Variance Share is the Shapley contribution of score-relevant metacognition to cross-student variation in predicted Part 2 scores. Knowledge Decoding Loss is the share of latent-knowledge variance left unexplained by a flexible conditional mean of knowledge given expected score. Ranking Reversals is the probability that two students facing the same sitting and test environment are ranked in opposite order by latent knowledge and predicted score. Points report posterior means; thick and thin intervals report 50 and 95 percent credible intervals, respectively.

5.3 The Widening Gender Gap

The score decomposition also quantifies how penalties for errors affect measured gender gaps. The reduced-form evidence in Section 3.4 showed that female students omit more than male students, especially at higher ability levels, without the corresponding hit-rate gains predicted by Rank-Preserving Metacognition. The structural estimates explain this pattern. Female students have lower metacognitive bias and slightly lower sensitivity: they answer more selectively than male students, but their selection does not better discriminate the knowledge state. Under formula scoring, this combination is particularly costly because recognizing low-knowledge states carries no direct score return, while recognizing and attempting high-knowledge states does. The model therefore predicts a larger gender score gap among higher-knowledge students.

Panel A of Figure 7 shows this pattern in the data. It reports observed Part 2 scores by gender and quintile of school mathematics grade, with quintiles constructed within sitting. Male students score more highly throughout the distribution, but the gap widens substantially with ability. The male–female difference rises from 0.165 points per question in the bottom quintile, $[0.089, 0.241]$, to 0.324 points in the top quintile, $[0.231, 0.417]$. Thus, the observed score gap roughly doubles between the bottom and top of the ability distribution.

Panel B therefore asks how much of the predicted gender gap comes from knowledge and how much from the metacognitive channel rewarded by formula scoring. For each posterior draw, I construct counterfactual combinations of male and female knowledge and positive metacognition. Again, since the two components enter scores multiplicatively, I use a symmetric Shapley decomposition. The predicted male–female gap is 0.224 expected points per retained Part 2 item, with interval $[0.189, 0.262]$. Differences in knowledge account for 0.134 points, with interval $[0.106, 0.164]$, or about 60 percent of the gap. The remaining



(a) Observed scores by school mathematics grade

(b) Structural decomposition

Figure 7: The Gender Score Gap

Notes: Panel A reports observed Part 2 scores by gender and school mathematics-grade quintile. Quintiles are constructed within sitting; intervals are 95 percent confidence intervals clustered by physical participant. Panel A uses the broad reduced-form analysis sample and all twelve Part 2 questions. Panel B reports the Shapley decomposition of the predicted male–female expected-score gap in the structural population into knowledge and score-relevant metacognition. Points report posterior means; thick and thin intervals report 50 and 95 percent credible intervals. Because the two panels target different samples and objects, Panel B provides a structural account of the mechanism rather than a mechanical decomposition of the observed gaps in Panel A.

0.090 points, with interval $[0.062, 0.120]$, or about 40 percent, come from differences in recognizing and attempting high-knowledge states. Formula scoring therefore turns gender differences in metacognition into measured performance differences over and above the underlying difference in knowledge.

I further decompose the structural gender gap by latent-knowledge quintile, separating differences in expected scores within knowledge strata from differences in the distribution of knowledge itself. The decomposition follows the logic of Oaxaca–Blinder (Blinder, 1973; Oaxaca, 1973), adapted to this nonlinear setting. For each posterior draw, let $m_g(b)$ denote the predicted score for gender g in latent-knowledge bin b, and $p_g(b)$ the corresponding population share. Using the male score schedule as reference,

$$\mathbb{E}[W_i \mid M] - \mathbb{E}[W_i \mid F] = \underbrace{\sum_b p_F(b)\,\{m_M(b) - m_F(b)\}}_{\text{within-knowledge score schedule}} + \underbrace{\sum_b \{p_M(b) - p_F(b)\}\, m_M(b)}_{\text{knowledge composition}}.$$

The first term measures the gender score gap among students in the same latent-knowledge bins; the second reflects differences in how male and female students are distributed across those bins. Appendix Figure A10 reports the results. Of the 0.224-point predicted gap, 0.074 points with interval $[0.050, 0.100]$, or about 33 percent, remain within latent-knowledge bins, while 0.150 points with interval $[0.118, 0.181]$, or about 67 percent, come from differences in knowledge composition. Latent knowledge therefore accounts for about two thirds of the aggregate gap. The remaining third arises among students with comparable knowledge, reflecting gender differences in metacognition.

6 Internal and External Validity

I conclude by providing additional evidence on the mechanisms underlying the paper's main results. I first exploit variation within the Iriberri and Rey-Biel (2021) data to assess whether students sort questions according to objective item difficulty. I then turn to two external settings: one in which confidence is observed directly (Bordalo et al., 2019) and one involving a reform in a national high-stakes college-entry examination (Coffman and Klinowski, 2020). Together, these exercises provide additional robustness for the findings and assess how the documented patterns in metacognitive sensitivity, confidence bias, and gender-by-ability heterogeneity extend beyond the main experimental setting.

6.1 Student Choice and Item Difficulty

The main analysis compares student choice between two scoring regimes. That comparison uses the experimental change in incentives, but also requires the questions assigned to the two parts of the test to be comparable. I therefore develop a complementary test that uses only choices within the penalty regime. Instead of varying the opportunity cost of answering, I use an observable proxy for a student's latent knowledge on a question: item easiness. Formally, I impose the following assumption.

Assumption 5 (Proxy)**.** The analyst observes a variable $R \in \mathbb{R}$ such that, for every student-type s,

(i) *Relevance.* If $r_2 > r_1$, then $(K \mid S = s, R = r_2) \succeq_{FOSD} (K \mid S = s, R = r_1)$.

(ii) *Exogeneity.* $D \perp R \mid (K, S = s)$.

Part (i) requires the proxy to rank questions by latent knowledge, in the weak sense of first-order stochastic dominance. Part (ii) requires the proxy to affect the decision to answer only through its relationship with latent knowledge. The proxy therefore cannot capture question features, such as framing, that directly shift confidence and selection holding knowledge fixed.

Under Assumption 5, the sorting logic of Lemma 1 implies that attempted questions should be easier, according to the proxy, than omitted questions. One way to express this implication is to treat the student's answer–omit decision as a binary classifier. Let $R_s^{(1)} \sim (R \mid D = 1, S = s)$ and $R_s^{(0)} \sim (R \mid D = 0, S = s)$ be independent draws of the proxy variable from the attempted and omitted pools, respectively, and define

$$\mathrm{AUC}_s := \mathbb{P}\Big(R_s^{(1)} > R_s^{(0)}\Big) + \frac{1}{2}\mathbb{P}\Big(R_s^{(1)} = R_s^{(0)}\Big).$$

The AUC is the probability that a randomly chosen attempted question has a higher proxy value than a randomly chosen omitted question, with ties split at random. Random sorting

gives $\text{AUC}_s = 1/2$; values below 1/2 indicate adverse sorting. Then, we have the following result.

Proposition 4 (Within-regime test)**.** *Suppose Assumption 1 and Assumption 5 hold, and that* $0 < \mathbb{P}(D = 1 \mid S = s) < 1$*. If students of type* s *have Rank-Preserving Metacognition, then*

$$(R \mid D = 1, S = s) \succeq_{FOSD} (R \mid D = 0, S = s) \qquad \textit{and} \qquad \text{AUC}_s \geq \frac{1}{2}.$$

The result follows from the same sorting argument as Lemma 1. Under Assumption 5(i) and RPM, both confidence and the observable proxy increase with latent knowledge. Selection into answering should therefore be positive not only on K, which is unobserved, but also on R, which is observed. Hence, this test replaces block comparability with the requirement that the proxy rank relative knowledge within the penalty block.

Estimation To bring the test to the data, I construct a ranking of the twelve Part 2 questions by easiness separately within each sitting. I estimate this ranking using the 903 students who attempt every question in the test. Restricting to this *full-completer sample* is important because correctness is observed only for attempted questions: using the full Part 2 sample would treat the endogenous omissions under study as ignorable.

I estimate relative item easiness using a Rasch model (De Ayala, 2013),

$$\Pr(Y_{ibj} = 1 \mid \alpha_{ib}, e_{bj}) = \Lambda(\alpha_{ib} + e_{bj}),$$

where α_{ib} captures student ability, e_{bj} item easiness, and $\Lambda(\cdot)$ is the logistic CDF. The common coefficient on ability implies a single ability-invariant ranking: if $e_{bj} > e_{bk}$, item j is easier than item k at every ability level. To estimate this ranking without estimating student ability, I condition on each student's total number of correct Part 2 answers. Once this total is fixed, the likelihood compares *which* questions the student answered correctly, rather than how many, and the student-specific ability term cancels. Relative easiness is therefore identified from systematic differences in which items are answered correctly among students with the same total score. Only relative easiness matters for the ranking, so I impose the normalization $\sum_j e_{bj} = 0$.

Because this ranking is estimated from a finite calibration sample, rather than using plug-in estimates of item easiness, I also propagate uncertainty about the ordering of items. To this end, I combine conditional Rasch likelihood above with weakly regularizing priors on item easiness and compute, for every pair j, k, the posterior probability $p_{bjk} = \Pr(e_{bj} > e_{bk} \mid \mathcal{C}_b)$, where $\mathcal{C}_b$ denotes the full-completer calibration data in sitting b. As a consequence, clearly ordered pairs receive weight close to zero or one, while uncertain comparisons receive weight closer to one half.[35]

[35]In practice, the ranking is sharp. Across the 792 within-sitting item pairs, the median posterior

For each student i who both attempts and omits at least one Part 2 question, let $\mathcal{A}_i$ and $\mathcal{M}_i$ denote the attempted and omitted sets. If item easiness were known, the student's AUC would be the fraction of attempted–omitted pairs for which the attempted item is easier. I instead average this statistic over posterior uncertainty in the item ranking:

$$\widetilde{\mathrm{AUC}}_i = \frac{1}{|\mathcal{A}_i||\mathcal{M}_i|} \sum_{j \in \mathcal{A}_i} \sum_{k \in \mathcal{M}_i} \Pr(e_{b(i)j} > e_{b(i)k} \mid \mathcal{C}_{b(i)}).$$

An attempted–omitted pair therefore receives weight close to one when the attempted item is very likely to be easier, close to zero when it is very likely to be harder, and close to one half when the ordering is uncertain. The resulting *soft* AUC preserves the ordinal interpretation of the plug-in statistic while carrying uncertainty in the estimated item ranking into the measure of sorting. Appendix D.3 provides additional details.

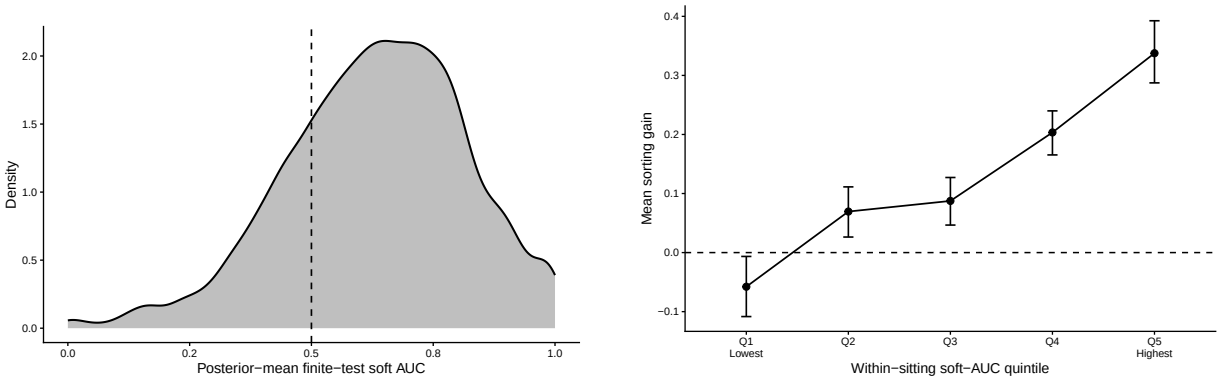
Validation of Full Completer Rankings Full completers are selected, so observing all of their responses solves the missing-outcome problem but does not by itself establish that their item ranking applies to students who omit. I assess this transportability using Part 1, where the large majority of students who are not full completers nevertheless answer all thirteen questions. Within each sitting, I compare the Rasch ranking estimated from full completers with rankings estimated from equally sized samples of these other full Part 1 responders. Across repeated matched splits, the equal-sitting mean Spearman correlation between the full-completer and non-full-completer rankings is 0.889 whereas the corresponding correlation between two equally sized non-full-completer samples is 0.898.[36] Hence, the relative item ordering estimated from them agrees with the ordering from other students almost as closely as two equally sized non-full-completer samples agree with one another. Table A13 reports the corresponding ranking-validation and selection statistics.

Results Figure 8 reports the results. Panel (a) shows the distribution of posterior-mean soft AUC among the 8,834 students for whom the statistic is defined. I first estimate the mean soft AUC separately within each sitting, combining uncertainty in the common item ranking with nonparametric resampling of students, and then aggregate the twelve sitting-specific estimates using a random-effects model. The resulting parent mean AUC is 0.634, with a 95 percent interval of $[0.590, 0.679]$. The corresponding plug-in estimate, which treats the conditional-Rasch ranking as known, is 0.643. Table A11 reports the parent estimates and prediction intervals for both measures. Students therefore sort positively on average within the penalty regime: an attempted question is substantially more likely than

confidence in one ordering direction is 0.985; 68 percent of pairs have confidence of at least 0.90, and the median exceeds 0.93 in every sitting. As a consequence, posterior-mean Bayesian ranking is also nearly identical to the plug-in conditional-Rasch ranking, with an equal-sitting mean Spearman correlation of 0.995.

[36] The corresponding Kendall correlations are 0.765 and 0.778, respectively.

Figure 8: Within-Regime Sorting by Item Difficulty



(a) Student-level soft AUC

(b) Sorting gains by AUC quintile

Notes: Panel (a) shows the distribution of posterior-mean soft AUC among students who attempt and omit at least one question in the penalty regime. The dashed vertical line marks random sorting, AUC = 0.5. Panel (b) reports mean individual sorting gains by within-sitting quintile of soft AUC among students whose number of omissions rises by at least one question between Parts 1 and 2. Sitting-specific means receive equal weight. Vertical bars report 95 percent intervals from a nonparametric student bootstrap within sitting-by-quintile cells. The horizontal dashed line marks zero sorting gain.

chance to be easier than an omitted question. Sorting is nevertheless far from sharp. The student-level distribution is wide, and 23.1 percent of students have a posterior-mean soft AUC at or below 0.5.

Panel (b) relates the within-regime measure to the experimental measure of selection quality. I form soft-AUC quintiles within each sitting and compute the mean between-regime sorting gain in each quintile. The relationship is monotone and large: mean sorting gains rise from -0.058 in the bottom AUC quintile, with a 95 percent interval of $[-0.108, -0.007]$, to 0.338 in the top quintile, with an interval of $[0.287, 0.393]$. The intermediate quintiles rise steadily as well. Table A12 reports the full quintile pattern and sample sizes. Students whose Part 2 choices more effectively separate easy from hard questions are therefore also those who convert penalty-induced omissions into larger gains in the quality of attempted answers. Requiring an increase of at least two omissions produces the same pattern.

The level of the soft AUC is identified entirely from choices within the penalty regime and is therefore unaffected by systematic differences between the experimental blocks. Its close relationship with the sorting-gain measure provides evidence that the within- and between-regime exercises capture the same behavioral margin. The within-regime evidence therefore points to the same conclusion as the experimental comparison.

6.2 Gender Differences in Self-Reported Confidence

The structural model infers how knowledge maps into confidence from students' test-taking choices. I next ask whether the same gender pattern—similar metacognitive sensitivity but a substantial difference in confidence bias—appears when confidence is observed

directly. I use the data from Bordalo et al. (2019), where participants take a multiple-choice quiz including different topics and report, for each question, the probability that their answer is correct.[37] This experiment therefore complements that of Iriberri and Rey-Biel (2021): on the one hand, confidence is directly observed; on the other, participants do not choose which questions to answer. I consequently construct observable analogues of the structural quantities estimated above.

Type-2 AUC I measure metacognitive sensitivity using the participant's *Type-2 AUC*, a standard measure in psychology and neuroscience (Fleming and Lau, 2014). Let $C_{ij} \in [0,1]$ denote participant i's reported confidence on question j, and let $Y_{ij} \in \{0,1\}$ indicate whether the answer is correct:

$$\text{AUC}_i^{II} = \mathbb{P}(C_i^1 > C_i^0) + \frac{1}{2}\mathbb{P}(C_i^1 = C_i^0),$$

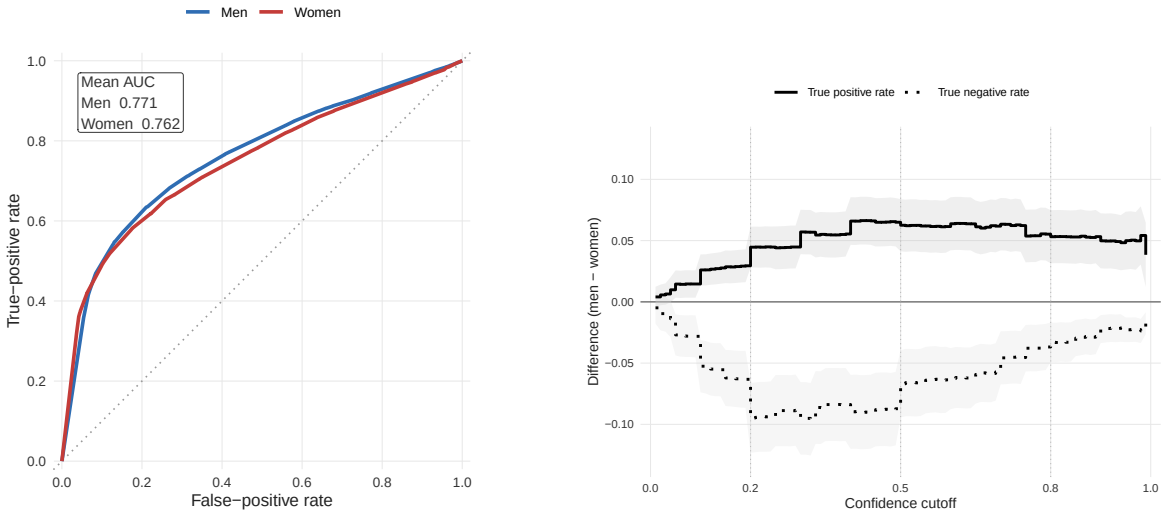
where C_i^1 is confidence on a randomly selected correct answer and C_i^0 confidence on a randomly selected incorrect answer. Type-2 AUC is thus the probability that a participant assigns higher confidence to a correct than to an incorrect answer, with half weight on ties. Equivalently, it is the area under the Type-2 ROC curve obtained by varying a confidence threshold and tracing the rate at which correct and incorrect responses exceed it. In this sense, Type-2 AUC measures how well confidence discriminates successful from unsuccessful performance and is closely related to the structural notion of metacognitive sensitivity in Section 4. Importantly, because it depends on the ranking of confidence rather than its overall level, it is invariant to a common shift in confidence bias.

Positive and Negative Metacognition Curves To capture the directional component of metacognition, I introduce a *virtual* confidence threshold τ. For each participant, define

$$\widetilde{\psi}_i^+(\tau) = \mathbb{P}(C_{ij} > \tau \mid Y_{ij} = 1, i), \qquad \widetilde{\psi}_i^-(\tau) = \mathbb{P}(C_{ij} \leq \tau \mid Y_{ij} = 0, i).$$

If confidence above τ were treated as a decision to answer and confidence below τ as a decision to omit, these objects would correspond to positive and negative metacognition in the structural model, up to conditioning on realized success and failure rather than latent knowledge. Varying τ over the full confidence scale therefore traces observable analogues of the two structural objects without taking a stance on any particular cutoff. Therefore, the area distance between the two curves can be taken as the analogue of bias. Indeed, a general upward shift in confidence raises $\widetilde{\psi}_i^+(\tau)$ and lowers $\widetilde{\psi}_i^-(\tau)$, even if the ranking of correct and incorrect answers —and hence Type-2 AUC—is unchanged.

[37] The dataset includes 45,278 question-level observations from 1,052 participants.



(a) Metacognitive sensitivity

(b) Directional metacognition

Figure 9: Gender Differences in Reported Confidence

Notes: Data are from Bordalo et al. (2019). Panel A reports gender-specific mean individual Type-2 ROC curves. Type-2 AUC is the probability that a correct response receives higher reported confidence than an incorrect response, with half weight on ties. Panel B reports male-minus-female differences in the positive analogue $\mathbb{P}(C > \tau \mid Y = 1)$ and the negative analogue $\mathbb{P}(C \leq \tau \mid Y = 0)$ over virtual confidence thresholds τ. Bands report pointwise 95 percent confidence intervals based on cross-participant variation.

Results Figure 9 reports the results, while Table A14 reports the corresponding estimates. The left panel shows little gender difference in metacognitive sensitivity. The two Type-2 ROC curves are close, and average Type-2 AUC is 0.771 for male participants and 0.762 for female participants. The difference is 0.010 and is not statistically significant. Confidence therefore discriminates correct from incorrect answers similarly well across genders.

The right panel reveals a much larger difference in the *location* of confidence. It plots male-minus-female differences in $\widetilde{\psi}^{+}(\tau)$ and $\widetilde{\psi}^{-}(\tau)$ as the virtual threshold varies. Across most of the confidence scale, the positive-metacognition gap is above zero while the negative-metacognition gap is below zero. Thus, at the same confidence threshold, male participants are more likely to place a correct answer above the threshold, but less likely to place an incorrect answer below it. At $\tau = 0.5$, for example, the male–female difference is +6.3 percentage points for $\widetilde{\psi}^{+}(\tau)$ and −6.7 points for $\widetilde{\psi}^{-}(\tau)$. This opposite-signed movement is what one would expect from a higher overall location of male confidence: it makes high-confidence classifications more frequent for both correct and incorrect answers, improving the positive metacognition component while worsening the negative one.[38]

Overall, even if the empirical objects are not exact counterparts of the structural

[38]The distinction can also be seen directly in average calibration. Reported confidence exceeds realized correctness by 13.7 percentage points for men and 10.4 points for women, a male–female difference of 3.4 points. This gap is essentially unchanged after accounting for the questions participants face and for individual performance: it is 3.1 points with question fixed effects and 3.4 points when I additionally control for leave-one-question-out ability. The directional pattern is similarly robust; at $\tau = 0.5$, the fully adjusted male–female differences remain +5.0 percentage points for the positive analogue and −5.3 points for the negative analogue. Table A15 reports the full set of specifications.

parameters, they provide external evidence that the patterns in sensitivity and bias that emerge from structural estimation replicate at the level of self-reported confidence in a different experimental context.

6.3 External Validity in a High-Stakes Test

The field experiment provides clean within-student variation in incentives, but the stakes are modest relative to standardized tests that determine access to educational opportunities. I therefore ask whether the same gender-by-ability pattern documented above also appears in a high-stakes setting. I use the Chilean University Selection Test (*Prueba de Selección Universitaria*, PSU), the national college-entry examination studied by Coffman and Klinowski (2020).

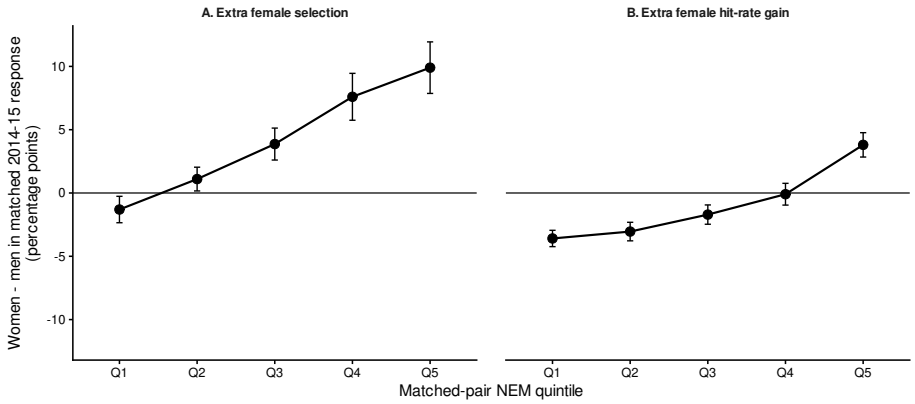
Until 2014, incorrect answers were penalized while omissions received zero points; beginning in 2015, the penalty was removed. Coffman and Klinowski (2020) show that the reform sharply reduced the gender gap in omissions, particularly among high-performing students. The PSU therefore provides a useful external-validity setting for my main results, although the two scoring regimes are observed across adjacent cohorts rather than within the same student.

Matching Design To make students observed under the two regimes as comparable as possible, I construct a matched synthetic panel around the 2015 reform. I restrict the sample to current high-school graduates taking the PSU for the first time. Within cells defined by gender, high school, age, and employment status, I match each 2014 mathematics-test taker, whenever a suitable match is available, to the nearest 2015 student on high-school GPA (*NEM*), school rank, parental education, and household income. I match with replacement and impose calipers on NEM and school rank. The procedure yields 93,213 matched pairs. Post-matching balance is good: the largest absolute standardized difference across the non-exact matching covariates is 0.025. Table A16 reports matching support and balance.

Identification of the reform effects requires conditional comparability across adjacent cohorts: after matching on predetermined student and school characteristics, comparable students taking the exam in 2015 provide the counterfactual behavior of 2014 students had they faced the no-penalty regime. To provide evidence on this identifying assumption, I report in Table A18 placebo comparisons for years before and after the scoring reform. The reform-year gender difference in omissions is substantially larger than the corresponding same-regime contrasts.

Hence, for each matched pair, I estimate the change between the penalty and no-penalty regimes in two outcomes: the share of questions omitted and the hit rate among attempted questions. I then compare these changes across gender. To account for residual differences in the matching covariates, I apply a regression bias correction (Abadie and Imbens, 2011)

Figure 10: Gender Differences in the Quantity and Quality Response by Prior Achievement



Notes: The figure reports female-minus-male differences in bias-corrected matched responses between the 2014 penalty regime and the 2015 no-penalty regime of the mathematics PSU, separately by matched-pair NEM quintile. Extra female selection is the gender difference in the change in the percentage of questions omitted; extra female hit-rate gain is the gender difference in the change in correctness conditional on attempting a question. NEM is determined before the PSU. Students are matched exactly on gender, high school, age, and employment status and by nearest neighbor on NEM, school rank, parental education, and household income, with replacement and calipers on NEM and school rank. Intervals are 95 percent confidence intervals with standard errors clustered by high school.

and cluster inference at the high-school level.

Results On average, the PSU reproduces the quantity–quality wedge documented in the field experiment. Relative to male students, female students increase omissions by an additional 2.6 percentage points under the penalty regime, while their relative hit-rate gain is −0.65 percentage points. Figure A11 reports these average responses.

The closer analogue to the gender–ability puzzle in the experiment is shown in Figure 10. I measure ability using matched-pair NEM, which is determined before the PSU and therefore unaffected by the scoring regime. Apart from replacing within-student comparisons with matched adjacent cohorts, the figure mirrors the earlier exercise by plotting female-minus-male quantity and quality responses across the ability distribution. The gender difference in omissions rises sharply with achievement, from −1.3 percentage points in the bottom NEM quintile to 9.9 points in the top quintile, confirming the gradient documented by Coffman and Klinowski (2020).

Importantly, the quantity–quality wedge is also visible through most of the NEM distribution. In quintiles 2 through 4, the additional female omission response rises from 1.1 to 7.6 percentage points, while the relative hit-rate response remains negative or essentially zero, moving only from −3.0 to −0.1 points. The main difference from the field experiment is at the top of the ability distribution: female hit-rate gains also increase with NEM and become positive in the top quintile, reaching 3.8 percentage points, although still far less than the corresponding 9.9-point omission response. Table A17 reports the full estimates.

Overall, the PSU reproduces both parts of the gender–ability pattern over most of the achievement distribution: higher-achieving women respond increasingly strongly on the

quantity margin, without a commensurate improvement in the quality of retained answers. The main departure is among the highest-achieving students, for whom the female quality response also becomes positive, though it remains substantially smaller than the selection response.

7 Conclusion

This paper argues that standardized tests may measure more than knowledge. When answering is costly, scores also depend on metacognition: whether students know what they know and what they do not know. I develop a framework that makes this dimension testable from choice data and show that students depart substantially from efficient question selection. Penalties induce large increases in omissions but much smaller gains in accuracy; many students sort questions only weakly, and some sort adversely. Structural estimates point to an imperfect mapping from knowledge into confidence and reveal meaningful gender differences in its shape. These differences affect both the interpretation of test scores and who benefits from alternative scoring rules.

More broadly, this paper introduces imperfect metacognition as a concept relevant for economics. Whereas standardized tests provide a particularly clean illustration of the paper's logic, the underlying mechanism extends beyond educational assessment. Imperfect metacognition may matter whenever agents choose which tasks to undertake, where to allocate effort, or what to delegate. In such settings, outcomes depend not only on ability, but also on how accurately agents perceive where their abilities lie. Understanding the economic consequences of this distinction defines a wider agenda for future work.

References

Abadie, A. and G. W. Imbens (2011). Bias-corrected matching estimators for average treatment effects. *Journal of business & economic statistics 29*(1), 1–11.

Abdulkadiroğlu, A., J. Angrist, and P. Pathak (2014). The elite illusion: Achievement effects at boston and new york exam schools. *Econometrica 82*(1), 137–196.

Ackerman, P. L. and R. Kanfer (2009). Test length and cognitive fatigue: an empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied 15*(2), 163.

Ackerman, R. and V. A. Thompson (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in cognitive sciences 21*(8), 607–617.

Aina, C. and F. Schneider (2025). Weighting competing models.

Akyol, Ş. P., J. Key, and K. Krishna (2016). Hit or miss? test taking behavior in multiple choice exams. Technical report, National Bureau of Economic Research.

Anelli, M. (2020). The returns to elite university education: A quasi-experimental analysis. *Journal of the European Economic Association 18*(6), 2824–2868.

Arenas, A. and C. Calsamiglia (2025). Gender differences in high-stakes performance and college admission policies. *Management Science 71*(10), 8413–8429.

Ash, E., S. Hansen, Y. Muvdi, and C. Marangon (2025). Large language models in economics. In *The Palgrave Handbook of Economics and Language*, pp. 191–210. Springer.

Augenblick, N., M. Backus, A. T. Little, and D. A. Moore (2025). Assumptions, disagreement, and overprecision: Theory and evidence. Technical report, Mimeo.

Augenblick, N., E. Lazarus, and M. Thaler (2024). Overinference from weak signals and underinference from strong signals. *Available at SSRN: https://ssrn.com/abstract=4315007*.

Augusto, L. M. (2010). Unconscious knowledge: A survey. *Advances in Cognitive Psychology 6*, 116.

Aumann, R. J. (1976). Agreeing to disagree. *The Annnals of Statistics*, 1236–1239.

Azmat, G., C. Calsamiglia, and N. Iriberri (2016). Gender differences in response to big stakes. *Journal of the European Economic Association 14*(6), 1372–1400.

Bagnoli, M. and T. Bergstrom (2005). Log-concave probability and its applications. *Economic theory 26*(2), 445–469.

Baldiga, K. (2014). Gender differences in willingness to guess. *Management Science 60*(2), 434–448.

Barber, B. M. and T. Odean (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *The quarterly journal of economics 116*(1), 261–292.

Bernheim, B. D., A. Lucia, K. Nielsen, and C. D. Sprenger (2026). When are decisions improvable? an evaluation of diagnostic methods. Technical report, National Bureau of Economic Research.

Bettinger, E. P., B. J. Evans, and D. G. Pope (2013). Improving college performance and retention the easy way: Unpacking the act exam. *American Economic Journal: Economic Policy 5*(2), 26–52.

Blinder, A. S. (1973). Wage discrimination: reduced form and structural estimates. *Journal of Human resources*, 436–455.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2019). Beliefs about gender. *American Economic Review 109*(3), 739–773.

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford, United Kingdom: Oxford University Press.

Bruhn, J., M. Gilraine, J. Ludwig, and S. Mullainathan (2025). Do test scores misrepresent test results? an item-by-item analysis. Technical report, National Bureau of Economic Research.

Bui, S. A., S. G. Craig, and S. A. Imberman (2014). Is gifted education a bright idea? assessing the impact of gifted and talented programs on students. *American Economic Journal: Economic Policy 6*(3), 30–62.

Camerer, C. and D. Lovallo (1999). Overconfidence and excess entry: An experimental approach. *American economic review 89*(1), 306–318.

Caplin, A., D. J. Deming, S. Leth-Petersen, and B. Weidmann (2023). Allocative skill. Technical report, National Bureau of Economic Research.

Chan, D. C., M. Gentzkow, and C. Yu (2022). Selection with variation in diagnostic skill: Evidence from radiologists. *The Quarterly Journal of Economics 137*(2), 729–783.

Chandra, A. and D. O. Staiger (2020). Identifying sources of inefficiency in healthcare. *The quarterly journal of economics 135*(2), 785–843.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014a). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American economic review 104*(9), 2593–2632.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014b). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American economic review 104*(9), 2633–2679.

Coffman, K. B. and D. Klinowski (2020). The impact of penalties for wrong answers on the gender gap in test scores. *Proceedings of the National Academy of Sciences 117*(16), 8794–8803.

Conde-Ruiz, J. I., J. J. Ganuza, and M. García (2020). Gender gap and multiple choice exams in public selection processes. *Hacienda Publica Espanola* (235), 11–28.

Currie, J. and W. B. MacLeod (2017). Diagnosing expertise: Human capital, decision making, and performance among physicians. *Journal of labor economics 35*(1), 1–43.

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

Deming, D. J. (2021). The growing importance of decision-making on the job. Technical report, National Bureau of Economic Research.

Deming, D. J. (2022). Four facts about human capital. *Journal of Economic Perspectives 36*(3), 75–102.

Deming, D. J. and M. I. Silliman (2024). Skills and human capital in the labor market. Technical report, National Bureau of Economic Research.

Díez-Rituerto, M., J. Gardeazabal, N. Iriberri, and P. Rey Biel (2024). Gender differences in willingness to guess revisited: Heterogeneity in a high stakes professional setting. Technical report, CEPR Discussion Papers.

Díez-Rituerto, M., J. Gardeazabal, N. Iriberri, and P. Rey-Biel (2025). Gender gaps in access to medical intern positions: The role of competition. *Journal of Labor Economics*. Accepted for publication.

Duckworth, A. L., C. Peterson, M. D. Matthews, and D. R. Kelly (2007). Grit: perseverance and passion for long-term goals. *Journal of personality and social psychology 92*(6), 1087.

Duflo, E., P. Dupas, and M. Kremer (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *American economic review 101*(5), 1739–1774.

Dvoretzky, A., J. Kiefer, and J. Wolfowitz (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 642–669.

Ebenstein, A., V. Lavy, and S. Roth (2016). The long-run economic consequences of high-stakes examinations: Evidence from transitory variation in pollution. *American Economic Journal: Applied Economics 8*(4), 36–65.

Ellis, R. S. (2012). *Entropy, large deviations, and statistical mechanics*. Springer Science & Business Media.

Enke, B. (2024). The cognitive turn in behavioral economics. Technical report, Mimeo Harvard.

Enke, B. and T. Graeber (2023). Cognitive uncertainty. *The Quarterly Journal of Economics 138*(4), 2021–2067.

Enke, B. and T. Graeber (2026). A note on when are decisions improvable? an evaluation of diagnostic methods.

Enke, B., T. Graeber, and R. Oprea (2023). Confidence, self-selection, and bias in the aggregate. *American Economic Review 113*(7), 1933–1966.

Erev, I., T. S. Wallsten, and D. V. Budescu (1994). Simultaneous over-and underconfidence: The role of error in judgment processes. *Psychological review 101*(3), 519.

Espinosa, M. P. and J. Gardeazabal (2020). The gender-bias effect of test scoring and framing: A concern for personnel selection and college admission. *The BE Journal of Economic Analysis & Policy 20*(3), 20190316.

Figlio, D. and U. Ozek (2024). The unintended consequences of test-based remediation. *American Economic Journal: Applied Economics 16*(1), 60–89.

Fleming, S. M. (2024). Metacognition and confidence: A review and synthesis. *Annual Review of Psychology 75*.

Fleming, S. M. and H. C. Lau (2014). How to measure metacognition. *Frontiers in human neuroscience 8*, 443.

Frydman, C. and L. J. Jin (2022). Efficient coding and risky choice. *The Quarterly Journal of Economics 137*(1), 161–213.

Funk, P. et al. (2023). Gender differences in academic performance the role of exam design in multiple-choice tests. Technical report.

Gabaix, X. (2019). Behavioral inattention. In *Handbook of behavioral economics: Applications and foundations 1*, Volume 2, pp. 261–343. Elsevier.

Galasso, V. and P. Profeta (2024). Gender differences in math tests: The role of time pressure. *The Economic Journal*.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Gonzalez, R. and G. Wu (1999). On the shape of the probability weighting function. *Cognitive psychology 38*(1), 129–166.

Griffin, D. and A. Tversky (1992). The weighing of evidence and the determinants of confidence. *Cognitive psychology 24*(3), 411–435.

Haggag, K., E. Leslie, D. Pope, and N. Pope (2024). The value of a higher act exam score. *Journal of Human Resources*.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, 153–161.

Heckman, J. J. and E. Vytlacil (2005). Structural equations, treatment effects, and econometric policy evaluation 1. *Econometrica 73*(3), 669–738.

Hoekstra, M. (2009). The effect of attending the flagship state university on earnings: A discontinuity-based approach. *The review of economics and statistics 91*(4), 717–724.

Holzinger, K. J. (1924). On scoring multiple response tests. *Journal of Educational Psychology 15*(7), 445–447.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Iriberri, N. and P. Rey-Biel (2019). Competitive pressure widens the gender gap in performance: Evidence from a two-stage competition in mathematics. *The Economic Journal 129*(620), 1863–1893.

Iriberri, N. and P. Rey-Biel (2021). Brave boys and play-it-safe girls: Gender differences in willingness to guess in a large scale natural field experiment. *European Economic Review 131*, 103603.

Jacob, B. and J. Rothstein (2016). The measurement of student ability in modern assessment systems. *Journal of Economic Perspectives 30*(3), 85–108.

Karle, H., D. Engelmann, and M. Peitz (2022). Student performance and loss aversion. *The Scandinavian Journal of Economics 124*(2), 420–456.

Khaw, M. W., Z. Li, and M. Woodford (2021). Cognitive imprecision and small-stakes risk aversion. *The review of economic studies 88*(4), 1979–2013.

Kline, P. and C. Walters (2021). Reasonable doubt: Experimental detection of job-level employment discrimination. *Econometrica 89*(2), 765–792.

Kline, P. and C. R. Walters (2019). On heckits, late, and numerical equivalence. *Econometrica 87*(2), 677–696.

Kripke, S. A. (1963). Semantical considerations on modal logic. *Acta Philosophica Fennica 16*, 83–94.

Krishna, K., P. Akyol, and E. Ozer (2025). Do time-constraints matter? how, why, and for whom? Technical report, National Bureau of Economic Research.

List, J. A. (2025). The experimentalist looks within: Toward an understanding of within-subject experimental designs. Technical report, National Bureau of Economic Research.

Macartney, H., R. McMillan, and U. Petronijevic (2018). *Teacher performance and accountability incentives*. National Bureau of Economic Research Cambridge, MA.

MacKinnon, J. G., M. Ø. Nielsen, and M. D. Webb (2023). Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics 232*(2), 272–299.

Malmendier, U. and G. Tate (2005). Ceo overconfidence and corporate investment. *The journal of finance 60*(6), 2661–2700.

Malmendier, U. and G. Tate (2008). Who makes acquisitions? ceo overconfidence and the market's reaction. *Journal of financial Economics 89*(1), 20–43.

Mazancieux, A., S. M. Fleming, C. Souchay, and C. J. Moulin (2020). Is there a g factor for metacognition? correlations in retrospective metacognitive sensitivity across tasks. *Journal of Experimental Psychology: General 149*(9), 1788.

McElreath, R. (2018). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.

Meager, R. (2019). Understanding the average impact of microcredit expansions: A bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics 11*(1), 57–91.

Meager, R. (2022). Aggregating distributional treatment effects: A bayesian hierarchical analysis of the microcredit literature. *American Economic Review 112*(6), 1818–1847.

Mislevy, R. J. and P.-K. Wu (1988). Inferring examinee ability when some item responses are missing. *ETS Research Report Series 1988*(2), i–75.

Mislevy, R. J. and P.-K. Wu (1996). Missing responses and irt ability estimation: Omits, choice, time limits, and adaptive testing. *ETS Research Report Series 1996*(2), i–36.

Moore, D. A. and P. J. Healy (2008). The trouble with overconfidence. *Psychological review 115*(2), 502.

Niederle, M. and L. Vesterlund (2007). Do women shy away from competition? do men compete too much? *The quarterly journal of economics 122*(3), 1067–1101.

Niederle, M. and L. Vesterlund (2010). Explaining the gender gap in math test scores: The role of competition. *Journal of economic perspectives 24*(2), 129–144.

Nielsen, E. (2019). Test questions, economic outcomes, and inequality.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International economic review*, 693–709.

Odean, T. (1998). Volume, volatility, price, and profit when all traders are above average. *The journal of finance 53*(6), 1887–1934.

Ortoleva, P. and E. Snowberg (2015). Overconfidence in political behavior. *American Economic Review 105*(2), 504–535.

Pekkarinen, T. (2015). Gender differences in behaviour under competitive pressure: Evidence on omission patterns in university entrance examinations. *Journal of Economic Behavior & Organization 115*, 94–110.

Polanyi, M. (1958). *Personal knowledge*. Routledge.

Prelec, D. (1998). The probability weighting function. *Econometrica*, 497–527.

Riener, G. and V. Wagner (2017). Shying away from demanding tasks? experimental evidence on gender differences in answering multiple-choice questions. *Economics of Education Review 59*, 43–62.

Rockoff, J. E., D. O. Staiger, T. J. Kane, and E. S. Taylor (2012). Information and employee evaluation: Evidence from a randomized intervention in public schools. *American Economic Review 102*(7), 3184–3213.

Rouault, M., P. Dayan, and S. M. Fleming (2019). Forming global estimates of self-performance from local confidence. *Nature communications 10*(1), 1141.

Rouault, M., A. McWilliams, M. G. Allen, and S. M. Fleming (2018). Human metacognition across domains: insights from individual differences and neuroimaging. *Personality neuroscience 1*, e17.

Ruiz-Tagle, C. (2024). Performance gaps in high-stakes testing: the role of textual context.

Shaked, M. and J. G. Shanthikumar (2007). *Stochastic orders*. Springer.

Spinnewijn, J. (2015). Unemployed but optimistic: Optimal insurance design with biased beliefs. *Journal of the European Economic Association 13*(1), 130–167.

Thurstone, L. L. (1919). A scoring method for mental tests. *Psychological Bulletin 16*(7), 235.

Vytlacil, E. (2002). Independence, monotonicity, and latent index models: An equivalence result. *Econometrica 70*(1), 331–341.

Wainer, H. (2011). Uneducated guesses: Using evidence to uncover misguided education policies.

Walters, C. (2024). Empirical bayes methods in labor economics. In *Handbook of Labor Economics*, Volume 5, pp. 183–260. Elsevier.

Woodford, M. (2020). Modeling imprecision in perception, valuation, and choice. *Annual Review of Economics 12*, 579–601.

Zimmerman, S. D. (2019). Elite colleges and upward mobility to top jobs and top incomes. *American Economic Review 109*(1), 1–47.

A Microfoundation

This appendix shows how Assumption 1 can arise from a multitasking model under imperfect metacognition. The microfoundation has two parts. The first is a portfolio-choice problem, which explains when confidence-based selection is monotone and when it is described by a fixed threshold, taking the latent pair (K, C) as primitive. The second is a hierarchical Bayesian updating model, which explains how the latent pair (K, C) may arise from processing cognitive signals — modeling reasoning — and meta-information about the informativeness of those signals. Throughout, I fix a student-type and a question-type and suppress that dependence in the notation.

A.1 Portfolio Choice Problem

Fix a student facing questions $j = 1, \ldots, N$. Before choosing which questions to answer, the student observes a confidence vector

$$c = (c_1, \ldots, c_N) \in [0, 1]^N.$$

Conditional on c_j, answering question j is subjectively equivalent to a Bernoulli lottery

$$B_j \mid c_j \sim \text{Bernoulli}(c_j),$$

with $(B_j)_{j=1}^N$ conditionally independent. For a portfolio $A \subseteq \{1, \ldots, N\}$, let

$$B(A) := \sum_{j \in A} B_j$$

denote the student's subjective number of correct answers in the attempted set. The student chooses A to maximize

$$V(A; c) := U\Big(\mathcal{L}(B(A) \mid c)\Big) - \chi(|A|), \tag{12}$$

where U evaluates the subjective distribution of total correct answers and χ is a cardinality cost. The first result requires only that, conditional on answering the same number of questions, the student prefers more correct answers to fewer.

Assumption 6 (Portfolio monotonicity)**.** For any portfolios $A, A' \subseteq \{1, \ldots, N\}$,

$$|A| = |A'| \quad \text{and} \quad B(A') \succeq_{FOSD} B(A) \quad \Longrightarrow \quad V(A'; c) \geq V(A; c),$$

with strict inequality whenever the dominance is strict.

Note that Assumption 6 is minimal: it does not require expected utility, separability, or any particular attitude toward portfolio risk.

Proposition A5 (Set-monotone selection)**.** *Suppose Assumption 6 holds. Then every optimal portfolio* $A^*(c) \in \arg\max_A V(A;c)$ *is confidence-monotone,*

$$j \in A^*(c),\ \ell \notin A^*(c) \quad \Longrightarrow \quad c_j \geq c_\ell.$$

Proposition A5 implies that the student first ranks questions by confidence and then chooses how far down the ranking to go. The stopping point, however, may depend on the entire realized confidence vector c. Thus monotonicity alone delivers *sorting by confidence*, but not a fixed threshold. To obtain a cutoff that does not depend on the realized confidence vector, one needs an additional restriction ruling out portfolio interactions in the answer/skip margin.

Assumption 7 (Separability)**.** For every confidence vector c, every portfolio $A \subseteq \{1,\dots,N\}$, and every $j \notin A$,

$$\operatorname{sgn}\Big(V(A\cup\{j\};c) - V(A;c)\Big) = \operatorname{sgn}\Big(V(\{j\};(c_j)) - V(\varnothing;\varnothing)\Big).$$

Moreover, the one-question value $c \longmapsto V(\{1\};(c))$ is continuous on $[0,1]$.

Assumption 7 says that the sign of the marginal gain from answering question j depends only on its own confidence level, and does so exactly as in the one-question problem. In this sense, the binary comparison between answering and skipping is independent of the background portfolio. The continuity clause is only a regularity condition ensuring that this comparison can be represented by a scalar cutoff.

Proposition A6 (Fixed cutoff under background independence)**.** *Suppose Assumption 6 and Assumption 7 hold. Then there exists a cutoff* $\tau \in [0,1]$ *such that, under the tie-breaking convention that the student skips whenever answering does not strictly improve the objective, the induced decision rule is*

$$D = \mathbf{1}\{C > \tau\}.$$

Proposition A5 and Proposition A6 separate two distinct ideas. Monotonicity alone implies that the student chooses an upper set in the confidence ranking, which is equivalent to a realized, potentially portfolio-dependent cutoff. Background independence then sharpens this to a fixed cutoff. This distinction is useful because some results in the paper require only monotone selection (see Section B.3) whereas Assumption 1 uses the stronger fixed-threshold representation.

Examples and Nonexamples

Suppose skipping yields o, a wrong answer yields m, and a correct answer yields h, where

$$h > o \geq m.$$

For a portfolio A, let total score be

$$W(A) = w_0 + o(N - |A|) + m\Big(|A| - B(A)\Big) + hB(A),$$

where w_0 is a baseline score shift.

Risk-neutral preferences If the student maximizes expected score, then for any $j \notin A$,

$$V(A \cup \{j\}; c) - V(A; c) = (h - m)c_j - (o - m).$$

Hence Assumption 7 holds, and the one-question gain is

$$g(c) = (h - m)c - (o - m).$$

Therefore

$$\tau = \frac{o - m}{h - m}.$$

CARA preferences Suppose instead that the student maximizes expected utility with

$$u(w) = 1 - e^{-\alpha w}, \qquad \alpha > 0.$$

Then

$$\mathbb{E}[u(W(A)) \mid c] = 1 - e^{-\alpha(w_0 + No)} \prod_{r \in A} \Big(e^{\alpha(o-m)} \Big[1 - c_r + c_r e^{-\alpha(h-m)}\Big] \Big).$$

Hence, for any $j \notin A$,

$$\begin{aligned} &\mathbb{E}[u(W(A \cup \{j\})) \mid c] - \mathbb{E}[u(W(A)) \mid c] \\ &\quad = e^{-\alpha(w_0 + No)} \prod_{r \in A} \Big(e^{\alpha(o-m)}[1 - c_r + c_r e^{-\alpha(h-m)}] \Big) \cdot \Big(1 - e^{\alpha(o-m)}[1 - c_j + c_j e^{-\alpha(h-m)}] \Big). \end{aligned}$$

The first factor is strictly positive, so the sign of the gain depends only on c_j. Therefore Assumption 7 holds, and the one-question gain has the same sign as

$$g(c) := 1 - e^{\alpha(o-m)}\Big[1 - c + ce^{-\alpha(h-m)}\Big].$$

Solving $g(c) = 0$ yields

$$\tau = \frac{1 - e^{-\alpha(o-m)}}{1 - e^{-\alpha(h-m)}}.$$

CRRA preferences Now suppose the student maximizes

$$V(A; c) = \mathbb{E}[u(W(A)) \mid c],$$

where

$$u(w) = \frac{w^{1-\rho}}{1-\rho}, \qquad \rho > 0,\ \rho \neq 1,$$

is CRRA utility on $\mathbb{R}_{++}$. Assume $W(A) > 0$ almost surely for every feasible portfolio.

Fix a portfolio A and a question $j \notin A$. Write

$$\Delta_j(A; c) := V(A \cup \{j\}; c) - V(A; c)$$

for the marginal gain from adding question j to portfolio A. Since adding j changes total score by $h - o$ if the student answers correctly and by $m - o$ if she answers incorrectly, we have

$$\Delta_j(A; c) = \mathbb{E}[c_j\, u(W(A) + h - o) + (1 - c_j)\, u(W(A) + m - o) - u(W(A)) \mid c]\,.$$

Under Assumption 7, the sign of $\Delta_j(A; c)$ would have to depend only on c_j, and not on the portfolio A. CRRA violates this property.

Capacity constraints If the student may answer at most $\bar{n}$ questions, then optimal selection remains monotone by Proposition A5, but Assumption 7 typically fails. The value of answering a given question depends on how many answer slots remain available, so the cutoff is realized and portfolio-dependent rather than fixed.

A.2 Hierarchical Bayesian Updating

I now turn to the informational microfoundation for the latent pair (K, C).

Fix a single match. Let $d_0 \in \Delta(\Omega)$ be a prior over a finite state space Ω of possible correct answers, and let $\lambda_0 \in \Delta(T)$ be a prior over a finite or Borel type space T. I interpret $\omega \in \Omega$ as the true answer and $t \in T$ as the latent informativeness type of the student's first-order reasoning process. For simplicity, assume that ω and t are independent under the prior.

The student does not observe ω directly. Instead, she reasons through an internal *first-order* signal

$$X \in \mathcal{X},$$

generated by a signal structure whose informativeness depends on t. Conditional on (ω, t), the first-order signal is generated by

$$X \sim \xi(\cdot \mid \omega, t).$$

Crucially, the student also receives a *meta-signal*

$$M \in \mathcal{M},$$

generated according to

$$M \sim \mu(\cdot \mid t).$$

The meta-signal conveys information about the latent type t, and hence about the informativeness of the first-order signal structure $\xi(\cdot \mid \omega, t)$. In that sense, the student is uncertain not only about the correct answer, but also about the quality of her own reasoning process. This meta-information is the source of imperfect metacognition in the model.

The student observes (X, M), but not t. Let

$$a(X, M) \in \Omega$$

denote the answer she would give if forced to answer the question. Under standard scoring rules, $a(X, M)$ is simply the posterior-optimal answer conditional on (X, M), with ties broken arbitrarily. Define the latent success indicator

$$Y^* := \mathbf{1}\{a(X, M) = \omega\}.$$

Thus $Y^* = 1$ iff the student would answer correctly if forced to attempt the question.

The two latent indices in the main text are now defined as conditional success probabilities computed at two different informational resolutions:

$$K := \mathbb{P}(Y^* = 1 \mid t, X, M), \tag{13}$$

$$C := \mathbb{P}(Y^* = 1 \mid X, M). \tag{14}$$

The index K is the objective probability of success on the match: it conditions on the realized type t of the first-order signal structure. The index C is subjective confidence: it averages over the remaining uncertainty about that type.

Equivalently, K is pinned down by the realized experiment, whereas C is pinned down by the posterior-predictive experiment induced by the information the student actually observes. If the meta-signal reveals t perfectly, then $C = K$ and metacognition is perfect. If the meta-signal is noisy, then the student only partially knows how informative her own reasoning is, and C becomes a noisy compression of K.

The next proposition states the reduced-form implication of this Bayesian model.

Proposition A7 (Martingale restriction)**.** *The pair* (K, C) *defined in Equation* (13)*–Equation* (14) *satisfies:*

(i) $Y^* \mid (K, C) \sim \text{Bernoulli}(K)$;

(ii) $C = \mathbb{E}[K \mid X, M]$ *and therefore* $\mathbb{E}[K \mid C] = C$.

Proposition A7 is the key bridge from the Bayesian model to the reduced form. Confidence cannot be arbitrarily related to knowledge. Rather, confidence must be a mean-preserving

contraction of knowledge: once one conditions on the student's confidence level, average objective performance must coincide with that confidence level. The next result shows that, under a natural symmetry condition, this martingale restriction is also essentially sufficient.

Proposition A8 (Converse implementation under symmetric experiments)**.** *Suppose the prior d_0 on Ω is uniform. Let $(\widetilde{K}, \widetilde{C})$ be any pair of random variables such that*

$$\widetilde{K} \in \left[\frac{1}{|\Omega|}, 1\right], \qquad \mathbb{E}[\widetilde{K} \mid \widetilde{C}] = \widetilde{C}.$$

Then there exists a Bayesian information structure with latent type t, first-order signal X, signal family $\xi(\cdot \mid \omega, t)$, and meta-signal M such that the induced pair (K, C) has the same distribution as $(\widetilde{K}, \widetilde{C})$.

Binary example: positive and negative metacognition

The role of meta-information is especially transparent in a binary version of the model, which also connects directly to the notions of positive and negative metacognition.

Suppose $\Omega = \{0, 1\}$, the prior on ω is uniform, and the latent type takes two values,

$$t \in \{L, H\}, \qquad \mathbb{P}(t = H) = \pi \in (0, 1).$$

Conditional on t, the first-order signal $X \in \{0, 1\}$ is generated by a binary symmetric signal structure:

$$\mathbb{P}(X = \omega \mid \omega, t = H) = \overline{k}, \qquad \mathbb{P}(X = \omega \mid \omega, t = L) = \underline{k},$$

where

$$\frac{1}{2} \leq \underline{k} < \overline{k} \leq 1.$$

Thus H and L correspond to high- and low-knowledge states, and under forced answering the student follows the signal X. The induced knowledge index therefore takes two values:

$$K = \begin{cases} \overline{k}, & t = H, \\ \underline{k}, & t = L. \end{cases}$$

Now let the meta-signal also be binary,

$$M \in \{\underline{m}, \overline{m}\},$$

where $\overline{m}$ corresponds to high felt confidence and $\underline{m}$ to low felt confidence. Define

$$\psi^+ := \mathbb{P}(M = \overline{m} \mid t = H), \qquad \psi^- := \mathbb{P}(M = \underline{m} \mid t = L).$$

The parameter ψ^+ is the probability of feeling high confidence when the underlying

knowledge state is high. It captures the student's ability to *know what she knows*, so it is natural to interpret it as *positive metacognition*. The parameter ψ^- is the probability of feeling low confidence when the underlying knowledge state is low. It captures the student's ability to *know what she does not know*, so it is natural to interpret it as *negative metacognition*.

By symmetry, the first-order signal X is uninformative about t, so confidence depends only on the meta-signal:

$$C = \mathbb{E}[K \mid M].$$

Thus the two confidence levels are

$$\overline{c} := \mathbb{E}[K \mid M = \overline{m}] = \frac{\pi\psi^+\overline{k} + (1-\pi)(1-\psi^-)\underline{k}}{\pi\psi^+ + (1-\pi)(1-\psi^-)}, \tag{15}$$

$$\underline{c} := \mathbb{E}[K \mid M = \underline{m}] = \frac{\pi(1-\psi^+)\overline{k} + (1-\pi)\psi^-\underline{k}}{\pi(1-\psi^+) + (1-\pi)\psi^-}. \tag{16}$$

These expressions make the role of positive and negative metacognition transparent. Increasing ψ^+ makes the high-confidence state more likely to come from high knowledge and the low-confidence state less likely to do so. Increasing ψ^- makes the low-confidence state more likely to come from low knowledge and the high-confidence state less likely to do so. In both cases, the wedge between $\overline{c}$ and $\underline{c}$ widens. Thus positive and negative metacognition are two distinct ways of making confidence more informative about underlying knowledge.

Several benchmark cases are immediate:

- If $\psi^+ = \psi^- = 1$, then the meta-signal reveals the knowledge state perfectly and $C = K$. Metacognition is perfect.
- If $\psi^+ + \psi^- = 1$, then the meta-signal is uninformative about the knowledge state, so

  $$\overline{c} = \underline{c} = \mathbb{E}[K] = \pi\overline{k} + (1-\pi)\underline{k}.$$

  Confidence carries no metacognitive content.
- If $\psi^+ + \psi^- > 1$, then high confidence is positively associated with high knowledge and low confidence with low knowledge. Selection based on confidence is informative in the correct direction.
- If $\psi^+ + \psi^- < 1$, then the association is reversed: high confidence is, on average, attached to lower knowledge states. Selection based on confidence is adversely informative.

A.3 Microfoundation of Assumption 1

I can now combine the choice and information sides. For each match, define K and C as in Equation (13)–Equation (14). Suppose, in addition, that the student's answer decision satisfies the fixed-cutoff representation from Proposition A6. Then

$$D = \mathbf{1}\{C > \tau\}.$$

Together with Proposition A7(i), this yields

$$Y^* \mid (K, C) \sim \text{Bernoulli}(K), \qquad D = \mathbf{1}\{C > \tau\}.$$

Restoring the suppressed dependence on types yields exactly the structural assumption in the main text. The latent pair (K, C) may depend on both student and question-type through the signal structure and the meta-signal, whereas the cutoff depends only on student-type because, on the choice side, question heterogeneity has already been absorbed into confidence:

$$Y^* \mid (K, C, S, Q) \sim \text{Bernoulli}(K), \qquad D = \mathbf{1}\{C > \tau(S)\}.$$

This microfoundation also clarifies the interpretation of the joint law of (K, C). The marginal law of K captures how informative the student's realized reasoning is across matches — that is, competence. The conditional law of $C \mid K$ captures how well the student knows how informative her own reasoning is — that is, metacognition.

B Testing Metacognition

I present some additional results on testing metacognitive quality from choice data, complementing Section 2.

B.1 Quantitative Bounds on Sorting Statistics

Proposition 1 and Proposition 4 establish sharp qualitative implications: under Rank-Preserving Metacognition, the relevant sorting statistics must satisfy specific sign restrictions. At the same time, RPM is intentionally permissive. In particular, weak RPM allows environments in which confidence and knowledge are positively related, but only weakly so. This raises a natural question: can stronger restrictions on the stochastic mapping from knowledge to confidence deliver sharper bounds on the same statistics? I now show that they can.

The key observation is that, by Definition 1, RPM is equivalent to supermodularity of the conditional log-survival function of confidence. A natural quantitative strengthening is therefore to require this object to be supermodular up to a bilinear term, with strength indexed by a parameter κ.

Definition 2 (κ-metacognition)**.** Fix $\kappa \geq 0$. I say that type s is *κ-metacognitive* if, for every $c_2 > c_1$ and $k_2 > k_1$ such that the four terms below are strictly positive,

$$\overline{F}_s(c_2 \mid k_2)\,\overline{F}_s(c_1 \mid k_1) \geq \exp\Big(\kappa(c_2 - c_1)(k_2 - k_1)\Big)\left(\overline{F}_s(c_2 \mid k_1)\,\overline{F}_s(c_1 \mid k_2)\right). \tag{17}$$

When $\kappa = 0$, this reduces to the qualitative RPM restriction in the main text. The next two results refine the qualitative tests and nest them as the special case $\kappa = 0$.

Proposition A9 (Quantitative between-regime test)**.** *Suppose Assumption 1–Assumption 3 hold. If students of type s are κ-metacognitive, then*

$$\Delta O_s > 0 \Longrightarrow \Delta H_s \geq \mathbb{V}(K \mid D(0) = 1, S = s)\,\frac{1 - e^{-2\kappa\delta_s}}{2}, \tag{18}$$

where $\delta_s := \tau(s,1) - \tau(s,0)$.

Equivalently, whenever $1 > O_{s,1} > O_{s,0}$, the sorting gain satisfies

$$\Gamma_s \geq \frac{1 - O_{s,0}}{O_{s,1} - O_{s,0}}\,\mathbb{V}(K \mid D(0) = 1, S = s)\,\frac{1 - e^{-2\kappa\delta_s}}{2}.$$

This lower bound highlights the same complementarity between knowledge dispersion and metacognitive quality that appears in Section 4 and is further developed through the notion of centrifugal spread in Section B.2. Sorting gains are larger when the student faces more

within-test variation in comparative advantage, and κ governs how strongly that variation is converted into performance gains under tighter selection.

The proof relies on an exponential-tilting argument, similar to those used in large deviations theory (for background, see Ellis, 2012). Under stronger κ-metacognition, tightening the regime shifts more weight toward higher-knowledge questions. The high-cost regime can therefore be viewed as an exponentially tilted version of the low-cost regime, with more mass on questions the student is better equipped to answer. The bound quantifies how much average performance must rise under that tilt, given the baseline dispersion of knowledge.

I next turn to a quantitative refinement of Proposition 4. For this result, I impose two additional regularity conditions.

Assumption 8. The following technical conditions hold:

(i) for each type s, the proxy distribution $R \mid S = s$ is continuous;

(ii) attempt probabilities are uniformly bounded away from zero across knowledge types, so that

$$\underline{a}_s := \inf_k \mathbb{P}(D = 1 \mid K = k, S = s) > 0.$$

Now let $U_s := F_{R|S=s}(R)$ denote the within-type rank transform of the proxy, so that $U_s \mid S = s \sim \mathrm{U}[0, 1]$. I say that the proxy is ℓ_s-strong for type s if there exists $\ell_s \geq 0$ such that the map

$$u \longmapsto \mathbb{E}[K \mid U_s = u, S = s] - \ell_s u$$

is weakly increasing. Suppose also that type s is κ-metacognitive.

Proposition A10 (Quantitative within-regime bound)**.** *Suppose Assumption 1, Assumption 5, and Assumption 8 hold, and let* $p_s := \mathbb{P}(D = 1 \mid S = s) \in (0, 1)$*. If student-type* s *is* κ*-metacognitive, then*

$$\mathrm{AUC}_s - \frac{1}{2} \geq \frac{\underline{a}_s \, \ell_s \, \tau_s}{12 \, p_s(1 - p_s)} \, \kappa. \tag{19}$$

The proof of Proposition A10 quantitatively refines the argument behind Proposition 4. The key observation is, once again, that κ-metacognition imposes a lower bound on how fast attempt probability may rise with latent knowledge. Proxy strength then transfers that slope from latent knowledge to the observable proxy rank. As a result, higher-ranked questions are more likely by an amount bounded away from zero. Combining this quantitative selection gradient with the covariance logic behind Lemma 1 yields the AUC bound in Equation (19).

B.2 An Identification Result

The tests in Section 2 derive observable restrictions from structural assumptions on the stochastic map from knowledge to confidence. I now turn to the converse question:

how much of that map can be recovered from sufficiently rich variation in the selection environment? The answer is that, if one could observe omit and hit-rates over a continuum of regimes, then the resulting *hit-omit curve* would identify progressively stronger monotonicity properties of the map $K \mapsto C$.

To make this precise, suppose that the binary regime in the main text is replaced by a continuously varying instrument $Z \in [0,1]$. Extend Assumption 2 and Assumption 3 in the natural way: conditional on (S,Q), the regime shifts only the answer threshold and leaves the joint distribution of (K,C) unchanged, while the composition of question-types remains balanced within student-type. Normalizing the regime within type s, one can therefore write the student decision rule as

$$D(z) = \mathbf{1}\{C > z\}, \qquad z \in [0,1].$$

Thus, z continuously indexes how selective the environment is: higher values of z induce more omission.

Fix now a student-type s and a distribution $\mu \in \Delta([0,1])$ of latent knowledge K for that type. This is the natural analogue of holding fixed the underlying population of matches and asking how the observable statistics respond as the regime varies. For each threshold z, let $O_{\mu,s}(z)$ and $H_{\mu,s}(z)$ denote the omission-rate and hit-rate generated by the rule $D(z)$, exactly as in the main text. Also define

$$H^0_\mu := \mathbb{E}_\mu[K \mid S = s].$$

This is the benchmark level of performance absent any selective answering: it is the hit-rate that would obtain if the student were forced to answer every question, or, equivalently, under the idealized limiting regime in which selection is shut down.

I then define the *hit-omit curve* as

$$\Gamma_{\mu,s} := \Big\{ \big(O_{\mu,s}(z),\, H_{\mu,s}(z) - H^0_\mu\big) :\ z \in [0,1],\ \mathbb{P}_\mu(D(z) = 1 \mid S = s) > 0 \Big\},$$

and the associated *selection surplus* at threshold z as

$$\gamma_{\mu,s}(z) := \Big(1 - O_{\mu,s}(z)\Big)\Big(H_{\mu,s}(z) - H^0_\mu\Big).$$

The hit-omit curve records, for each threshold, the gain in average quality delivered by confidence-based selection relative to the nonselective benchmark H^0_μ. The surplus $\gamma_{\mu,s}(z)$ rescales that gain by the share of questions that remain selected.

To characterize what these objects identify, it is useful to distinguish three nested

monotonicity notions for the stochastic map $K \mapsto C$ within type s:

$$\begin{aligned}
&\textit{Expectation monotonicity}: && k_2 > k_1 \Longrightarrow \mathbb{E}[C \mid K = k_2, S = s] \geq \mathbb{E}[C \mid K = k_1, S = s],\\
&\textit{Stochastic monotonicity}: && k_2 > k_1 \Longrightarrow (C \mid K = k_2, S = s) \succeq_{FOSD} (C \mid K = k_1, S = s),\\
&\textit{Hazard-rate monotonicity}: && k_2 > k_1 \Longrightarrow (C \mid K = k_2, S = s) \succeq_{HRO} (C \mid K = k_1, S = s).
\end{aligned}$$

The third condition is exactly RPM.

Proposition A11 (Identification from the hit-omit curve)**.** *Fix a student-type s. Then:*

(i) $K \mapsto C$ is expectation monotone if and only if, for every $\mu \in \Delta([0,1])$, the signed area under the hit-omit curve is nonnegative:

$$\int_0^1 \gamma_{\mu,s}(z)\, dz \geq 0.$$

(ii) $K \mapsto C$ is stochastically monotone if and only if, for every $\mu \in \Delta([0,1])$, the hit-omit curve lies weakly above the horizontal axis.

(iii) $K \mapsto C$ is hazard-rate monotone if and only if, for every $\mu \in \Delta([0,1])$, the hit-omit curve lies weakly above the horizontal axis and is weakly increasing.

The result complements the main text by showing how the same observables recover increasingly strong structure once the threshold varies continuously. Expectation monotonicity is identified by an average property of the hit-omit curve: the signed area under it is nonnegative. Stochastic monotonicity sharpens this into a pointwise restriction: the curve must lie weakly above the horizontal axis at every threshold. Hazard-rate monotonicity sharpens it again into a shape restriction: as the threshold rises and selection becomes stricter, the hit-rate gain must itself rise weakly. Thus, the geometry of the hit-omit curve encodes the strength of the underlying confidence ranking.

This perspective also clarifies the relationship with the binary tests in the main text. A between-regime comparison considers only two thresholds, say $z_0 < z_1$. In the present language, that comparison is just a secant of the continuous hit-omit curve. If hazard-rate monotonicity holds, then $z \mapsto H_{\mu,s}(z)$ is increasing, so every such secant has nonnegative slope: whenever omission rises between z_0 and z_1, the hit-rate cannot fall. The RPM test in the main text is therefore the two-point implication of a global monotonicity property that becomes visible only under continuous threshold variation.

The proof uses the same sorting logic as the main results, but applies it at different levels of strength. In every case, the key object is the covariance between latent knowledge and threshold selection. For stochastic monotonicity, nonnegative hit-rate gains are equivalent to nonnegative covariance between K and the survival indicator $\mathbf{1}\{C > z\}$ at each threshold. Expectation monotonicity integrates this covariance over thresholds, collapsing the full

selection schedule into the covariance between K and $\mathbb{E}[C \mid K, S = s]$. Hazard-rate monotonicity, in turn, uses the same retention logic as the proof of the between-regime test: tightening the threshold from z_1 to z_2 defines a retention decision within the pool selected at z_1, and hazard-rate ordering implies that average latent knowledge must rise.

B.3 Sample-Level Tests

The main text studies population implications of the fixed-threshold representation

$$D = \mathbf{1}\{C > \tau(S)\}.$$

The multitasking microfoundation shows, however, that this representation is stronger than what preference monotonicity alone delivers, but also needs a separability assumption. In general, even if students rank questions by confidence, the realized stopping rule may depend on the entire confidence vector and hence vary with the realized confidence levels. Once the cutoff is unstable in this way, it becomes conceptually difficult to think about identification. For that reason, the relevant objects become sample ones: how omission and hit-rates behave in long tests and in large student samples. I next show how Proposition 1 and Proposition 4 extend to this case.

Throughout this subsection, I introduce explicit indexing for students ($i = 1, \ldots, M$) and questions ($j = 1, \ldots, N$), as I will take limits along these dimensions. I first assume the sample-level content of Proposition A5, replacing the fixed-threshold representation Assumption 1.ii.

Assumption 9 (Confidence-monotone selection)**.** For every student i and test length N, there exists a measurable map $\mathcal{A}_{i,N} : [0,1]^N \to 2^{\{1,\ldots,N\}}$ such that the attempted set in regime $z = 1$ satisfies $A_{i,N}(1) = \mathcal{A}_{i,N}\big(C_{i1}(1), \ldots, C_{iN}(1)\big)$ a.s., and, for every confidence vector $c \in [0,1]^N$,

$$j \in \mathcal{A}_{i,N}(c),\ k \notin \mathcal{A}_{i,N}(c) \quad \Longrightarrow \quad c_j \geq c_k.$$

Thus the student's choice is determined by the realized confidence vector, and the chosen portfolio is an upper set in that ranking. This is weaker than a fixed threshold, because the stopping point may still depend on the whole vector c. Next, to compare selective and nonselective environments, I impose a sample analogue of Assumption 3, the comparability condition in the main text.

Assumption 10 (Comparable testing designs)**.** The following hold.

(i) *Long-test design.* Conditional on $S_i = s$, the array

$$\big(K_{ij}(0), Y^*_{ij}(0), K_{ij}(1), Y^*_{ij}(1)\big)_{j \geq 1}$$

is i.i.d. across j, and the marginal law of $K_{ij}(z) \mid S_i = s$ is the same in the two regimes and non-degenerate.

(ii) *Large-cohort design.* For each fixed N, the student-level vectors

$$\Big((K_{ij}(0), Y^*_{ij}(0), K_{ij}(1), Y^*_{ij}(1))_{j=1}^{N}, S_i\Big)_{i\geq 1}$$

are i.i.d. across i, the support of S_i is finite, and for every student-type s, the marginal law of $K_{ij}(z) \mid S_i = s$ is the same in the two regimes and non-degenerate.

This is the same economic content as in the main text: regime variation should change selection incentives, not the underlying distribution of latent knowledge within student-type. Compared with the main text, I also need a stronger benchmark regime. Without a fixed threshold, comparing two selective regimes is difficult because both may reflect different and unstable portfolio-dependent cutoffs. To isolate selection cleanly, I therefore compare the selective regime to a nonselective benchmark in which students answer everything.

Assumption 11 (Nonselective benchmark)**.** In regime $z = 0$, students answer every question:

$$D_{ij}(0) = 1 \qquad \forall i, j.$$

For the large-cohort comparison, I also rule out composition changes generated by students who answer no question in the selective regime.

Assumption 12 (Nonempty selective portfolios)**.** For every student i and test length N,

$$|A_{i,N}(1)| \geq 1 \qquad \text{a.s.}$$

This assumption matches the empirical working sample, which retains students with at least one attempted question in each block. It is not needed for the long-test result, where the event under study already bounds the attempted share away from zero. It is needed for the large-cohort result because hit-rates must otherwise be averaged over a regime-1-selected subpopulation whose latent knowledge need not be comparable across regimes.

Finally, for technical simplicity, I rule out knife-edge ties in confidence.

Assumption 13 (No ties in confidence)**.** Conditional on $S_i = s$, the distribution of $C_{ij}(1)$ is atomless.

This is only a regularity condition. It ensures that, in finite tests, the confidence ranking is unique almost surely, so omission reflects selection rather than arbitrary tie-breaking.

Finally, I focus on the deterministic version of Rank-Preserving Metacognition (informally, Equation (2) in the main text) because, once confidence is noisy conditional on knowledge,

realized sample paths need not sort questions by latent quality even if RPM holds on average. The deterministic benchmark restores that pathwise ranking and therefore delivers sharp sample predictions.

Definition 3 (Deterministic Rank-Preserving Metacognition)**.** I say that student-type s satisfies *deterministic Rank-Preserving Metacognition* if there exists a non-constant, weakly increasing map $\psi_s : [0,1] \to [0,1]$ such that, for every j,

$$C_{ij}(1) = \psi_s\Big(K_{ij}(1)\Big) \quad \text{a.s. conditional on } S_i = s.$$

Under Definition 3, confidence-monotone selection implies that, conditional on the realized number of attempts, the attempted questions are those with the highest latent knowledge. In this sense, even without a fixed cutoff, omission in regime 1 still reflects positive sorting on latent knowledge.

For student i, let

$$A_{i,N}(1) := \{j \le N : \ D_{ij}(1) = 1\}$$

denote the set of attempted questions in regime 1. Define the regime-1 omission-rate and the regime-specific hit-rates by

$$O_{i,N}(1) := 1 - \frac{|A_{i,N}(1)|}{N}, \qquad H_{i,N}(1) := \frac{1}{|A_{i,N}(1)|} \sum_{j \in A_{i,N}(1)} Y^*_{ij}(1), \qquad H_{i,N}(0) := \frac{1}{N} \sum_{j=1}^{N} Y^*_{ij}(0),$$

whenever $A_{i,N}(1) \neq \varnothing$. For a sample of M students, let

$$R_{i,N} := \mathbf{1}\{|A_{i,N}(1)| \ge 1\},$$

and average over the subsample of students who attempt at least one question in regime 1:

$$O_{M,N}(1) := \frac{\sum_{i=1}^{M} R_{i,N} O_{i,N}(1)}{\sum_{i=1}^{M} R_{i,N}}, \qquad H_{M,N}(z) := \frac{\sum_{i=1}^{M} R_{i,N} H_{i,N}(z)}{\sum_{i=1}^{M} R_{i,N}}, \quad z \in \{0,1\},$$

whenever the denominator is positive. Under Assumption 12, $R_{i,N} = 1$ almost surely, so the large-cohort objects average over the full student population.

Proposition A12 (Sample analogue of the between-regime test)**.** *Suppose Assumption 9, Assumption 10, Assumption 11, and Assumption 13 hold. Fix any $\varepsilon \in (0, 1/2)$. Then:*

(i) If student i has type s and type s satisfies Definition 3, then

$$\lim_{N \to \infty} \mathbb{P}\Big(\{O_{i,N}(1) \in [\varepsilon, 1-\varepsilon]\} \cap \{H_{i,N}(1) - H_{i,N}(0) \le 0\} \,\Big|\, S_i = s \Big) = 0.$$

(ii) If, in addition, Assumption 12 holds and every student-type in the population satisfies

Definition 3, then for any fixed $N \geq 2$,

$$\lim_{M\to\infty} \mathbb{P}\Big(\{O_{M,N}(1) \in [\varepsilon, 1-\varepsilon]\} \cap \{H_{M,N}(1) - H_{M,N}(0) \leq 0\}\Big) = 0.$$

Thus, even when the answer rule is not described by a fixed threshold, deterministic Rank-Preserving Metacognition implies that omission without positive selection becomes asymptotically unlikely. In the long-test result, concentration across questions removes finite-test noise. In the large-cohort result, nonempty selective portfolios keep the comparison population fixed, and a law of large numbers averages the positive within-student sorting gain across students.

The proof uses the same sorting logic as the main text, but now at the sample level. The key step is to decompose the hit-rate differential into a deterministic *selection term* and a stochastic *noise term*. Under deterministic rank preservation, omitting a nontrivial share of questions means that the student keeps an upper tail of the latent knowledge distribution, so the selection term is bounded away from zero. The Dvoretzky–Kiefer–Wolfowitz inequality (Dvoretzky et al., 1956) controls this upper-tail mean uniformly over all attempted shares bounded away from zero and one. Hoeffding bounds for bounded independent variables (see, e.g., Boucheron et al., 2013) then show that the remaining knowledge and Bernoulli sampling noise vanish exponentially fast. For the large-cohort statement, the same sorting gain is averaged over students and combined with a law of large numbers.

I next record the within-regime sample analogue. An upper-set restriction orders attempted and omitted questions at any fixed confidence vector, but alone does not ensure that a question remains selected when its own confidence rises and the stopping point changes. I therefore impose the corresponding own-confidence monotonicity condition.

Assumption 14 (Own-confidence monotonicity)**.** For every student i, test length N, question j, vector $c_{-j} \in [0,1]^{N-1}$, and $c'_j \geq c_j$,

$$j \in \mathcal{A}_{i,N}(c_j, c_{-j}) \quad \Longrightarrow \quad j \in \mathcal{A}_{i,N}(c'_j, c_{-j}).$$

Economically, holding confidence in every other answer fixed, becoming more confident in question j cannot cause the student to abandon that question. The condition is satisfied by fixed-threshold rules and by rules that answer a fixed number of the highest-confidence questions, while still allowing the stopping point more generally to depend on the rest of the confidence vector.

I keep exactly the same proxy condition as in the main text, restated at sample-level.

Assumption 15 (Sample proxy conditions)**.** Fix a student-type s. Conditional on $S_i = s$, the pairs

$$(K_{ij}(1), R_{ij})_{j=1}^{N}$$

are i.i.d., and:

(i) $r_2 > r_1 \Longrightarrow (K \mid S = s, R = r_2) \succeq_{FOSD} (K \mid S = s, R = r_1)$;

(ii) $D_{ij}(1) \perp R_{ij} \mid (K_{ij}(1), S_i = s)$.

Finally, to define the sample version of the Area Under the Curve, let J be drawn uniformly from $\{1, \ldots, N\}$, independently of the data, and define

$$R^{(1)}_{s,N} \sim (R_{iJ} \mid D_{iJ}(1) = 1, S_i = s), \qquad R^{(0)}_{s,N} \sim (R_{iJ} \mid D_{iJ}(1) = 0, S_i = s).$$

Proposition A13 (Sample AUC under proxy relevance)**.** *Suppose Assumption 9, Assumption 14, and Assumption 15 hold. If type s satisfies Definition 3 and*

$$0 < \mathbb{P}(D_{iJ}(1) = 1 \mid S_i = s) < 1,$$

then

$$(R_{iJ} \mid D_{iJ}(1) = 1, S_i = s) \succeq_{FOSD} (R_{iJ} \mid D_{iJ}(1) = 0, S_i = s),$$

and therefore

$$\mathrm{AUC}_{s,N} := \mathbb{P}\Big(R^{(1)}_{s,N} > R^{(0)}_{s,N}\Big) + \frac{1}{2}\mathbb{P}\Big(R^{(1)}_{s,N} = R^{(0)}_{s,N}\Big) \geq \frac{1}{2}.$$

The proof again mirrors the logic of the main text. Deterministic rank preservation and own-confidence monotonicity imply that a question's probability of being selected is weakly increasing in its own latent knowledge, even though the realized stopping point may depend on the whole portfolio. Proxy relevance then transfers this monotonicity from knowledge to the observable proxy, and Lemma 1 yields the conclusion.

C Proofs

This Appendix Section collects proofs of all the results.

C.1 Main Results

Proof of Lemma 1. Let $p := \mathbb{P}(A = 1) \in (0, 1)$. Fix $t \in \mathbb{R}$ and let $h_t(x) := \mathbf{1}\{x > t\}$.

First observe that

$$\mathbb{P}(T > t \mid A = 1) - \mathbb{P}(T > t \mid A = 0) = \frac{\mathrm{Cov}(h_t(T), \eta(T))}{p(1-p)}.$$

Indeed, since $\mathbb{E}[A \mid T] = \eta(T)$, one has $\mathbb{E}[h_t(T)A] = \mathbb{E}[h_t(T)\eta(T)]$. Also, $\mathbb{E}[h_t(T)A] = p\,\mathbb{E}[h_t(T) \mid A = 1]$ and $\mathbb{E}[h_t(T)] = p\,\mathbb{E}[h_t(T) \mid A = 1] + (1-p)\,\mathbb{E}[h_t(T) \mid A = 0]$. Combining these identities gives the equality.

Second, observe that we can write

$$\mathrm{Cov}(h_t(T), \eta(T)) = \frac{1}{2}\,\mathbb{E}\big[(h_t(T) - h_t(T'))(\eta(T) - \eta(T'))\big]$$

for an independent copy T' of T.

Combining these two observations we obtain the result. Since both h_t and η are weakly increasing, $(h_t(T) - h_t(T'))(\eta(T) - \eta(T')) \geq 0$ a.s.. Hence $\mathrm{Cov}(h_t(T), \eta(T)) \geq 0$, and therefore $\mathbb{P}(T > t \mid A = 1) \geq \mathbb{P}(T > t \mid A = 0) \qquad \forall t \in \mathbb{R}$, i.e. $(T \mid A = 1) \succeq_{FOSD} (T \mid A = 0)$. For the strict part, since T is nondegenerate, then there exists t_0 such that both $\mathbb{P}(T \leq t_0) > 0$ and $\mathbb{P}(T > t_0) > 0$. For this choice of t_0, on the event $\{T > t_0 \geq T'\}$ one has $h_{t_0}(T) - h_{t_0}(T') = 1$ and, because η is strictly increasing on the support of T, also $\eta(T) - \eta(T') > 0$. Since this event has strictly positive probability, it follows that $\mathrm{Cov}(h_{t_0}(T), \eta(T)) > 0$. Therefore $\mathbb{P}(T > t_0 \mid A = 1) > \mathbb{P}(T > t_0 \mid A = 0)$ which implies $(T \mid A = 1) \succ_{FOSD} (T \mid A = 0)$. □

Proof of Proposition 1. Fix a student-type s, and write $\tau_z := \tau(s, z)$ for $z \in \{0, 1\}$. By Assumption 2 and Assumption 3, the conditional law of (K, C) given $S = s$ is invariant across regimes. Hence, within type s, observed regime comparisons coincide with comparisons across the potential regimes generated by the thresholds τ_0 and τ_1. In particular,

$$1 - O_{s,z} = \mathbb{P}(D = 1 \mid S = s, Z = z) = \mathbb{P}(C > \tau_z \mid S = s),$$

and, using Equation (1), Assumption 1(i) and the law of iterated expectations,

$$H_{s,z} = \mathbb{P}(Y = 1 \mid D = 1, S = s, Z = z) = \mathbb{P}(Y^* = 1 \mid D = 1, S = s, Z = z) = \\ = \mathbb{E}[K \mid D(z) = 1, S = s].$$

Now suppose $\Delta O_s > 0$. Then $\mathbb{P}(C > \tau_1 \mid S = s) = 1 - O_{s,1} < 1 - O_{s,0} = \mathbb{P}(C > \tau_0 \mid S = s)$ and since $c \mapsto \mathbb{P}(C > c \mid S = s)$ is weakly decreasing, this implies $\tau_1 > \tau_0$. Consider the sub-population of matches

$$\mathcal{R}_s := \{D(0) = 1, S = s\} = \{C > \tau_0, S = s\}.$$

Within this subpopulation, define the selector $A := D(1)$ and the score $T := K$. Conditional on $(K, S) = (k, s)$ and on belonging to $\mathcal{R}_s$, the probability of remaining selected in regime 1 is

$$\eta_s(k) := \mathbb{P}(D(1) = 1 \mid D(0) = 1, K = k, S = s) = \frac{\overline{F}_s(\tau_1 \mid k)}{\overline{F}_s(\tau_0 \mid k)},$$

where $\overline{F}_s(c \mid k) := \mathbb{P}(C > c \mid S = s, K = k)$. Applying the definition of RPM with $c_2 = \tau_1$ and $c_1 = \tau_0$ shows that $k \mapsto \eta_s(k)$ is weakly increasing, and, if RPM holds strictly, then $k \mapsto \eta_s(k)$ is strictly increasing.

Applying Lemma Lemma 1 on $\mathcal{R}_s$ then yields

$$(K \mid D(1) = 1, S = s) \succeq_{FOSD} (K \mid D(0) = 1, D(1) = 0, S = s),$$

with strict dominance under strict RPM and $\mathbb{V}(K \mid D(0) = 1, S = s) > 0$. Taking expectations gives

$$\mathbb{E}[K \mid D(1) = 1, S = s] \geq \mathbb{E}[K \mid D(0) = 1, D(1) = 0, S = s].$$

Partitioning $\{D(0) = 1\} = \{D(0) = 1, D(1) = 0\} \sqcup \{D(0) = 1, D(1) = 1\}$ so we have that $\mathbb{E}[K \mid D(0) = 1, S = s]$ is a convex combination of $\mathbb{E}[K \mid D(1) = 1, S = s]$ and $\mathbb{E}[K \mid D(0) = 1, D(1) = 0, S = s]$. Hence

$$\mathbb{E}[K \mid D(1) = 1, S = s] \geq \mathbb{E}[K \mid D(0) = 1, S = s].$$

Recalling that these two conditional are precisely $H_{s,1}$ and $H_{s,0}$, we obtain

$$\Delta H_s = H_{s,1} - H_{s,0} \geq 0.$$

If RPM holds strictly and $\mathbb{V}(K \mid D(0) = 1, S = s) > 0$, then the FOSD-dominance above is strict. Moreover, $\Delta O_s > 0$ implies that the screened-out group $\{D(0) = 1, D(1) =$

$0, S = s\}$ has positive probability, while the assumption that $H_{s,1}$ is well-defined implies that the retained group $\{D(1) = 1, S = s\}$ also has positive probability. Hence $\mathbb{E}[K \mid D(0) = 1, S = s]$ is a nondegenerate convex combination of two distinct means, namely $\mathbb{E}[K \mid D(1) = 1, S = s] \neq \mathbb{E}[K \mid D(0) = 1, D(1) = 0, S = s]$. Therefore, $\mathbb{E}[K \mid D(1) = 1, S = s] > \mathbb{E}[K \mid D(0) = 1, S = s]$, and thus $H_{s,1} > H_{s,0}$. □

Proof of Corollary 1. Suppose $\sum_s w_s \Delta O_s > 0$. By Assumption 4, either $\tau(s, 1) \geq \tau(s, 0)$ for all s, or $\tau(s, 1) \leq \tau(s, 0)$ for all s. The second case is impossible under the premise, because then each type would weakly attempt more in regime 1, implying $\Delta O_s \leq 0$ for all s and hence $\sum_s w_s \Delta O_s \leq 0$. Therefore it must be that $\tau(s, 1) \geq \tau(s, 0) \qquad \forall s$.

Fix any type s. If $\Delta O_s > 0$, then Proposition 1 implies $\Delta H_s \geq 0$. If instead $\Delta O_s = 0$, then $\mathbb{P}(C > \tau(s, 1) \mid S = s) = \mathbb{P}(C > \tau(s, 0) \mid S = s)$. Since $\tau(s, 1) \geq \tau(s, 0)$ and $c \mapsto \mathbb{P}(C > c \mid S = s)$ is weakly decreasing, the two threshold rules induce the same attempt set up to null events, so $H_{s,1} = H_{s,0}$ and therefore $\Delta H_s = 0$. Thus, for every s, $\Delta H_s \geq 0$. Multiplying by w_s and summing yields $\sum_s w_s \Delta H_s \geq 0$.

For the strict part, the premise $\sum_s w_s \Delta O_s > 0$ implies that there exists at least one type s with $w_s > 0$ and $\Delta O_s > 0$. For that type, strict RPM and $\mathbb{V}(K \mid D(0) = 1, S = s) > 0$ imply by Proposition 1 that $\Delta H_s > 0$. Since all other types satisfy $\Delta H_s \geq 0$, it follows that $\sum_s w_s \Delta H_s > 0$. □

Proof of Proposition 4. Fix a student-type s and suppress the index throughout. Under Assumption 1(ii), the answer decision is $D = \mathbf{1}\{C > \tau(s)\}$. Hence, conditional on latent knowledge $K = k$, the answer probability is

$$a(k) := \mathbb{P}(D = 1 \mid K = k, S = s) = \overline{F}_s(\tau(s) \mid k).$$

Because type s has Rank-Preserving Metacognition, the family $C \mid (K = k, S = s)$ is increasing in k in the hazard-rate order, and therefore also in the first-order stochastic order. In particular, $a(k)$ is weakly increasing in k.

Now let R be the proxy score. By Assumption 5(ii), $D \perp R \mid (K, S = s)$, so by iterated expectations

$$\mathbb{P}(D = 1 \mid R = r, S = s) = \mathbb{E}[a(K) \mid R = r, S = s].$$

Since $a(\cdot)$ is weakly increasing and Assumption 5(i) says that the conditional law of $K \mid (R = r, S = s)$ shifts upward in the first-order stochastic order as r rises, it follows that

$$\eta(r) := \mathbb{P}(D = 1 \mid R = r, S = s)$$

is weakly increasing. Lemma 1 can therefore be applied with selector $A = D$ and score

$T = R$, conditional on type s. It follows that

$$(R \mid D = 1, S = s) \succeq_{FOSD} (R \mid D = 0, S = s).$$

To conclude, let $R_s^{(1)} \sim (R \mid D = 1, S = s)$, $R_s^{(0)} \sim (R \mid D = 0, S = s)$, and let $R_s^{(0)\prime}$ be an independent copy of $R_s^{(0)}$. For each fixed r_0, the function $r \mapsto \mathbf{1}\{r > r_0\} + \frac{1}{2}\mathbf{1}\{r = r_0\}$ is weakly increasing. Hence first-order stochastic dominance implies

$$\mathbb{E}\Big[\mathbf{1}\{R_s^{(1)} > r_0\} + \frac{1}{2}\mathbf{1}\{R_s^{(1)} = r_0\}\Big] \geq \mathbb{E}\Big[\mathbf{1}\{R_s^{(0)} > r_0\} + \frac{1}{2}\mathbf{1}\{R_s^{(0)} = r_0\}\Big].$$

Taking expectations over $r_0 = R_s^{(0)}$ yields

$$\mathrm{AUC}_s \geq \mathbb{P}\Big(R_s^{(0)} > R_s^{(0)\prime}\Big) + \frac{1}{2}\mathbb{P}\Big(R_s^{(0)} = R_s^{(0)\prime}\Big).$$

Since $R_s^{(0)}$ and $R_s^{(0)\prime}$ are i.i.d., the right-hand side equals $1/2$. Therefore $\mathrm{AUC}_s \geq 1/2$. □

Proof of Proposition 2. Fix a match type (s, q). If $O_{s,q,1} = 0$, then the student answers every question of that match type under regime 1. Hence selective answering coincides with the nonselective benchmark, and $H_{s,q,1} = H_{s,q,0}$.

Now suppose $O_{s,q,1} > 0$. In the binary model, the threshold rule selects questions according to confidence. Thus, in the selective regime, answered questions are precisely the high-confidence questions. Therefore $1 - O_{s,q,1}$ is the probability of the event $C = \bar{c}$, conditional on the match type (s, q), and $H_{s,q,1}$ is the average value of K on that event. Hence

$$(1 - O_{s,q,1})H_{s,q,1} = \mathbb{E}[K\mathbf{1}\{C = \bar{c}\} \mid S = s, Q = q]\,.$$

By contrast, in the nonselective benchmark the hit-rate is simply the unconditional average value of K within the same match type:

$$H_{s,q,0} = \mathbb{E}[K \mid S = s, Q = q].$$

Multiplying the difference $H_{s,q,1} - H_{s,q,0}$ by the answered share $1 - O_{s,q,1}$, we therefore obtain

$$\begin{aligned}&(1 - O_{s,q,1})(H_{s,q,1} - H_{s,q,0}) = \\ &\mathbb{E}[K\mathbf{1}\{C = \bar{c}\} \mid S = s, Q = q] - \mathbb{E}[K \mid S = s, Q = q]\mathbb{P}(C = \bar{c} \mid S = s, Q = q).\end{aligned}$$

It remains only to evaluate the two terms on the right-hand side. In the binary model,

$$\mathbb{P}(C = \bar{c} \mid S = s, Q = q) = \pi_{s,q}\psi_{s,q}^{+} + (1 - \pi_{s,q})(1 - \psi_{s,q}^{-}),$$

while

$$\mathbb{E}[K\mathbf{1}\{C=\overline{c}\} \mid S=s, Q=q] = \pi_{s,q}\psi^{+}_{s,q}\overline{k} + (1-\pi_{s,q})(1-\psi^{-}_{s,q})\underline{k}.$$

Substituting these expressions, and using

$$\mathbb{E}[K \mid S=s, Q=q] = \pi_{s,q}\overline{k} + (1-\pi_{s,q})\underline{k},$$

gives, after collecting terms,

$$(1-O_{s,q,1})(H_{s,q,1}-H_{s,q,0}) = (\overline{k}-\underline{k})\pi_{s,q}(1-\pi_{s,q})(\psi^{+}_{s,q}+\psi^{-}_{s,q}-1).$$

This is exactly Equation (9). □

Proof of Proposition 3. Fix a student type s and a regime z. If $O_{s,z}=0$, then the student answers every question. Since a correct answer pays h_z, an incorrect answer pays zero, and K is the probability of a correct answer, the expected score per question is

$$W_{s,z} = h_z\mathbb{E}_{\pi_s}[k].$$

Now suppose $O_{s,z}>0$. In the binary model, selective answering is governed by the confidence threshold. Thus high-confidence questions are answered and low-confidence questions are omitted. The expected score can therefore be written directly by conditioning on the latent knowledge state:

$$W_{s,z} = \pi_s\Big[\psi^{+}_s h_z\overline{k} + (1-\psi^{+}_s)o_z\Big] + (1-\pi_s)\Big[(1-\psi^{-}_s)h_z\underline{k} + \psi^{-}_s o_z\Big].$$

The first bracket is the expected score in the high-knowledge state: with probability ψ^{+}_s the student is confident and answers, while with probability $1-\psi^{+}_s$ she omits. The second bracket is the analogous expression in the low-knowledge state.

Rearranging this expression around the payoff from omitting in the high-knowledge state and answering in the low-knowledge state gives

$$W_{s,z} = \pi_s o_z + (1-\pi_s)h_z\underline{k} + \pi_s\psi^{+}_s g_z(\overline{k}) + (1-\pi_s)\psi^{-}_s g_z(\underline{k}).$$

To express the same score in terms of sensitivity and bias, use the identities

$$\psi^{+}_s = \frac{1+\psi^{\Sigma}_s+\psi^{\Delta}_s}{2}, \qquad \psi^{-}_s = \frac{1+\psi^{\Sigma}_s-\psi^{\Delta}_s}{2}.$$

Substituting these identities into the previous expression and collecting the constant, sensitivity, and bias terms yields

$$W_{s,z} = \frac{h_z\mathbb{E}_{\pi_s}[k]+o_z}{2} + \frac{1}{2}\left[\mathbb{E}_{\pi_s}[g_z(k)]\psi^{\Sigma}_s + \Big(h_z\mathbb{E}_{\pi_s}[k]-o_z\Big)\psi^{\Delta}_s\right].$$

Using the notation introduced in the statement of the proposition, this is exactly

$$W_{s,z} = K_z(\pi_s) + \frac{1}{2}\left[A_z(\pi_s)\psi_s^{\Sigma} + B_z(\pi_s)\psi_s^{\Delta}\right],$$

which proves Equation (10). □

C.2 Additional Results

Proof of Proposition A5. Fix c, and let $A^* = A^*(c)$ be optimal. Suppose, toward a contradiction, that there exist $j \in A^*$ and $\ell \notin A^*$ such that $c_\ell > c_j$. Define

$$A' := (A^* \setminus \{j\}) \cup \{\ell\}.$$

Then $|A'| = |A^*|$, so the cost term is unchanged. Let

$$Y := \sum_{r \in A^* \setminus \{j\}} B_r.$$

Then

$$B(A^*) = Y + B_j, \qquad B(A') = Y + B_\ell.$$

Since $c_\ell > c_j$, the Bernoulli variable B_ℓ strictly first-order stochastically dominates B_j. Because Y is independent of both, $Y + B_\ell$ strictly first-order stochastically dominates $Y + B_j$. Hence, by Assumption 6,

$$V(A'; c) > V(A^*; c),$$

contradicting optimality of A^*. □

Proof of Proposition A6. Define the one-question gain function

$$g(c) := V(\{1\}; (c)) - V(\varnothing; \varnothing).$$

Because Bernoulli$(c') \succeq_{FOSD}$ Bernoulli(c) whenever $c' > c$, Assumption 6 implies that g is weakly increasing on $[0, 1]$. By Assumption 7, g is also continuous.

Now define

$$\tau := \sup\{c \in [0, 1] : g(c) \leq 0\},$$

with the convention $\tau = 0$ if $g(c) > 0$ for all c, and $\tau = 1$ if $g(c) \leq 0$ for all c. Since g is continuous and weakly increasing, it follows that

$$g(c) > 0 \quad \Longleftrightarrow \quad c > \tau.$$

Consider any realized confidence vector c. If $c_j > \tau$, then $g(c_j) > 0$, so by Assumption 7

$$V(A \cup \{j\}; c) > V(A; c) \qquad \text{for every } A \not\ni j.$$

Hence any optimal portfolio must include j. Conversely, if $c_j < \tau$, then $g(c_j) < 0$, so for every $A \not\ni j$,

$$V(A \cup \{j\}; c) < V(A; c).$$

Therefore no optimal portfolio can include j. Finally, if $c_j = \tau$, then $g(c_j) = 0$ whenever $\tau \in (0, 1)$, and under the tie-breaking convention the student skips. Thus the induced decision rule is

$$D = \mathbf{1}\{C > \tau\}.$$

□

Proof of Proposition A7. For part (i), since $K = \mathbb{E}[Y^* \mid t, X, M]$, the tower property gives

$$\mathbb{P}(Y^* = 1 \mid K, C) = \mathbb{E}[\mathbb{E}[Y^* \mid t, X, M] \mid K, C] = \mathbb{E}[K \mid K, C] = K.$$

Because $Y^* \in \{0, 1\}$, this is equivalent to

$$Y^* \mid (K, C) \sim \text{Bernoulli}(K).$$

For part (ii), again by the tower property,

$$\mathbb{E}[K \mid X, M] = \mathbb{E}[\mathbb{E}[Y^* \mid t, X, M] \mid X, M] = \mathbb{E}[Y^* \mid X, M] = C.$$

Since C is measurable with respect to (X, M), conditioning once more yields

$$\mathbb{E}[K \mid C] = \mathbb{E}[\mathbb{E}[K \mid X, M] \mid C] = \mathbb{E}[C \mid C] = C.$$

□

Proof of Proposition A8. Set the latent type equal to the target knowledge level:

$$t := \widetilde{K}.$$

Let the prior on t be the marginal law of $\widetilde{K}$. Let the meta-signal be

$$M := \widetilde{C},$$

with conditional law equal to a regular conditional distribution of $\widetilde{C}$ given $\widetilde{K}$.

Take the first-order signal space to be $\mathcal{X} = \Omega$. For each realized type $t = k$, define the

symmetric signal structure

$$\xi_k(x \mid \omega) = \begin{cases} k, & x = \omega, \\ \dfrac{1-k}{|\Omega|-1}, & x \neq \omega. \end{cases}$$

Because $k \geq 1/|\Omega|$, the posterior-optimal forced answer is $a(x, m) = x$.

Now fix a realization (x, m). If the student knew k, the posterior probability that $\omega = x$ would be exactly k. Hence the induced knowledge index is

$$K = k = t.$$

Next, under the uniform prior the unconditional distribution of X does not depend on k:

$$\mathbb{P}(X = x \mid t = k) = \frac{1}{|\Omega|} \qquad \forall x,\ k.$$

Hence observing X provides no additional information about t. The student's posterior over t after observing $(X, M) = (x, m)$ is therefore the same as after observing $M = m$ alone. Under the symmetric signal structure above, the posterior probability that $\omega = x$ is the posterior mean of k. Therefore the induced confidence index is

$$C = \mathbb{E}[t \mid M].$$

By construction, $(t, M) \overset{d}{=} (\widetilde{K}, \widetilde{C})$, so the martingale condition implies

$$C = \mathbb{E}[\widetilde{K} \mid \widetilde{C}] = \widetilde{C} = M.$$

Hence $(K, C) = (t, M)$ in distribution, and therefore

$$(K, C) \overset{d}{=} (\widetilde{K}, \widetilde{C}).$$

□

Proof of Proposition A9. Write $\tau_z := \tau(s, z)$ and let $\nu_0 := \mathcal{L}(K \mid D(0) = 1, S = s)$ be the law of knowledge among the pooled answered in the low-cost regime. Because $\Delta O_s > 0$, one has $\delta_s > 0$, as in the Proof of Proposition 1. Now work inside the low-cost attempt pool $\{D(0) = 1, S = s\}$, i.e. with respect to the reference measure ν_0. Define

$$\eta(k) := \mathbb{P}(D(1) = 1 \mid D(0) = 1, K = k, S = s) = \frac{\overline{F}_s(\tau_1 \mid k)}{\overline{F}_s(\tau_0 \mid k)}.$$

Applying Definition 2 with $c_2 = \tau_1$ and $c_1 = \tau_0$ gives, for every $k_2 > k_1$,

$$\log \eta(k_2) - \log \eta(k_1) \geq \kappa \delta_s (k_2 - k_1).$$

Hence $r(k) := \eta(k) e^{-ak}$, where $a := \kappa \delta_s$ is weakly increasing.

Next, consider a measure ν_a obtained "tilting" ν_0,

$$d\nu_a := \frac{e^{ak} d\nu_0}{\int e^{au} d\nu_0}.$$

Since

$$H_{s,1} = \frac{\int k\, \eta(k)\, d\nu_0}{\int \eta(k)\, d\nu_0} = \frac{\int k\, r(k)\, d\nu_a}{\int r(k)\, d\nu_a},$$

and both $k \mapsto k$ and $k \mapsto r(k)$ are weakly increasing, Chebyshev's covariance inequality yields $H_{s,1} \geq \int k\, d\nu_a$. Moreover, $H_{s,0} = \int k\, d\nu_0$. Therefore

$$\Delta H_s \geq \int k\, d\nu_a - \int k\, d\nu_0.$$

To bound the right-hand side, define the moment generating function $M(t) := \int e^{tk}\, d\nu_0$, for $t \in [0, a]$. Then

$$\frac{d}{dt} \log M(t) = \mathbb{E}_{\nu_t}[K] \quad \text{and} \quad \frac{d^2}{dt^2} \log M(t) = \mathbb{V}_{\nu_t}(K),$$

so

$$\int k\, d\nu_a - \int k\, d\nu_0 = \int_0^a \mathbb{V}_{\nu_t}(K)\, dt.$$

Now let K' be an independent copy of K under ν_t. The pairwise representation of the variance gives

$$\mathbb{V}_{\nu_t}(K) = \frac{1}{2}\, \mathbb{E}_{\nu_t \otimes \nu_t}\Big[(K - K')^2\Big].$$

Moreover,

$$\frac{d(\nu_t \otimes \nu_t)}{d(\nu_0 \otimes \nu_0)}(k, k') = \frac{e^{t(k+k')}}{M(t)^2} \geq e^{-2t},$$

because $k, k' \in [0, 1]$ and $M(t) \leq e^t$. Hence

$$\mathbb{V}_{\nu_t}(K) \geq e^{-2t} \mathbb{V}_{\nu_0}(K).$$

Integrating from 0 to a yields

$$\Delta H_s \geq \mathbb{V}_{\nu_0}(K) \int_0^a e^{-2t}\, dt = \mathbb{V}(K \mid D(0) = 1, S = s)\, \frac{1 - e^{-2a}}{2},$$

which is exactly Equation (18) because $a = \kappa \delta_s$. □

Proof of Proposition A10. Write

$$a_s(k) := \mathbb{P}(D = 1 \mid K = k, S = s) = \overline{F}_s(\tau_s \mid k).$$

Applying Definition 2 with $c_2 = \tau_s$ and $c_1 = 0$, and using $\overline{F}_s(0 \mid k) = 1$, yields

$$\log a_s(k_2) - \log a_s(k_1) \geq \kappa \tau_s (k_2 - k_1) \qquad \forall\, k_2 > k_1.$$

Therefore, for every $k_2 > k_1$,

$$a_s(k_2) - a_s(k_1) = a_s(k_1)\Big(\exp\Big(\log a_s(k_2) - \log a_s(k_1)\Big) - 1\Big) \geq \underline{a}_s\, \kappa \tau_s\, (k_2 - k_1),$$

because $e^x - 1 \geq x$ for $x \geq 0$. Hence the function

$$b_s(k) := a_s(k) - m_s k, \qquad m_s := \underline{a}_s\, \kappa \tau_s,$$

is weakly increasing. Let

$$\pi_s(u) := \mathbb{P}(D = 1 \mid U_s = u, S = s).$$

By Assumption 5 (Exogeneity),

$$\pi_s(u) = \mathbb{E}[a_s(K) \mid U_s = u, S = s] = m_s\, \mathbb{E}[K \mid U_s = u, S = s] + \mathbb{E}[b_s(K) \mid U_s = u, S = s].$$

Since the proxy is ℓ_s-string the function

$$u \longmapsto \mathbb{E}[K \mid U_s = u, S = s] - \ell_s u$$

is weakly increasing. By Assumption 5 (Validity), the conditional law of $K \mid U_s = u, S = s$ shifts upward in first-order stochastic order as u rises; since b_s is increasing, the function

$$u \longmapsto \mathbb{E}[b_s(K) \mid U_s = u, S = s]$$

is weakly increasing as well. Consequently,

$$u \longmapsto \pi_s(u) - m_s \ell_s u$$

is weakly increasing. Now compute

$$\mathbb{E}[U_s \mid D = 1, S = s] - \mathbb{E}[U_s \mid D = 0, S = s] = \frac{\mathrm{Cov}(U_s, \pi_s(U_s) \mid S = s)}{p_s(1 - p_s)}.$$

Since $\pi_s(u) = m_s \ell_s u + g_s(u)$ for some weakly increasing function g_s, Chebyshev's covariance

inequality gives

$$\mathrm{Cov}(U_s, \pi_s(U_s) \mid S = s) \geq m_s \ell_s \mathbb{V}(U_s \mid S = s).$$

Hence

$$\mathbb{E}[U_s \mid D = 1, S = s] - \mathbb{E}[U_s \mid D = 0, S = s] \geq \frac{m_s \ell_s \mathbb{V}(U_s \mid S = s)}{p_s(1 - p_s)}.$$

It remains to identify the left-hand side. Let $U_s^{(1)} \sim (U_s \mid D = 1, S = s)$ and $U_s^{(0)} \sim (U_s \mid D = 0, S = s)$ be independent. By the definition of the mid-rank transform,

$$\mathbb{E}[U_s \mid D = 1, S = s] = \frac{p_s}{2} + (1 - p_s)\mathrm{AUC}_s,$$

and similarly

$$\mathbb{E}[U_s \mid D = 0, S = s] = p_s(1 - \mathrm{AUC}_s) + \frac{1 - p_s}{2}.$$

Subtracting gives

$$\mathbb{E}[U_s \mid D = 1, S = s] - \mathbb{E}[U_s \mid D = 0, S = s] = \mathrm{AUC}_s - \frac{1}{2}.$$

Substituting $m_s = \underline{a}_s \kappa \tau_s$ and $\mathbb{V}(U_s \mid S = s) = 1/12$ gives the result. □

Proof of Proposition A11. Fix a student-type s and a knowledge distribution $\mu \in \Delta([0, 1])$. For any threshold $z \in [0, 1]$ such that $\mathbb{P}_\mu(D(z) = 1 \mid S = s) > 0$, the selection surplus satisfies

$$\gamma_{\mu,s}(z) = \Big(1 - O_{\mu,s}(z)\Big)\Big(H_{\mu,s}(z) - H_\mu^0\Big) = \mathrm{Cov}_\mu\Big(K, \mathbf{1}\{D(z) = 1\} \mid S = s\Big). \tag{20}$$

Indeed,

$$\mathrm{Cov}_\mu\Big(K, \mathbf{1}\{D(z) = 1\} \mid S = s\Big) = \mathbb{P}_\mu(D(z) = 1 \mid S = s)\Big(\mathbb{E}_\mu[K \mid D(z) = 1, S = s] - H_\mu^0\Big).$$

A second identity follows from the law of total covariance. For any bounded measurable function g,

$$\mathrm{Cov}_\mu\Big(K, g(C) \mid S = s\Big) = \mathrm{Cov}_\mu\Big(K, \mathbb{E}[g(C) \mid K, S = s] \mid S = s\Big). \tag{21}$$

Since K is measurable with respect to itself, the conditional covariance term vanishes. Taking $g(c) = \mathbf{1}\{c > z\}$, and using $D(z) = \mathbf{1}\{C > z\}$, yields

$$\gamma_{\mu,s}(z) = \mathrm{Cov}_\mu\Big(K, \overline{F}_s(z \mid K) \mid S = s\Big), \tag{22}$$

where $\overline{F}_s(z \mid k) = \mathbb{P}(C > z \mid K = k, S = s)$.

Finally, I repeatedly use the following elementary characterization:

$$h \text{ is weakly increasing } \iff \mathrm{Cov}_\mu\Big(K, h(K) \mid S = s\Big) \geq 0 \quad \text{for every } \mu \in \Delta([0,1]). \tag{23}$$

The forward implication is Chebyshev's association inequality. For the reverse implication, if h were not weakly increasing, there would exist $k_2 > k_1$ such that $h(k_2) < h(k_1)$. Taking $\mu = \frac{1}{2}\delta_{k_1} + \frac{1}{2}\delta_{k_2}$ would then yield a negative covariance.

Point (i). By Equation (20) and Fubini's theorem,

$$\int_0^1 \gamma_{\mu,s}(z)\, dz = \mathrm{Cov}_\mu\Big(K, \int_0^1 \mathbf{1}\{D(z) = 1\}\, dz \;\Big|\; S = s\Big).$$

Since $D(z) = \mathbf{1}\{C > z\}$ and $C \in [0,1]$,

$$\int_0^1 \mathbf{1}\{D(z) = 1\}\, dz = \int_0^1 \mathbf{1}\{C > z\}\, dz = C.$$

Hence

$$\int_0^1 \gamma_{\mu,s}(z)\, dz = \mathrm{Cov}_\mu(K, C \mid S = s).$$

Applying Equation (21) with $g(c) = c$,

$$\mathrm{Cov}_\mu(K, C \mid S = s) = \mathrm{Cov}_\mu\Big(K, \mathbb{E}[C \mid K, S = s] \mid S = s\Big).$$

Therefore, by Equation (23), the integral is nonnegative for every μ if and only if the map

$$k \longmapsto \mathbb{E}[C \mid K = k, S = s]$$

is weakly increasing, which is exactly expectation monotonicity.

Point (ii). Fix a threshold z with $\mathbb{P}_\mu(D(z) = 1 \mid S = s) > 0$. Since $1 - O_{\mu,s}(z) > 0$, one has

$$H_{\mu,s}(z) \geq H^0_\mu \iff \gamma_{\mu,s}(z) \geq 0.$$

By Equation (22), this is equivalent to

$$\mathrm{Cov}_\mu\Big(K, \overline{F}_s(z \mid K) \mid S = s\Big) \geq 0.$$

Hence, by Equation (23), the inequality $H_{\mu,s}(z) \geq H^0_\mu$ holds for every μ if and only if the map

$$k \longmapsto \overline{F}_s(z \mid k)$$

is weakly increasing. Requiring this for every $z \in [0,1]$ is exactly stochastic monotonicity.

Point (iii). Suppose first that $K \mapsto C$ is hazard-rate monotone. Fix μ and two thresholds $z_2 > z_1$ such that $\mathbb{P}_\mu(D(z_1) = 1 \mid S = s) > 0$. Let μ_{z_1} denote the law of K conditional on

$D(z_1) = 1$. Then

$$H_{\mu,s}(z_2) = \frac{\mathbb{E}_{\mu_{z_1}}[K\,\rho_{z_1,z_2}(K) \mid S = s]}{\mathbb{E}_{\mu_{z_1}}[\rho_{z_1,z_2}(K) \mid S = s]}, \qquad \rho_{z_1,z_2}(k) := \frac{\overline{F}_s(z_2 \mid k)}{\overline{F}_s(z_1 \mid k)}.$$

By hazard-rate monotonicity, $k \mapsto \rho_{z_1,z_2}(k)$ is weakly increasing wherever well defined. Applying Chebyshev's association inequality under the measure μ_{z_1} yields

$$H_{\mu,s}(z_2) \geq H_{\mu,s}(z_1).$$

Thus $z \mapsto H_{\mu,s}(z)$ is weakly increasing.

Conversely, suppose that for every μ, the map $z \mapsto H_{\mu,s}(z)$ is weakly increasing on the set of thresholds with positive selection. Fix $k_2 > k_1$ and let

$$\mu = \lambda\delta_{k_1} + (1-\lambda)\delta_{k_2} \qquad \text{for some } \lambda \in (0,1).$$

Write

$$s_i(z) := \overline{F}_s(z \mid k_i), \qquad i = 1, 2.$$

Whenever $s_1(z) > 0$,

$$H_{\mu,s}(z) = \frac{\lambda s_1(z) k_1 + (1-\lambda) s_2(z) k_2}{\lambda s_1(z) + (1-\lambda) s_2(z)} = \frac{\lambda k_1 + (1-\lambda)\sigma(z) k_2}{\lambda + (1-\lambda)\sigma(z)},$$

where

$$\sigma(z) := \frac{s_2(z)}{s_1(z)}.$$

For fixed $\lambda \in (0,1)$ and $k_2 > k_1$, the last expression is strictly increasing in $\sigma(z)$. Hence, monotonicity of $z \mapsto H_{\mu,s}(z)$ implies that $z \mapsto \sigma(z)$ is weakly increasing. Therefore, for every $z_2 > z_1$ such that $s_1(z_1), s_1(z_2) > 0$,

$$\frac{\overline{F}_s(z_2 \mid k_2)}{\overline{F}_s(z_2 \mid k_1)} \geq \frac{\overline{F}_s(z_1 \mid k_2)}{\overline{F}_s(z_1 \mid k_1)},$$

or equivalently,

$$\overline{F}_s(z_2 \mid k_2)\,\overline{F}_s(z_1 \mid k_1) \geq \overline{F}_s(z_2 \mid k_1)\,\overline{F}_s(z_1 \mid k_2).$$

Since $k_2 > k_1$ and $z_2 > z_1$ were arbitrary, this is exactly hazard-rate monotonicity. □

Proof of Proposition A12. Part (i). Fix a student i of type s, suppress the student index throughout, and condition all probabilities below on $S = s$. Let

$$A_N := |A_N(1)|, \qquad E_N := \{O_N(1) \in [\varepsilon, 1-\varepsilon]\} = \left\{\frac{A_N}{N} \in [\varepsilon, 1-\varepsilon]\right\}.$$

By Assumption 10(i), the marginal law of latent knowledge is the same in the two regimes,

so

$$\mu_s := \mathbb{E}[K(0) \mid S = s] = \mathbb{E}[K(1) \mid S = s].$$

Since type s satisfies Definition 3, there exists a nonconstant weakly increasing map ψ_s such that

$$C(1) = \psi_s(K(1)) \qquad \text{a.s. conditional on } S = s.$$

Define

$$m_s(c) := \mathbb{E}[K(1) \mid C(1) = c, S = s], \qquad X := m_s(C(1)).$$

Because $C(1)$ is a weakly increasing deterministic transformation of $K(1)$, the map m_s is weakly increasing on the support of $C(1)$. Moreover, Assumption 13 and the nondegeneracy of $K(1) \mid S = s$ imply that X is nonconstant. By iterated expectations,

$$\mathbb{E}[X \mid S = s] = \mathbb{E}[K(1) \mid S = s] = \mu_s.$$

Now decompose

$$H_N(1) - H_N(0) = \Delta_N + \nu_N,$$

where

$$\Delta_N := \frac{1}{A_N} \sum_{j \in A_N(1)} m_s(C_j(1)) - \mu_s$$

and

$$\nu_N := \frac{1}{A_N} \sum_{j \in A_N(1)} \Big(K_j(1) - m_s(C_j(1))\Big) + \frac{1}{A_N} \sum_{j \in A_N(1)} \Big(Y_j^*(1) - K_j(1)\Big) - \frac{1}{N} \sum_{j=1}^{N} \Big(K_j(0) - \mu_s\Big) - \frac{1}{N} \sum_{j=1}^{N} \Big(Y_j^*(0) - K_j(0)\Big).$$

I first study Δ_N. By Assumption 9 and Assumption 13, the selected set consists almost surely of the A_N highest confidence realizations. Since m_s is weakly increasing, these are also the A_N highest realizations of $X_j := m_s(C_j(1))$, up to irrelevant ties in X_j. Let $F_{X,s}$ denote the cdf of $X \mid S = s$, and define for $\alpha \in (0, 1)$

$$g_s(\alpha) := \frac{1}{\alpha} \int_{1-\alpha}^{1} F_{X,s}^{-1}(u)\, du.$$

Because X is bounded and nonconstant, g_s is continuous on $(0, 1)$ and

$$g_s(\alpha) > \mathbb{E}[X \mid S = s] = \mu_s \qquad \forall \alpha \in (0, 1).$$

Hence

$$\gamma_{s,\varepsilon} := \min_{\alpha \in [\varepsilon, 1-\varepsilon]} \Big(g_s(\alpha) - \mu_s\Big) > 0.$$

Let $\widehat{F}_{X,N}$ denote the empirical cdf of $(X_j)_{j=1}^N$, and set $\alpha_N := A_N/N$. On E_N, the selected average is the empirical upper-tail mean:

$$\frac{1}{A_N}\sum_{j\in A_N(1)} X_j = \frac{1}{\alpha_N}\int_{1-\alpha_N}^1 \widehat{F}^{-1}_{X,N}(u)\,du.$$

For every $\alpha \in [\varepsilon, 1-\varepsilon]$,

$$\begin{aligned}\left|\frac{1}{\alpha}\int_{1-\alpha}^1 \left(\widehat{F}^{-1}_{X,N}(u) - F^{-1}_{X,s}(u)\right) du\right| &\le \frac{1}{\varepsilon}\int_0^1 \left|\widehat{F}^{-1}_{X,N}(u) - F^{-1}_{X,s}(u)\right| du \\ &= \frac{1}{\varepsilon} W_1(\widehat{F}_{X,N}, F_{X,s}) \\ &\le \frac{1}{\varepsilon}\|\widehat{F}_{X,N} - F_{X,s}\|_\infty,\end{aligned}$$

where the last inequality uses that $X \in [0,1]$. The Dvoretzky–Kiefer–Wolfowitz inequality therefore gives, for a universal constant $C_0 < \infty$,

$$\mathbb{P}\left(E_N \cap \left\{\Delta_N \le \frac{\gamma_{s,\varepsilon}}{2}\right\} \,\middle|\, S = s\right) \le C_0 \exp\left(-\frac{N\varepsilon^2\gamma^2_{s,\varepsilon}}{2}\right) \longrightarrow 0.$$

I next control ν_N. Conditional on the confidence vector $(C_j(1))_{j=1}^N$, the selected set is deterministic, and the variables

$$K_j(1) - m_s(C_j(1))$$

are independent across j, centered, and have conditional range at most one. Therefore, on E_N, Hoeffding's inequality gives for every $t > 0$,

$$\mathbb{P}\left(E_N \cap \left\{\left|\frac{1}{A_N}\sum_{j\in A_N(1)} (K_j(1) - m_s(C_j(1)))\right| > t\right\} \,\middle|\, S = s\right) \le 2e^{-2\varepsilon N t^2}.$$

Conditional instead on the full regime-1 knowledge vector, the selection set remains deterministic and, by Assumption 1.i, the variables

$$Y_j^*(1) - K_j(1)$$

are independent across j, centered, and have conditional range one. Hence

$$\mathbb{P}\left(E_N \cap \left\{\left|\frac{1}{A_N}\sum_{j\in A_N(1)} (Y_j^*(1) - K_j(1))\right| > t\right\} \,\middle|\, S = s\right) \le 2e^{-2\varepsilon N t^2}.$$

For regime 0, Assumption 11 ensures that all N questions are answered. The variables $K_j(0) - \mu_s$ are independent, centered, and have range one, while conditional on the regime-0

knowledge vector the variables $Y_j^*(0) - K_j(0)$ are likewise independent, centered, and have range one. Thus

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{j=1}^{N}(K_j(0)-\mu_s)\right| > t \,\middle|\, S=s\right) \le 2e^{-2Nt^2}$$

and

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{j=1}^{N}(Y_j^*(0)-K_j(0))\right| > t \,\middle|\, S=s\right) \le 2e^{-2Nt^2}.$$

Choosing $t = \gamma_{s,\varepsilon}/8$ and applying the union bound gives

$$\mathbb{P}\Big(E_N \cap \Big\{|\nu_N| > \frac{\gamma_{s,\varepsilon}}{2}\Big\} \,\Big|\, S=s\Big) \le 4e^{-N\varepsilon\gamma_{s,\varepsilon}^2/32} + 4e^{-N\gamma_{s,\varepsilon}^2/32} \longrightarrow 0.$$

Finally,

$$E_N \cap \{H_N(1) - H_N(0) \le 0\} \subseteq E_N \cap \Big\{\Delta_N \le \frac{\gamma_{s,\varepsilon}}{2}\Big\} \;\cup\; E_N \cap \Big\{|\nu_N| > \frac{\gamma_{s,\varepsilon}}{2}\Big\},$$

so the desired probability converges to zero.

Part (ii). Fix $N \ge 2$. By Assumption 12, $R_{i,N} = 1$ almost surely for every student. Define

$$G_i := \frac{1}{|A_{i,N}(1)|}\sum_{j\in A_{i,N}(1)} K_{ij}(1) - \frac{1}{N}\sum_{j=1}^{N} K_{ij}(1).$$

By Assumption 9, Assumption 13, and deterministic RPM, the attempted set contains the questions with the highest latent knowledge. Consequently,

$$G_i \ge 0 \qquad \text{a.s.},$$

with strict inequality whenever $0 < |A_{i,N}(1)| < N$.

By Assumption 1.i and iterated expectations,

$$\begin{aligned}\mathbb{E}[H_{i,N}(1) - H_{i,N}(0) \mid S_i = s] &= \mathbb{E}\left[\frac{1}{|A_{i,N}(1)|}\sum_{j\in A_{i,N}(1)} K_{ij}(1) - \frac{1}{N}\sum_{j=1}^{N} K_{ij}(0) \,\middle|\, S_i = s\right] \\ &= \mathbb{E}[G_i \mid S_i = s],\end{aligned}$$

where the second equality uses the equality of the marginal knowledge distributions across regimes in Assumption 10(ii). Thus every student-type has a weakly nonnegative expected hit-rate gain, and the gain is strictly positive for any type that omits with positive probability.

Let

$$\omega := \mathbb{E}[O_{i,N}(1)], \qquad \zeta := \mathbb{E}[H_{i,N}(1) - H_{i,N}(0)] = \mathbb{E}[G_i].$$

By Assumption 10(ii), the student-level vectors are i.i.d. across i. Hence the strong law of

large numbers gives

$$O_{M,N}(1) \xrightarrow{a.s.} \omega, \qquad H_{M,N}(1) - H_{M,N}(0) \xrightarrow{a.s.} \zeta.$$

If $\omega \notin [\varepsilon, 1-\varepsilon]$, then

$$\mathbb{P}\Big(O_{M,N}(1) \in [\varepsilon, 1-\varepsilon]\Big) \to 0.$$

If instead $\omega \in [\varepsilon, 1-\varepsilon]$, then $\omega > 0$. Since selective portfolios are nonempty, this implies

$$\mathbb{P}(0 < |A_{i,N}(1)| < N) > 0.$$

Therefore $\zeta = \mathbb{E}[G_i] > 0$, and

$$\mathbb{P}\Big(H_{M,N}(1) - H_{M,N}(0) \le 0\Big) \to 0.$$

In either case,

$$\mathbb{P}\Big(\{O_{M,N}(1) \in [\varepsilon, 1-\varepsilon]\} \cap \{H_{M,N}(1) - H_{M,N}(0) \le 0\}\Big) \to 0.$$

□

Proof of Proposition A13. Fix a student i of type s, and let J be uniform on $\{1, \dots, N\}$, independently of the data. For each question j, define

$$a_{j,s,N}(k) := \mathbb{P}(D_{ij}(1) = 1 \mid K_{ij}(1) = k, S_i = s).$$

I first claim that each $a_{j,s,N}$ is weakly increasing in k.

Fix a realization c_{-j} of the other $N-1$ confidences. Since type s satisfies Definition 3,

$$C_{ij}(1) = \psi_s(K_{ij}(1))$$

for some weakly increasing map ψ_s. If $k_2 > k_1$, then $\psi_s(k_2) \ge \psi_s(k_1)$. By Assumption 14,

$$j \in \mathcal{A}_{i,N}(\psi_s(k_1), c_{-j}) \quad \Longrightarrow \quad j \in \mathcal{A}_{i,N}(\psi_s(k_2), c_{-j}).$$

Thus the selection indicator is weakly increasing in the question's own knowledge for every fixed c_{-j}. Because the question-level pairs are i.i.d. under Assumption 15, the distribution of $C_{i,-j}(1)$ does not depend on $K_{ij}(1)$, conditional on $S_i = s$. Averaging over $C_{i,-j}(1)$ proves that $a_{j,s,N}(k)$ is weakly increasing.

For the randomly sampled slot J, define

$$a_{s,N}(k) := \mathbb{P}(D_{iJ}(1) = 1 \mid K_{iJ}(1) = k, S_i = s) = \frac{1}{N}\sum_{j=1}^{N} a_{j,s,N}(k).$$

This function is also weakly increasing. Now let R denote the proxy for the randomly sampled slot. By Assumption 15(ii) and iterated expectations,

$$\mathbb{P}(D_{iJ}(1) = 1 \mid R = r, S_i = s) = \mathbb{E}[a_{s,N}(K) \mid R = r, S_i = s].$$

Since $a_{s,N}$ is weakly increasing and Assumption 15(i) implies that $K \mid (R = r, S = s)$ shifts upward in the first-order stochastic order as r rises, the map

$$r \longmapsto \mathbb{P}(D_{iJ}(1) = 1 \mid R = r, S_i = s)$$

is weakly increasing.

Lemma Lemma 1 therefore applies with selector $A = D_{iJ}(1)$ and score $T = R_{iJ}$, conditional on $S_i = s$. It yields

$$(R_{iJ} \mid D_{iJ}(1) = 1, S_i = s) \succeq_{FOSD} (R_{iJ} \mid D_{iJ}(1) = 0, S_i = s).$$

The AUC conclusion then follows exactly as in the proof of Proposition 4:

$$\mathrm{AUC}_{s,N} \geq \frac{1}{2}.$$

□

D Details on Empirical Methods

D.1 Reduced-Form Analysis

This Appendix describes the implementation of the Bayesian reduced-form model in Section 3.3. It covers the hierarchy and priors, the construction of probability-scale treatment effects and figures, the distributions across observed and counterfactual sittings, and the computational diagnostics.

Model and covariates I estimate the model in Equation (8) at the student-question level. Repeated participation by the same physical student in different years is treated as a separate student observation. The covariance matrices are parameterized as

$$\Sigma_S = \operatorname{diag}(\sigma_S) R_S \operatorname{diag}(\sigma_S), \qquad \Sigma_Q = \operatorname{diag}(\sigma_Q) R_Q \operatorname{diag}(\sigma_Q).$$

The student design X_i contains an indicator for female, standardized school mathematics grade and its missingness indicator, standardized prior participation, an indicator for the upper grade within age level, an indicator for non-public school, and a school-type missingness indicator. Missing mathematics grade is coded as zero on the standardized scale and accompanied by its missingness indicator. Missing school type is assigned the reference category and likewise accompanied by a missingness indicator.

The sitting design G_b contains an intercept, indicators for 2017 and 2018, and indicators for Levels 2–4; no year-by-level interactions are included. Residual sitting innovations are independent across the four components of m_b.

The question design W_j contains mathematical domain, context presence, reasoning-path structure, curricularity, answer format, image presence, standardized log stem length, an indicator for near-duplicate answer options, standardized position within block, and correct-answer letter.

Priors and posterior computation The prior specification is

$$M_{kl} \sim N(0, 1.5^2), \qquad B_{kl}, \gamma_{O,l}, \gamma_{H,l} \sim N(0, 0.75^2),$$

$$\sigma_S, \sigma_Q \sim \text{HalfNormal}(0, 1), \qquad \tau \sim \text{HalfNormal}(0, 0.5),$$

and

$$R_S, R_Q \sim \text{LKJcorr}(2).$$

The standardized innovations z_i, ζ_b, and q_j have independent standard Normal priors.

Estimation I estimate the model in Stan using non-centered parameterizations for the student, sitting, and item hierarchies and Cholesky-factor parameterizations for Σ_S and Σ_Q.

I use four independent chains, each with 1,000 warmup and 2,000 retained iterations, for 8,000 retained posterior draws. Sampling uses `adapt_delta=0.90` and `max_treedepth=10`. All reported posterior summaries use the full set of draws; subsampling is used only for the graphical and diagnostic calculations described below.

Probability-scale treatment effects Within each posterior draw m, I evaluate every student under both scoring regimes over the same 25-question reference set. For $z \in \{0, 1\}$,

$$o_{ij}^{(m)}(z) = \text{logit}^{-1}\left[a_i^{O,(m)} + \left(z - \tfrac{1}{2}\right) t_i^{O,(m)} + W_j' \gamma_O^{(m)} + u_j^{O,(m)}\right],$$

and

$$h_{ij}^{(m)}(z) = \text{logit}^{-1}\left[a_i^{H,(m)} + \left(z - \tfrac{1}{2}\right) t_i^{H,(m)} + W_j' \gamma_H^{(m)} + u_j^{H,(m)}\right].$$

Let $\mathcal{J}_{b(i)}$ denote the 25-question reference set associated with student i's sitting. Define

$$O_i^{(m)}(z) = \frac{1}{25} \sum_{j \in \mathcal{J}_{b(i)}} o_{ij}^{(m)}(z),$$

and

$$H_i^{(m)}(z) = \frac{\sum_{j \in \mathcal{J}_{b(i)}} [1 - o_{ij}^{(m)}(z)] h_{ij}^{(m)}(z)}{\sum_{j \in \mathcal{J}_{b(i)}} [1 - o_{ij}^{(m)}(z)]}.$$

The corresponding student-level effects are

$$\Delta O_i^{(m)} = O_i^{(m)}(1) - O_i^{(m)}(0), \qquad \Delta H_i^{(m)} = H_i^{(m)}(1) - H_i^{(m)}(0).$$

Measures of individual heterogeneity are calculated across students within each posterior draw and then summarized across draws. In particular,

$$p_{\text{fail}}^{(m)} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left\{\Delta O_i^{(m)} > 0,\ \Delta H_i^{(m)} \leq 0\right\}.$$

The standard deviations and correlation of ΔO_i and ΔH_i are calculated in the same way.

Joint distribution of individual effects Within posterior draw m, the observed-population empirical distribution is

$$F_N^{(m)} = \frac{1}{N} \sum_{i=1}^{N} \delta_{(\Delta O_i^{(m)}, \Delta H_i^{(m)})}.$$

The black contours in Panel A of Figure 2 represent the posterior mixture

$$E\left[F_N^{(m)} \mid \mathcal{D}\right],$$

rather than the distribution of student-specific posterior means.

To construct the figure, I retain approximately 25 student effects from each posterior draw and pool them across draws. A fixed random subsample of 100,000 student-draw observations is then used to estimate a bivariate Gaussian kernel density for $(100\Delta O, 100\Delta H)$. The bandwidth in each dimension is the normal-reference bandwidth multiplied by 1.15, and the density is evaluated on a 260×260 grid over $[-100, 100]^2$. Contour thresholds define highest-density regions containing 25, 45, 65, 82, and 93 percent of the smoothed posterior-mixture mass. The black point reports the posterior mean of the observed-population average treatment effects.

The grey circles show the raw student-level changes. Raw $(\Delta O_i, \Delta H_i)$ observations are grouped into a rectangular grid over the displayed region, with circle area increasing in the number of observations in each bin. This binning is used only to make overlapping raw observations visible.

Sorting gain For posterior draws satisfying $\Delta O_i^{(m)} > 0.01$, define

$$\Gamma_i^{(m)} = \frac{1 - O_i^{(m)}(0)}{\Delta O_i^{(m)}} \Delta H_i^{(m)}.$$

The cutoff avoids numerical instability when the omission response is arbitrarily close to zero and is not imposed on the sign-based RPM statistic. Let

$$I_i^{(m)} = \mathbf{1}\{\Delta O_i^{(m)} > 0.01\}.$$

The within-draw sorting-gain distribution is

$$G_N^{(m)} = \frac{\sum_{i=1}^N I_i^{(m)} \delta_{\Gamma_i^{(m)}}}{\sum_{i=1}^N I_i^{(m)}}.$$

The black density in Panel B of Figure 2 represents the posterior mixture $E[G_N^{(m)} \mid \mathcal{D}]$. For display, the raw histogram and posterior density are restricted to $[-1, 1]$ and normalized over that interval. Numerical posterior summaries use the full, untruncated distribution of defined sorting gains.

Observed-sitting effects The partially pooled treatment effect for an observed sitting b is obtained, within each posterior draw, by averaging the individual probability-scale effects of students in that sitting:

$$\Delta O_b^{(m)} = \frac{1}{N_b} \sum_{i:b(i)=b} \Delta O_i^{(m)}, \qquad \Delta H_b^{(m)} = \frac{1}{N_b} \sum_{i:b(i)=b} \Delta H_i^{(m)}.$$

These quantities retain the fitted sitting innovation $\zeta_b^{(m)}$ and therefore describe the 12 realized test environments after partial pooling. Note that they differ from the parent distribution below, which integrates over a new residual sitting shock.

Parent distribution across sittings The upper hierarchy defines a distribution of treatment effects across comparable test environments. For posterior draw m, year-by-level cell b, and outcome margin $q \in \{O, H\}$, let $T_{bq}^{(m)}(\xi)$ denote the average probability-scale treatment effect after replacing the fitted sitting innovation by $\xi \sim N_4(0, I)$. Student composition, posterior student deviations, the 25-item reference set, item covariates, and item effects are otherwise held at their observed values within the cell.

The resulting parent distribution is the equal-weight mixture

$$\mathcal{P}_q^{(m)} = \frac{1}{12} \sum_{b=1}^{12} \mathcal{L}\left[T_{bq}^{(m)}(\Xi) \mid \Xi \sim N_4(0, I)\right].$$

Its mean and standard deviation are

$$\mu_{P,q}^{(m)} = E_{\mathcal{P}_q^{(m)}}[T], \qquad \sigma_{P,q}^{(m)} = \sqrt{\mathrm{Var}_{\mathcal{P}_q^{(m)}}(T)}.$$

The parent variance combines residual sitting heterogeneity with systematic variation over the observed year-by-level support:

$$\mathrm{Var}_{\mathcal{P}_q^{(m)}}(T) = E_b[\mathrm{Var}(T \mid b, m)] + \mathrm{Var}_b[E(T \mid b, m)].$$

The reported parent standard deviation is therefore not a direct probability-scale transformation of τ_q.

I evaluate the Gaussian integral using a 16-node tensor Gauss–Hermite rule and repeat the calculation on a dispersed subset of posterior draws using an 81-node rule. Table A19 reports the comparison. The maximum absolute discrepancy is below 0.01 percentage points for either parent mean and below 0.27 percentage points for either parent standard deviation.

Prediction for a new sitting Prediction for a new sitting additionally integrates over a realized residual sitting shock. Within each posterior draw and each of the 12 observed year-by-level cells, I draw

$$\Xi_b^{(m)} \sim N_4(0, I)$$

and set

$$m_b^{\mathrm{new},(m)} = M^{(m)} G_b + \mathrm{diag}(\tau^{(m)}) \Xi_b^{(m)}.$$

I then recompute the probability-scale treatment effect using the observed student composition and item reference set of that cell. Equal-weight pooling across the 12 year-by-level

cells and posterior draws gives the distribution reported as "New-sitting prediction" in the Appendix hierarchy table.

The observed-population average treatment effect and the parent mean use different weighting schemes. The former weights student-sittings by their empirical frequency; the latter gives equal weight to the 12 year-by-level cells and integrates over their residual sitting distribution.

MCMC diagnostics Table A20 reports diagnostics from all 8,000 retained draws. The production fit has zero divergent transitions and zero maximum-treedepth hits. The maximum rank-normalized $\widehat{R}$ is 1.005, the minimum bulk effective sample size is 750, the minimum tail effective sample size is 1,315, and the minimum chain-level E-BFMI is 0.664. Appendix Figures A12 and A13 report selected trace and rank plots.

Prior and posterior predictive checks Prior predictive simulations propagate draws from the full prior through the observed student and item design matrices and generate omission and conditional-correctness probabilities under both scoring regimes. The checks therefore assess the joint implications of the prior on the probability scale rather than only at $X = W = 0$.

Posterior predictive checks use 200 posterior draws and complete replicated response datasets simulated conditional on the fitted hierarchy. Appendix Figures A14 and A15 compare observed and replicated rates by sitting and scoring regime. All 24 sitting-by-regime omission rates and all 24 conditional hit-rates lie within their central 95 percent posterior predictive intervals.

Appendix Figure A16 compares observed and replicated dispersion across students and items. Seven of the eight dispersion statistics lie within their central 95 percent posterior predictive intervals. The exception is cross-item hit-rate dispersion under the omission-reward regime, which lies slightly above the upper posterior-predictive bound.

D.2 Structural Estimation and Policy Implications

This Appendix describes the implementation of the structural model in Section 4 and the construction of the posterior objects used in Section 5. I first describe the likelihood and priors, then computation and diagnostics, and finally the construction of the structural and policy figures.

Likelihood and Implementation The estimation fixes $\underline{k} = 0.20$ and $\overline{k} = 1$. Let $R_{ij} \in \{O, W, C\}$ denote omission, an attempted incorrect answer, and an attempted correct answer. In Part 1, selected student-sittings have no omissions, so the probability of a correct answer is

$$\Pr(R_{ij} = C \mid z = 0) = 0.20 + 0.80\pi_{ij}.$$

In Part 2, the three outcome probabilities are

$$\begin{aligned}
\Pr(R_{ij} = O \mid z = 1) &= \pi_{ij}(1 - \psi_{ij}^{+}) + (1 - \pi_{ij})\psi_{ij}^{-}, \\
\Pr(R_{ij} = C \mid z = 1) &= \pi_{ij}\psi_{ij}^{+} + 0.20(1 - \pi_{ij})(1 - \psi_{ij}^{-}), \\
\Pr(R_{ij} = W \mid z = 1) &= 0.80(1 - \pi_{ij})(1 - \psi_{ij}^{-}).
\end{aligned}$$

Part-1 observations therefore enter through a Bernoulli likelihood, while Part-2 observations enter through the three-category likelihood above.

The student design X_i contains gender, school mathematics grade, number of prior participations, an indicator for the upper grade within an age level, and an indicator for non-public school attendance. All covariates are centered within sitting; mathematics grade is also standardized by its within-sitting standard deviation. Missing mathematics grades are replaced by the corresponding sitting mean and missing school type by the sitting mode.

The item design W_j contains indicators for mathematical domain, context, reasoning-path structure, curricularity, answer-option format, and correct-answer position, together with controls for the presence of an image, log stem length, near-duplicate answer options, and position within the test block. The resulting design has 22 columns. Categorical indicators and the image and near-duplicate controls are centered so that $W_j = 0$ represents the average retained item. Stem length is standardized globally and position within regime.

Residual item effects are centered separately within each sitting-by-regime block. For a block with J_{bz} retained items, the centered standard-normal residual is multiplied by $(1 - 1/J_{bz})^{-1/2}$ before applying its estimated scale. Residual item effects therefore sum exactly to zero within block while retaining unit marginal variance before scaling.

Priors, Computation, and Diagnostics The prior specification is common across the three structural indices $r \in \{\pi, +, -\}$. Grand means and covariate coefficients satisfy

$$\mu_r \sim N(0, 1.25^2), \qquad \beta_r \sim N(0, 0.50^2), \qquad \gamma_r \sim N(0, 0.40^2).$$

Year and level effects receive sum-to-zero Gaussian priors with marginal standard deviations 0.40 and 0.55, respectively. Residual sitting standard deviations satisfy

$$\tau_r \sim N^{+}(0, 0.35^2).$$

Writing $\Sigma_S = D_S R_S D_S$, the student standard deviations and correlation matrix have priors

$$\sigma_{S,r} \sim N^{+}(0, 0.80^2), \qquad R_S \sim \mathrm{LKJ}(2),$$

while residual item standard deviations satisfy

$$\sigma_{Q,r} \sim N^{+}(0, 0.50^2).$$

Here N^{+} denotes a normal distribution restricted to the positive half-line. All standardized sitting, student, and item shocks have independent standard-normal priors. Student and sitting effects are implemented in non-centered form.

I estimate the model in Stan using four independent chains, each with 1,000 warm-up and 6,000 retained iterations, for 24,000 retained draws in total. Sampling uses `adapt_delta`=0.95 and maximum tree depth 12. Likelihood evaluation is parallelized with `reduce_sum`, using a grain size of 500 and eight threads per chain. The full chains are used for convergence diagnostics. Paper results use a joint subset of 1,200 draws spaced approximately uniformly through the combined posterior. Parameters and local effects are always taken from the same draw, preserving posterior dependence.

The production run produced no divergent transitions or maximum-tree-depth hits. The minimum E-BFMI across chains is 0.625. Across the full parameter vector, the maximum $\widehat{R}$ is 1.0116, the minimum bulk and tail effective sample sizes are 174 and 254, and the maximum ratio of Monte Carlo standard error to posterior standard deviation is 0.077. The weakest mixing occurs in the Cholesky terms governing the student-level association between ψ^{+} and ψ^{-}, and in the corresponding implied correlation. Trace and rank plots for the main hierarchical parameters, together with the complete parameter-level diagnostic output, are reported with the additional Appendix diagnostics.

Prior predictive checks use 6,000 draws from the prior and the actual centered student and item design rows. For each draw, sampled parameters are propagated through the structural probabilities for Part-1 correctness, Part-2 omission, and Part-2 correctness conditional on attempt. These checks therefore reflect the joint scale of the full prior specification rather than an artificial $X = W = 0$ case.

Posterior predictive checks use 80 posterior draws and the fitted student, sitting, and item effects. At each draw, I simulate complete response data and compare the observed sample with its replicated distribution for Part-1 hit rates, Part-2 omission and conditional hit rates, the shares of students answering or omitting all retained Part-2 questions, student-level rate distributions, sitting-level rates, and retained-item calibration.

The model-fit exercise reported in Figure A8 uses 200 posterior draws and the fitted, rather than parent-standardized, student and item effects. Student-profile cells are defined by gender, mathematics-grade quintile, and prior-participation status, giving 20 profiles. Item cells are defined by sitting and mathematical domain. Within each cell, I compare observed rates with posterior mean predictions for Part-1 correctness, Part-2 omission, and Part-2 correctness conditional on attempt. Conditional correctness is computed as the sum of predicted correct probabilities divided by the sum of predicted attempt probabilities

within the cell. The RMSEs reported in the main text are unweighted across cells. A separate calibration check applies the experimental 5/1/0 scoring rule and compares observed and predicted average Part-2 scores across the same 20 student profiles.

Construction of Structural Objects The structural figures describe the *parent population* implied by the hierarchy rather than the empirical distribution of fitted student effects. I construct this population separately within each posterior draw. I retain the observed distribution of sittings and student covariates, set the residual sitting shock to its parent mean, $\zeta_b^r = 0$, and replace each fitted student residual by

$$\widetilde{s}_i = D_S R_S^{1/2} z_i, \qquad z_i \sim N(0, I_3).$$

For numerical integration, I generate a fixed set of antithetic standard-normal draws z_i and $-z_i$ and reuse it across posterior draws. This integrates over the estimated parent distribution of student heterogeneity without adding simulation noise to posterior comparisons. Unless otherwise stated, structural population objects use the average-item convention, setting the item component to zero.

The three primitive probabilities are first obtained on the probability scale,

$$\pi_i = \Lambda(a_i^\pi), \qquad \psi_i^+ = \Lambda(a_i^+), \qquad \psi_i^- = \Lambda(a_i^-),$$

and then rotated into

$$\psi_i^\Sigma = \psi_i^+ + \psi_i^- - 1, \qquad \psi_i^\Delta = \psi_i^+ - \psi_i^-.$$

Population distributions For Figure 4, I compute the mean, standard deviation, quantiles, and pairwise correlations of $(\pi_i, \psi_i^\Sigma, \psi_i^\Delta)$ across the parent population within each posterior draw. These scalar quantities use the unsmoothed simulated population. For the displayed distributions, I estimate a kernel density within each draw using the default Gaussian bandwidth multiplied by 1.10, on a 161-point grid over $[0, 1]$ for knowledge and $[-1, 1]$ for sensitivity and bias. The solid line reports the pointwise posterior median density; the two bands report pointwise 50 and 95 percent posterior intervals.

Gender heterogeneity For Figure 5, I construct standardized male and female profiles for every student-sitting. Because gender is centered within sitting, its design value is set to $-\bar{F}_b$ under the male profile and $1 - \bar{F}_b$ under the female profile, where $\bar{F}_b$ is the observed female share in sitting b. All other covariates, the year-by-level composition, and the parent student-residual draw are held fixed. I transform each counterfactual profile to $(\pi, \psi^\Sigma, \psi^\Delta)$, average across the parent population, and repeat across posterior draws. Points in the figure report posterior means; thick and thin intervals contain 50 and 95 percent of posterior

draws.

Other observed heterogeneity Figure A9 uses the same standardization procedure. For mathematics grade, I compare -0.5 with $+0.5$ on the within-sitting standardized scale, a one-standard-deviation contrast. Prior participation compares zero with one previous participation; grade compares the lower with the upper grade within level; and school type compares public with non-public school. Each contrast changes only the covariate of interest, holding all other observed characteristics and the parent residual draw fixed. Reported differences are population averages of the individual predicted contrasts within each posterior draw.

Construction of Policy Objects Unlike the average-item structural figures above, the policy exercises evaluate each student against the actual retained Part-2 item set in the corresponding sitting. Let $\mathcal{J}_i$ denote this set and $J_i = |\mathcal{J}_i|$. Within each posterior draw, I reconstruct π_{ij} and ψ_{ij}^+, including estimated item covariates and residual item effects, and define

$$K_i = \frac{1}{J_i}\sum_{j\in\mathcal{J}_i}\pi_{ij}, \qquad M_i = \frac{\sum_{j\in\mathcal{J}_i}\pi_{ij}\psi_{ij}^+}{\sum_{j\in\mathcal{J}_i}\pi_{ij}}, \qquad W_i = \frac{1}{J_i}\sum_{j\in\mathcal{J}_i}\left(1+4\pi_{ij}\psi_{ij}^+\right) = 1+4K_iM_i.$$

Student effects use the same parent standardization as in Section D.2.

Score decoding The three quantities in Figure 6 are calculated separately within each posterior draw. For the variance decomposition, let $\bar{K}$ and $\bar{M}$ denote the corresponding population means and define

$$W_i^K = 1+4K_i\bar{M}, \qquad W_i^M = 1+4\bar{K}M_i.$$

With $V = \mathbb{V}(W_i)$, $V_K = \mathbb{V}(W_i^K)$, and $V_M = \mathbb{V}(W_i^M)$, the symmetric Shapley contributions are

$$\phi_K = \frac{1}{2}\left[V_K + (V - V_M)\right], \qquad \phi_M = \frac{1}{2}\left[V_M + (V - V_K)\right].$$

The metacognition share plotted in the figure is ϕ_M/V.

For the knowledge decoding loss, I approximate $\mathbb{E}[K_i \mid W_i]$ by sorting students on W_i, forming 50 equal-frequency bins, and assigning each student the mean K_i in their score bin. The residual sum of squares divided by the total sum of squares in K_i gives L_K. The additional Appendix reports the analogous reverse quantity L_M and repeats both calculations using 25 and 100 bins.

For ranking reversals, I use a fixed sample of 60,000 student pairs. The first student is sampled uniformly from the selected structural population and the second uniformly from a different student in the same sitting. The same pairs are used for every posterior draw.

Exact ties in either K or W are excluded, and a reversal occurs when

$$(K_i - K_{i'})(W_i - W_{i'}) < 0.$$

The supporting ranking figure divides valid pairs within each posterior draw into five equal-frequency groups according to $|K_i - K_{i'}|$ and reports the reversal probability in each group. All comparisons therefore hold the test environment fixed and use predicted scores rather than noisy realization.

Observed gender-score figure Panel A of Figure 7 is constructed directly from the data and does not use the selected structural sample. I retain observations with complete omission coding, at least one attempted question in each test part, and observed gender and school mathematics grade. The score uses all 12 Part-2 questions, assigning 5 points to a correct answer, 1 to an omission, and 0 to an incorrect answer, and is averaged across questions within student-sitting. Mathematics-grade quintiles are constructed separately within each sitting using average ranks. The plotted points are gender-specific cell means. Their 95 percent confidence intervals use CR1 standard errors clustered by physical participant, allowing repeated appearances of the same participant across sittings.

Structural gender-score decomposition Panel B of Figure 7 uses the selected structural population and the retained Part-2 item environment. Within each posterior draw, I construct the standardized male and female student profiles described above. Let

$$W_i^{ab} = \frac{1}{J_i} \sum_{j \in \mathcal{J}_i} \left(1 + 4\pi_{ij}^{a} \psi_{ij}^{+,b}\right), \qquad a, b \in \{F, M\},$$

where the first superscript indexes the gender profile used for knowledge and the second the profile used for recognition of knowledge. The two Shapley components are

$$\Delta_K = \frac{1}{2}\mathbb{E}_i\left[(W_i^{MF} - W_i^{FF}) + (W_i^{MM} - W_i^{FM})\right],$$
$$\Delta_M = \frac{1}{2}\mathbb{E}_i\left[(W_i^{FM} - W_i^{FF}) + (W_i^{MM} - W_i^{MF})\right].$$

By construction, $\Delta_K + \Delta_M = \mathbb{E}_i[W_i^{MM} - W_i^{FF}]$. The "metacognition" component in this decomposition therefore refers specifically to recognition of knowledge, ψ^+, the metacognitive primitive that enters expected formula scores under the experimental calibration.

Latent-knowledge schedules and Oaxaca–Blinder decomposition The supporting gender schedule and Oaxaca–Blinder decomposition use the same standardized female and male structural populations. Within each posterior draw, I pool female and male latent-knowledge values and divide the resulting $2N$ observations into five equal-frequency

bins. The same cutoffs are then applied to both gender profiles. For each bin b, $m_g(b)$ is the mean predicted score for gender profile g, and $p_g(b)$ is that profile's population share. The latent-knowledge schedule plots $m_F(b)$ and $m_M(b)$ across these pooled quintiles.

The Oaxaca–Blinder decomposition reported in Figure A10 uses the female knowledge distribution to weight the within-knowledge score-schedule component and the male score schedule to value the knowledge-composition component, as defined in the main text. The baseline uses five pooled knowledge bins. I repeat the decomposition using 10 and 20 bins as a discretization check.

Unless otherwise noted, posterior structural and policy figures use the same 1,200 joint posterior draws. Posterior points report means across draws; thick and thin intervals report 50 and 95 percent posterior intervals. Density figures instead report the pointwise posterior median and pointwise 50 and 95 percent bands. The observed gender-score figure is the sole exception: its uncertainty is frequentist and clustered by physical participant.

D.3 Rasch Calibration of Item Easiness

This Appendix describes the item-easiness calibration used in the within-regime test. The target is the relative ordering of the twelve Part 2 items within each of the twelve sittings. I calibrate this ordering using the 903 students who attempt every question in the test, for whom correctness is observed on every Part 2 item.

Conditional Rasch model Let $Y_{ibj} \in \{0, 1\}$ indicate whether full completer i answers Part 2 item $j \in \{1, \ldots, 12\}$ correctly in sitting b. I specify

$$\Pr(Y_{ibj} = 1 \mid \alpha_{ib}, e_{bj}) = \Lambda(\alpha_{ib} + e_{bj}),$$

where α_{ib} is student ability, e_{bj} is item easiness, and

$$\Lambda(x) = \frac{\exp(x)}{1 + \exp(x)}.$$

Larger values of e_{bj} correspond to easier items. Because the coefficient on ability is common across items,

$$e_{bj} > e_{bk} \quad \Longrightarrow \quad \Pr(Y_{ibj} = 1 \mid \alpha_{ib}) > \Pr(Y_{ibk} = 1 \mid \alpha_{ib})$$

for every value of α_{ib}. The Rasch model therefore delivers a single ability-invariant ranking of the twelve items.

Let

$$C_{ib} = \sum_{j=1}^{12} Y_{ibj}$$

denote the student's Part 2 total score. For a response vector y containing c correct answers,

conditioning on $C_{ib} = c$ eliminates the student-specific ability parameter:

$$\Pr(Y_{ib} = y \mid C_{ib} = c, e_b) = \frac{\exp\left(\sum_{j=1}^{12} y_j e_{bj}\right)}{\sum_{a \in \{0,1\}^{12}: \sum_j a_j = c} \exp\left(\sum_{j=1}^{12} a_j e_{bj}\right)}.$$

The reason is that the term involving ability, $\alpha_{ib} \sum_j y_j = \alpha_{ib} c$, is common to every response pattern with total score c and cancels from the conditional likelihood. Equivalently, for two equal-score patterns that differ only in whether item j or item k is answered correctly,

$$\frac{\Pr(Y_{ibj} = 1, Y_{ibk} = 0 \mid C_{ib}, e_b)}{\Pr(Y_{ibj} = 0, Y_{ibk} = 1 \mid C_{ib}, e_b)} = \exp(e_{bj} - e_{bk}).$$

Relative item easiness can therefore be estimated without specifying or estimating the distribution of student ability.

Only differences in e_{bj} are identified: adding the same constant to every item easiness parameter leaves the conditional likelihood unchanged. I therefore impose the location normalization

$$\sum_{j=1}^{12} e_{bj} = 0.$$

This normalization fixes the origin of the easiness scale but has no effect on the item ranking.

Students with $C_{ib} = 0$ or $C_{ib} = 12$ have only one feasible response pattern conditional on their total score and therefore contribute no information about relative item easiness. Of the 903 full completers, 879 have informative Part 2 scores. I evaluate the exact conditional likelihood using item-level correct-answer totals and the number of students at each score $c = 1, \ldots, 11$.

Bayesian estimation To quantify uncertainty in the estimated item ranking, I combine the exact conditional likelihood with a weakly regularizing Gaussian prior on the sum-to-zero easiness vector,

$$p(e_b) \propto \prod_{j=1}^{12} \exp\left\{-\frac{e_{bj}^2}{2(1.5)^2}\right\}, \qquad \sum_{j=1}^{12} e_{bj} = 0.$$

The Bayesian layer is used to propagate finite-calibration-sample uncertainty in relative item easiness. It does not provide the identifying restriction: identification of relative easiness comes from the conditional Rasch structure above.

Each sitting is estimated separately in Stan using four chains, 1,000 warm-up iterations and 1,000 retained iterations per chain, `adapt_delta`=0.95, and maximum tree depth 12. All twelve fits have zero divergent transitions and zero maximum-tree-depth hits. Across all item parameters and sittings, the largest $\widehat{R}$ is 1.005, while the minimum bulk and tail

effective sample sizes are 3,530 and 2,159, respectively.

Pairwise ranking probabilities For each pair of Part 2 items j and k, define

$$p_{bjk} := \Pr(e_{bj} > e_{bk} \mid \mathcal{C}_b),$$

where $\mathcal{C}_b$ denotes the full-completer calibration data in sitting b. I estimate p_{bjk} as the fraction of posterior draws for which $e_{bj} > e_{bk}$. These probabilities enter the measurement-aware AUC:

$$\widetilde{\mathrm{AUC}}_{ib} = \frac{1}{|\mathcal{A}_{ib}||\mathcal{M}_{ib}|} \sum_{j \in \mathcal{A}_{ib}} \sum_{k \in \mathcal{M}_{ib}} p_{bjk},$$

where $\mathcal{A}_{ib}$ and $\mathcal{M}_{ib}$ are the sets of questions attempted and omitted by student i, respectively. Equivalently,

$$\widetilde{\mathrm{AUC}}_{ib} = \mathbb{E}[\mathrm{AUC}_{ib}(e_b) \mid \mathcal{C}_b]\,.$$

Thus the soft AUC is the posterior expectation of the AUC that would be computed if the latent item-easiness vector were known.

The rankings are generally precise. Define posterior confidence in the ordering of a pair as

$$q_{bjk} := \max\{p_{bjk}, 1 - p_{bjk}\}.$$

Across the

$$12 \times \binom{12}{2} = 792$$

within-sitting item pairs, the median q_{bjk} is 0.985; 68 percent of pairs have $q_{bjk} \geq 0.90$, and 61 percent have $q_{bjk} \geq 0.95$. The median posterior confidence exceeds 0.93 in every sitting. The soft AUC therefore differs materially from a deterministic ranking mainly for pairs whose relative easiness is weakly determined by the calibration sample.

Plug-in benchmark For comparison, I also maximize the same exact conditional Rasch likelihood and construct a plug-in AUC that treats the estimated item ranking as known. The posterior-mean Bayesian and plug-in rankings are nearly identical: their equal-sitting mean Spearman correlation is 0.995. The corresponding parent AUC is 0.643, compared with 0.634 for the measurement-aware soft AUC. I use the latter throughout the main analysis because it carries uncertainty about the item ranking into the sorting measure.

The posterior uncertainty described here is conditional on the Rasch measurement model. In particular, propagating uncertainty in e_b addresses uncertainty arising from the finite calibration sample; it does not by itself establish the common-discrimination restriction of the Rasch model or the transportability of the full-completer ranking to students who omit. The latter is assessed separately using the Part 1 validation exercise described in the main text.

E Details on Item Annotation

This Appendix describes the construction of the item characteristics used in the empirical analysis. The annotation corpus contains all 300 competition questions. For each item, the source material consists of the student-visible stem and five answer alternatives, the official correct-answer letter and proposed solution, and any associated question or solution images. The pipeline separates deterministic measurements from semantic annotations produced by a large language model (LLM), and uses repeated annotation runs to assess the stability of the latter.

Source preparation and deterministic measures Before LLM annotation, I normalize the parsed source text and construct characteristics that can be measured mechanically. These include stem and solution length, sentence and line counts, counts of mathematical-expression spans, question and solution image counts, correct-answer position, and position within the test block. I also construct rule-based classifications of answer-option format, dictionary-based candidate flags for meta-options, and surface-similarity flags for potentially near-duplicate alternatives. Source preflight checks identify missing or malformed text and image inputs. None of these quantities is elicited from the LLM.

The final answer-option format combines deterministic and LLM information. When all five alternatives are text-readable and nonvisual, the deterministic classification is authoritative. The LLM classification is used for visual or otherwise unresolved answer menus. The remaining semantic option characteristics, including the presence of meta-options and near-duplicate alternatives, are obtained from the annotation procedure described below.

Leakage-controlled LLM annotation Each item is processed by three separate annotation modules, with the information supplied to each module restricted to what is needed for its task. All modules use a frozen ontology, closed response categories, and strict structured-output schemas.

The *context module* receives only the student-visible stem, answer alternatives, and question images. It classifies whether the item is formal or contextualized and, when contextualized, assigns a primary context. It does not receive the official answer or solution, year, competition level, question position, or any information on student performance or behavior.

The *mathematics module* receives the stem, alternatives, official answer, proposed solution, and question and solution images. It classifies the primary mathematical domain, identifies up to three minimal prerequisite concepts, and labels the reasoning-path structure, principal structural device, and quality of the official solution. The module does not receive the competition level. In particular, curricularity is not elicited from the model: the prompt asks for the least advanced prerequisite set required by the official solution without revealing

whether those prerequisites are expected at the item's competition level.

The *options module* receives the answer alternatives, relevant question images, and only the stem information needed to interpret the role of those images. It classifies semantic answer-option format and identifies meta-options and near-duplicate alternatives. As above, it receives no empirical information on item difficulty, correctness, omission behavior, or student characteristics.

Across all three modules, the actual item identifier is replaced by the generic identifier `ITEM`. The model therefore cannot infer year, level, or test position from the identifier. No annotation module receives empirical correctness rates, omission rates, estimated treatment effects, gender differences, or other outcomes subsequently analyzed in the paper.

Model and repeated runs The complete corpus is annotated three times using byte-identical source inputs, prompts, schemas, ontology, lookup tables, and model configuration. All three runs use `gpt-5.6-sol` through the OpenAI Responses API in Batch mode, with high-detail image inputs and strict JSON-schema outputs. The context and options modules use low reasoning effort and the mathematics module uses medium reasoning effort. The frozen pipeline and ontology versions are 1.3.0 and 1.2, respectively.

Each run contains $300 \times 3 = 900$ requested module outputs. Across the three repeated runs, 2,692 of 2,700 outputs parse successfully. The eight failures are incomplete responses caused by reaching the maximum output-token limit. Because these failures occur in different items and modules across runs, they do not leave any item without a final annotation.

The three runs are stochastic repetitions of the same frozen annotation procedure rather than independent human raters. Their agreement therefore measures the test–retest stability of the measurement procedure; it does not by itself rule out systematic classification error shared across runs.

Consensus and derived measures For single-label and binary LLM fields, the final annotation is the two-of-three majority across repeated runs. If one output is missing, agreement between the two available runs determines the label. Cases without a mechanical majority are sent to a separate adjudication pass using the same frozen codebook and the same module-specific information restrictions. I then manually review the adjudicated decision against the source item and codebook before accepting it. All original run-level annotations and adjudication decisions are retained in the audit files.

Prerequisite concepts are aggregated separately tag by tag: a prerequisite is retained when it appears in at least two of the three runs. One item requires adjudication of the prerequisite set. The resulting consensus prerequisites are then mapped through a frozen curriculum crosswalk to the earliest competition level at which each prerequisite is expected. Let L_j^{req} denote the highest required level among the consensus prerequisites and L_j the

observed competition level. Curricularity is defined mechanically as

$$\text{Curricularity}_j = \begin{cases} \text{prior curriculum}, & L_j^{\text{req}} < L_j, \\ \text{current curriculum}, & L_j^{\text{req}} = L_j, \\ \text{beyond curriculum}, & L_j^{\text{req}} > L_j. \end{cases}$$

Thus competition level never enters the LLM's assessment of mathematical prerequisites; it is introduced only after annotation through this deterministic comparison. Re-deriving curricularity from the final consensus prerequisite set reproduces the final curricularity label for all 300 items.

Gender coding is constructed analogously rather than elicited directly from the model. The context module first determines the item's primary context. A frozen context-to-stereotype mapping then assigns that context to *female-coded*, *male-coded*, or *neutral*. Character names or genders therefore cannot directly determine the stereotype measure. Boundary, mixed, and residual contexts carry a separate uncertainty flag.

Adjudication and uncertainty flags Mechanical consensus resolves nearly all fields. Eight of the 300 items require adjudication of at least one field, covering ten field-level decisions in total. These cases concern prerequisite tags, primary context or its derived gender coding, structural devices, official-solution quality, and option diagnostics. The adjudication LLM re-reads the permitted source material under the frozen module rules and resolves the disagreement without access to empirical outcomes; I subsequently review each resolution manually. No final analytical label remains unresolved.

The pipeline separately preserves source-quality flags, category-boundary flags, run-level review flags, and consensus agreement counts. These indicators remain part of the audit trail rather than being used to overwrite disagreement mechanically.

Repeated-run reliability Agreement across the three frozen runs is high for the principal analytical variables. Exact three-run agreement is 95.7 percent for context presence, 96.3 percent for mathematical domain, 95.0 percent for curricularity, 87.2 percent for reasoning-path structure, and 85.2 percent for structural device. The corresponding Fleiss' κ statistics are 0.937, 0.969, 0.927, 0.854, and 0.836, respectively. Primary context has 90.0 percent exact agreement and $\kappa = 0.878$, while the derived gender-stereotype classification has 96.3 percent agreement and $\kappa = 0.899$. Final option format agrees across all three runs for 99.3 percent of items.

Prerequisite sets are naturally less stable because they are multi-label objects. Exact agreement on the complete prerequisite set is 63.4 percent, while mean pairwise Jaccard similarity is 0.874 and the median is one for each pair of runs. The derived curricularity measure is substantially more stable than the underlying exact tag set. The replication files

report field-level raw agreement, all pairwise Cohen's κ coefficients, Fleiss' κ, consensus rates, and the complete set of non-unanimous annotations.

These statistics assess repeated-run reliability rather than agreement with an external human gold standard. Systematic errors induced by the common model, ontology, or prompts may therefore persist across all three runs. The construction mitigates this concern through information restrictions, deterministic derivation where possible, explicit uncertainty flags, and retention of the complete annotation and adjudication audit trail.

Reproducibility The replication package stores the frozen prompts and JSON schemas for each module, the complete ontology and decision rules, the curriculum and stereotype lookup tables, source and image checksums, run-level configuration and metadata, all three raw annotations, parse errors, consensus and non-unanimity files, adjudication records, and the final analysis-ready item file. The final dataset contains one record for each of the 300 questions and preserves the provenance of each variable as source-derived, deterministic, LLM consensus, deterministically derived from an LLM annotation, or quality-assurance metadata.

F Tables and Figures

Table A1: Student Covariate Descriptives for Reduced-Form and Structural Samples

	Main reduced-form sample			Structural sample		
Variable	N	Mean	SD	N	Mean	SD
Female	9,807	0.341	0.474	6,789	0.343	0.475
School math grade (standardized)	8,838	0.029	0.954	6,115	0.037	0.948
Prior participations before current contest	9,807	0.161	0.420	6,789	0.169	0.427
Any prior participation	9,807	0.141	0.348	6,789	0.148	0.355
Upper grade within contest level	9,807	0.558	0.497	6,789	0.549	0.498
Non-public school	9,740	0.504	0.500	6,750	0.499	0.500
Risk measure (1–5)	5,261	2.043	1.108	3,642	1.964	0.992
School math grade missing	9,807	0.099	0.298	6,789	0.099	0.299
School type missing	9,807	0.007	0.082	6,789	0.006	0.076
Risk measure missing	9,807	0.464	0.499	6,789	0.464	0.499

Notes: The main analysis sample contains 9,807 students and the structural sample 6,789; the row-specific N reports nonmissing observations. Binary-variable means are proportions. Missingness indicators are shown because missing school-math and school-type indicators enter the reduced-form Bayesian specification, while risk missingness documents the complete-case scope of the frequentist controls robustness.

Table A2: Item-Characteristic Balance across Scoring-Rule Blocks

Family	Item characteristic	No-reward mean	Formula mean	Std. diff.	Permutation p
Question complexity	Log(1 + stem words)	3.370	3.371	0.002	0.985
	Log(1 + stem math spans)	1.035	1.005	-0.050	0.684
	Question image present	0.368	0.278	-0.197	0.109
Option format	Option format: mixed	0.069	0.097	0.102	0.403
	Option format: symbolic	0.021	0.075	0.264	0.031
	Option format: textual	0.049	0.042	-0.032	0.816
	Option format: visual	0.007	0.005	-0.028	1.000
Answer key	Correct answer: B	0.174	0.147	-0.075	0.544
	Correct answer: C	0.195	0.227	0.080	0.511
	Correct answer: D	0.252	0.190	-0.153	0.205
	Correct answer: E	0.158	0.195	0.099	0.418
Domain	Domain: number theory digits	0.189	0.171	-0.046	0.711
	Domain: algebra functions sequences	0.189	0.219	0.078	0.522
	Domain: geometry measurement spatial	0.308	0.301	-0.016	0.898
	Domain: discrete reasoning	0.124	0.161	0.105	0.392
Content	Contextualized item	0.414	0.354	-0.137	0.267
Curricularity	Curricularity: current curriculum	0.262	0.289	0.065	0.599
	Curricularity: beyond curriculum	0.035	0.094	0.253	0.038
Reasoning path	Reasoning: decomposition aggregation	0.248	0.160	-0.220	0.076
	Reasoning: case enumeration	0.089	0.089	0.000	0.999
	Reasoning: constraint elimination	0.138	0.098	-0.127	0.315
Content	Any structural device	0.378	0.400	0.046	0.711
Option format	Near-duplicate options	0.064	0.079	0.060	0.630
Solution quality	Official-solution quality flag	0.131	0.119	-0.036	0.765

Notes: Positive differences mean formula-scoring minus no-reward. Statistics first compare the 13 and 12 items within each year-by-level sitting, then average sitting differences using the student composition of the 9,807-student main sample. Standardized differences divide by the corresponding pooled within-sitting item SD. Unadjusted p-values use 50,000 within-sitting permutations preserving 13 versus 12 items. Item placement is not documented as randomized, so these are structured permutation-reference diagnostics, not literal design-based randomization inference. The single joint permutation p-value for all listed covariates is 0.370. Reference categories for categorical controls are omitted from the table. Position within block is excluded because it is a deterministic order control, not an item-composition characteristic.

Table A3: Pooled Regime-Effect Inference Robustness Summary

Outcome	Inference check	Estimate	SE	95% CI	p
Omission rate	Main student + item CV1	34.68	1.67	[31.39, 37.96]	<0.001
Hit rate among attempts	Main student + item CV1	2.59	2.40	[-2.14, 7.32]	0.282
Omission rate	Two-way CV3 maximum-SE rule	34.68	1.74	[31.25, 38.10]	<0.001
Hit rate among attempts	Two-way CV3 maximum-SE rule	2.59	2.61	[-2.54, 7.72]	0.321
Omission rate	Sitting CV3	34.68	2.84	[28.42, 40.93]	<0.001
Hit rate among attempts	Sitting CV3	2.59	3.21	[-4.46, 9.65]	0.436
Omission rate	Sitting WCR-S	34.68			<0.001
Hit rate among attempts	Sitting WCR-S	2.59			0.445

Table A4: Posterior Individual-Effect Heterogeneity and Sorting-Gain Summaries

Quantity	Unit	Posterior mean	2.5%	25%	Median	75%	97.5%
Mean omission effect	pp	34.69	33.91	34.41	34.69	34.96	35.49
SD omission effect	pp	22.37	21.98	22.23	22.37	22.50	22.76
Mean hit-rate effect	pp	3.79	2.46	3.34	3.79	4.24	5.15
SD hit-rate effect	pp	11.56	10.27	11.12	11.56	11.98	12.83
Corr.(omission, hit-rate effects)	correlation	0.148	0.069	0.120	0.147	0.175	0.231
RPM-failure share	percent	37.41	32.43	35.61	37.35	39.14	42.65
Sorting gain defined	percent	96.95	96.46	96.79	96.96	97.12	97.44
Mean sorting gain	gain	0.147	0.059	0.115	0.146	0.177	0.240
SD sorting gain	gain	1.107	0.906	1.029	1.102	1.179	1.333
Median sorting gain	gain	0.101	0.055	0.085	0.100	0.116	0.148
Sorting gain <= 0	percent	38.41	33.30	36.59	38.35	40.20	43.78

Table A5: Bayesian Population Hierarchy Summaries

Outcome	Parent mean	Parent SD	New-sitting prediction	Observed-population ATE
Omission rate	34.90 [31.62, 38.00]	10.61 [8.32, 13.57]	[13.66, 53.16]	34.69 [33.91, 35.49]
Hit rate among attempts	3.49 [-3.40, 10.24]	13.86 [8.85, 21.21]	[-25.50, 32.49]	3.79 [2.46, 5.15]

Table A6: Observable heterogeneity in regime responses

	Omission response	Hit-rate response
Panel A: Full 2016–2018 sample		
Female	5.14	-0.71
	(0.60)	(0.78)
Math grade (SD)	0.91	0.97
	(0.31)	(0.34)
Repeat participant	-3.32	-0.26
	(0.88)	(1.07)
Non-public school	-3.76	-0.71
	(0.52)	(0.53)
Student observations	8,787	
Panel B: 2017–2018 survey sample		
Female	3.43	-0.73
	(0.77)	(0.98)
Math grade (SD)	0.85	0.99
	(0.39)	(0.43)
Repeat participant	-3.07	0.15
	(0.90)	(1.17)
Non-public school	-3.03	-1.83
	(0.62)	(0.65)
Willingness to answer (SD)	-6.77	-2.05
	(0.51)	(0.43)
Student observations	4,665	

Notes: Entries are percentage-point changes in the response to the omission-reward regime associated with the indicated student characteristic, estimated jointly within each panel. Mathematics grade is standardized within year-by-level sitting; willingness to answer is standardized in the survey sample. All specifications use weighted student-question regressions with student fixed effects and sitting-specific regime effects. Omission observations receive weight $1/J_z$ and attempted correctness observations receive weight $1/A_{iz}$, preserving equal student weight within each regime. Standard errors in parentheses are clustered two ways by physical participant and form-specific item. Coefficients are descriptive heterogeneity and should not be interpreted as causal effects of student characteristics.

Table A7: Binary Gender-by-Math-Ability Regime-Response Coefficients

Outcome	Term	Estimate	SE	95% CI	p
Omission rate	Female	3.91	0.90	[2.14, 5.67]	<0.001
Omission rate	Above-mean math grade	0.32	0.69	[-1.04, 1.68]	0.644
Omission rate	Female x above-mean math grade	2.35	1.08	[0.22, 4.48]	0.031
Hit rate among attempts	Female	-1.32	0.99	[-3.28, 0.64]	0.186
Hit rate among attempts	Above-mean math grade	1.09	0.72	[-0.34, 2.51]	0.134
Hit rate among attempts	Female x above-mean math grade	1.17	1.09	[-0.97, 3.31]	0.282

Table A8: Willingness to Answer by Gender and Ability

	(1) Willingness to answer	(2) Omission response (pp)	(3) Hit-rate response (pp)
Panel A. Gender gap in willingness to answer			
Female	-0.174^{***}		
	(0.033)		
Math grade (SD)	-0.044^{**}		
	(0.020)		
Female × Math grade (SD)	0.024		
	(0.032)		
Student observations	4,693		
Physical participants	4,379		
Panel B. Gender-by-ability heterogeneity conditional on willingness			
Female		1.768	−1.871
		(1.149)	(1.259)
Above-mean math grade		0.257	0.874
		(0.898)	(0.920)
Female × Above-mean math grade		2.864^{**}	2.004
		(1.382)	(1.533)
Student observations		4,665	4,665
Physical participants		4,353	4,353
Student-item observations		116,625	91,696

Notes: Panel A reports the 2017–2018 survey regression of willingness to answer under the omission-reward rule on female, school mathematics grade standardized within year-by-level sitting, and their interaction, with sitting fixed effects. Higher values indicate greater willingness to answer. Standard errors are clustered by physical participant. Panel B reports weighted student-item regressions of the regime response on female, an indicator for above-mean within-sitting mathematics grade, and their interaction. The regressions additionally control for regime interactions with willingness to answer, previous participation, and non-public-school status, and include student fixed effects and sitting-specific regime effects. Omission observations are weighted by the inverse number of questions in each regime; hit-rate observations are restricted to attempted questions and weighted by the inverse number of attempts in each regime. Standard errors in Panel B are two-way clustered by physical participant and form-specific item. Coefficients in columns (2)–(3) are percentage points. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Table A9: Persistence of Structural Metacognitive Profiles

Trait	Years	N	Spearman correlation [95% CI]
Sensitivity	2016–2017	321	0.349 [0.243, 0.443]
	2017–2018	312	0.358 [0.256, 0.453]
	2016–2018	199	0.355 [0.233, 0.465]
Bias	2016–2017	321	0.429 [0.333, 0.517]
	2017–2018	312	0.430 [0.330, 0.524]
	2016–2018	199	0.439 [0.317, 0.555]

Notes: The table reports Spearman correlations of posterior-mean residual student profiles across test administrations for participants observed in both years. Profiles are evaluated in a common grand-mean environment and therefore exclude year, level, residual sitting heterogeneity, observed student covariates, and item effects. The structural model treats each student-sitting separately and does not use participant identifiers to link observations across years. Intervals are 95 percent participant-bootstrap confidence intervals based on 2,000 replications. Two participants with ambiguous duplicate person-year observations are excluded.

Table A10: Structural Bias and Subjective Performance Forecasts

	(1)	(2)
Realized correct answers	0.452	0.163
	(0.032)	(0.057)
Knowledge	-0.010	0.701
	(0.184)	(0.222)
Sensitivity	-0.421	-0.248
	(0.166)	(0.165)
Bias	1.004	1.315
	(0.109)	(0.114)
Student covariates	No	Yes
Sitting fixed effects	Yes	Yes
Observations	3,310	3,310
R^2	0.381	0.401
Outcome mean	12.596	12.596

Notes: The dependent variable is the student's forecast of the number of correct answers. Both specifications condition on the realized number of correct answers and include year-by-level sitting fixed effects. Knowledge, sensitivity, and bias are posterior-mean common-environment residual profiles and are standardized to unit standard deviation within the validation sample. Column (2) adds the observed student covariates used in the structural model. The structural model is estimated without using the performance forecast. Standard errors, clustered by physical participant, are reported in parentheses.

Table A11: Within-Regime AUC Estimates

Method	Parent mean AUC	95% interval	Prediction interval
Plug-in conditional Rasch benchmark	0.643	[0.597, 0.688]	[0.479, 0.806]
Measurement-aware soft AUC	0.634	[0.590, 0.679]	[0.484, 0.785]

Notes: The table reports parent AUC estimates obtained by combining the twelve sitting-specific estimates using a random-effects model. The 95 percent intervals refer to the parent mean, while the prediction intervals incorporate between-sitting heterogeneity. The plug-in benchmark treats the estimated conditional-Rasch ranking as known; the measurement-aware soft AUC propagates posterior uncertainty about pairwise item orderings.

Table A12: Sorting Gains by Within-Sitting Soft-AUC Quintile

AUC quintile	Mean soft AUC	Mean sorting gain	95% interval	N
Q1	0.380	-0.058	[-0.108, -0.007]	1,678
Q2	0.549	0.069	[0.026, 0.111]	1,666
Q3	0.647	0.087	[0.047, 0.127]	1,683
Q4	0.735	0.203	[0.165, 0.240]	1,677
Q5	0.861	0.338	[0.287, 0.393]	1,682

Notes: The table reports mean soft AUC and mean individual sorting gains by within-sitting quintile of soft AUC among students whose number of omissions rises by at least one question between Parts 1 and 2. Sitting-specific means receive equal weight. Brackets report 95 percent intervals from a nonparametric student bootstrap within sitting-by-quintile cells.

Table A13: Validation of Full-Completer Item Rankings

	Ranking agreement		
Comparison	Spearman	Kendall	
Panel A: Part 1 item-ranking validation			
Full completers vs. other responders	0.889	0.765	
Other responders vs. other responders	0.898	0.778	
Panel B: Selection into full completion			
	Full completers	Other responders	Difference
Part 1 hit rate	0.488	0.451	0.037
Number of students	903	6,838	

Notes: Panel A validates the Part 1 item ranking estimated from full completers. Within each of the twelve sittings, the full-completer ranking is compared with rankings estimated from equally sized samples of other students who answer every Part 1 question. The first row reports equal-sitting mean agreement between full-completer and other-responder rankings; the second reports agreement between two equally sized samples of other responders. Panel B reports pooled Part 1 performance among students who answer every Part 1 question.

Table A14: Metacognitive Sensitivity and Confidence Bias by Gender

		Men	Women	M−F
Panel A. Threshold-free measures				
	Type-2 AUC	0.771	0.762	0.010*
		(0.004)	(0.004)	(0.006)
	Overconfidence	0.137	0.104	0.034***
		(0.005)	(0.006)	(0.008)
Panel B. Directional measures				
$\tau = 0.2$	Positive	0.921	0.876	0.045***
		(0.005)	(0.007)	(0.009)
	Negative	0.223	0.318	-0.095***
		(0.009)	(0.011)	(0.014)
$\tau = 0.5$	Positive	0.773	0.710	0.063***
		(0.007)	(0.008)	(0.011)
	Negative	0.581	0.648	-0.067***
		(0.010)	(0.010)	(0.014)
$\tau = 0.8$	Positive	0.574	0.521	0.053***
		(0.008)	(0.008)	(0.011)
	Negative	0.845	0.878	-0.033***
		(0.006)	(0.005)	(0.008)

Notes: The unit is the participant. Standard errors are in parentheses. The final column reports men minus women; stars test equality of gender means using Welch two-sample inference: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Type-2 AUC is the probability that a correct answer receives higher confidence than an incorrect answer, with half weight on ties. Overconfidence is mean confidence minus mean correctness. Positive is $\Pr(C > \tau \mid Y = 1)$; Negative is $\Pr(C \leq \tau \mid Y = 0)$. Confidence is rounded to the elicited two-decimal scale before thresholding.

Table A15: Robustness of Gender Differences in Confidence Bias

		Raw	Question FE	Question FE + LOO ability
Panel A. Overall confidence bias				
	Overconfidence	0.034*** (0.008)	0.031*** (0.008)	0.034*** (0.008)
Panel B. Directional measures				
$\tau = 0.2$	Positive	0.045*** (0.008)	0.049*** (0.008)	0.037*** (0.008)
	Negative	-0.095*** (0.014)	-0.090*** (0.014)	-0.076*** (0.014)
$\tau = 0.5$	Positive	0.063*** (0.011)	0.069*** (0.010)	0.050*** (0.010)
	Negative	-0.067*** (0.014)	-0.062*** (0.013)	-0.053*** (0.014)
$\tau = 0.8$	Positive	0.053*** (0.011)	0.059*** (0.011)	0.040*** (0.010)
	Negative	-0.033*** (0.008)	-0.030*** (0.008)	-0.027*** (0.008)

Notes: Entries are men minus women; CR1 standard errors clustered by participant are in parentheses. Each participant receives equal total weight within the relevant sample. LOO ability is correctness on all other confidence-elicitation questions. Positive is $\Pr(C > \tau \mid Y = 1)$; Negative is $\Pr(C \leq \tau \mid Y = 0)$. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A16: PSU Matching Diagnostics

	2014–2015 matched sample	
Panel A: Matching support		
Eligible 2014 students	151,169	
Matched pairs	93,213	
Overall match rate	61.7%	
Female match rate	62.8%	
Male match rate	60.3%	
Unique 2015 donors used	58,826	
Maximum donor reuse	15	
Panel B: Covariate balance		
	Before matching	After matching
NEM	0.019	0.002
School rank	0.040	0.001
Parental education	0.010	0.025
Household income	0.045	0.001

Notes: Panel A reports support for the matched synthetic-panel design around the 2015 PSU reform. Matching is exact on gender, high school, age, and employment status and uses one-nearest-neighbor matching with replacement on NEM, school rank, parental education, and household income, with calipers on NEM and school rank. Panel B reports absolute standardized mean differences for the non-exact matching covariates.

Table A17: Gender Differences in PSU Responses by Prior Achievement

NEM quintile	Mean NEM	Extra female selection (pp)	Extra female hit-rate gain (pp)
Q1	413.2	−1.31	−3.59
		[−2.35, −0.26]	[−4.24, −2.95]
Q2	475.8	1.10	−3.05
		[0.16, 2.04]	[−3.78, −2.31]
Q3	523.5	3.86	−1.71
		[2.60, 5.13]	[−2.47, −0.94]
Q4	582.1	7.60	−0.10
		[5.75, 9.45]	[−0.96, 0.76]
Q5	677.4	9.90	3.80
		[7.86, 11.94]	[2.84, 4.76]

Notes: Entries report female-minus-male differences in bias-corrected matched responses between the 2014 penalty regime and the 2015 no-penalty regime. Quintiles are formed from matched-pair average NEM using common cutoffs for women and men. Brackets report 95 percent confidence intervals based on standard errors clustered by high school.

Table A18: PSU Adjacent-Cohort Gender Contrasts

Matched cohorts	Extra female selection (pp)	Extra female hit-rate gain (pp)
2013–2014	0.80	−0.06
	[0.26, 1.34]	[−0.45, 0.32]
2014–2015	2.61	−0.65
	[1.32, 3.91]	[−1.12, −0.19]
2015–2016	0.25	−1.33
	[0.14, 0.37]	[−1.70, −0.96]

Notes: Entries report female-minus-male differences in bias-corrected matched adjacent-cohort responses. The 2014–2015 comparison spans the removal of the wrong-answer penalty; 2013–2014 and 2015–2016 are same-regime comparisons. Brackets report 95 percent confidence intervals with standard errors clustered by high school.

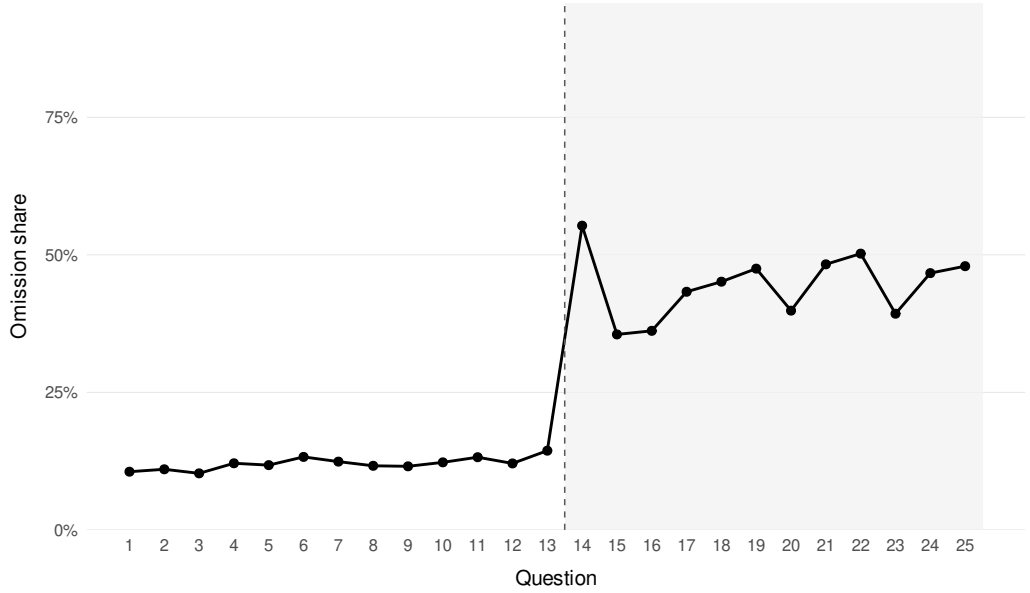
Table A19: Gauss–Hermite Parent-Distribution Cubature Validation

metric	mean_abs_difference	q95_abs_difference	max_abs_difference
parent_corr_effects	0.0012165924517021209	0.0025694306060432518	0.0031614719598265784
parent_cov_effects	3.196046012273191e-6	9.221752577465622e-6	4.021669607239312e-5
parent_mean_hit	9.020269812295195e-6	2.7187998497918244e-5	4.213718888630702e-5
parent_mean_omission	7.633540832939142e-6	3.566376897587421e-5	7.308223802970915e-5
parent_sd_hit	7.680947747632735e-4	0.0016437692996526584	0.002661913977816538
parent_sd_omission	1.618309502127678e-4	5.189979319214769e-4	0.0018213714738168352
parent_var_hit	2.399811414641959e-4	6.040643438114283e-4	0.0011861105347061446
parent_var_omission	3.852332583976496e-5	1.0844662312750937e-4	5.53143505671698e-4
residual_parent_cov_effects	2.6139027277662725e-6	7.993171173468565e-6	3.945530788379125e-5
residual_parent_sd_hit	9.419292262284136e-4	0.002013644171334865	0.0033527367069330904
residual_parent_sd_omission	2.7760896850930687e-4	8.025305060646461e-4	0.0025629639410915117
residual_parent_var_hit	2.4316130392248368e-4	6.170700365925996e-4	0.0012154168592325085
residual_parent_var_omission	4.0539310032505236e-5	1.1254290750163935e-4	5.892603371975332e-4
year_level_cov_effects	7.593926370418741e-7	2.995050796492523e-6	3.730553678166215e-6
year_level_sd_hit	1.6133708814029812e-5	6.36669964992411e-5	1.1311445559405997e-4
year_level_sd_omission	1.0580191575660768e-5	2.5926310115299822e-5	1.820283861089511e-4
year_level_var_hit	3.1902026353523775e-6	1.2944548748291739e-5	2.9306324526360422e-5
year_level_var_omission	2.015984192740333e-6	5.468553673575787e-6	3.6116831525835266e-5

Table A20: Bayesian Reduced-Form MCMC Diagnostics

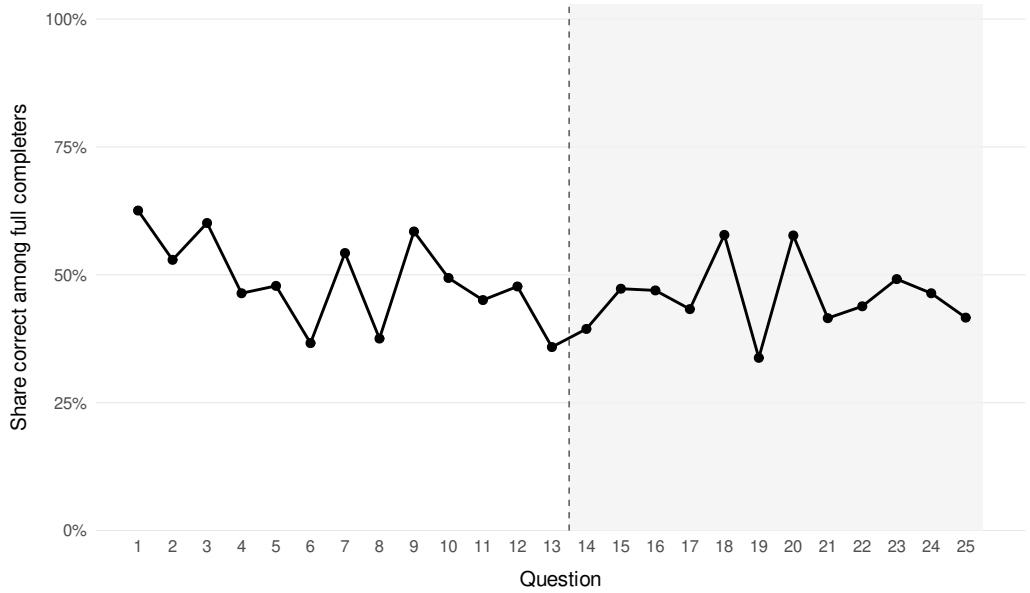
Diagnostic	Value
Posterior draws	8000
Chains	4
Divergences	0
Maximum-treedepth hits	0
Maximum R-hat	1.005
Minimum bulk ESS	750
Minimum tail ESS	1315
Minimum E-BFMI	0.664

Figure A1: Omission rate by question



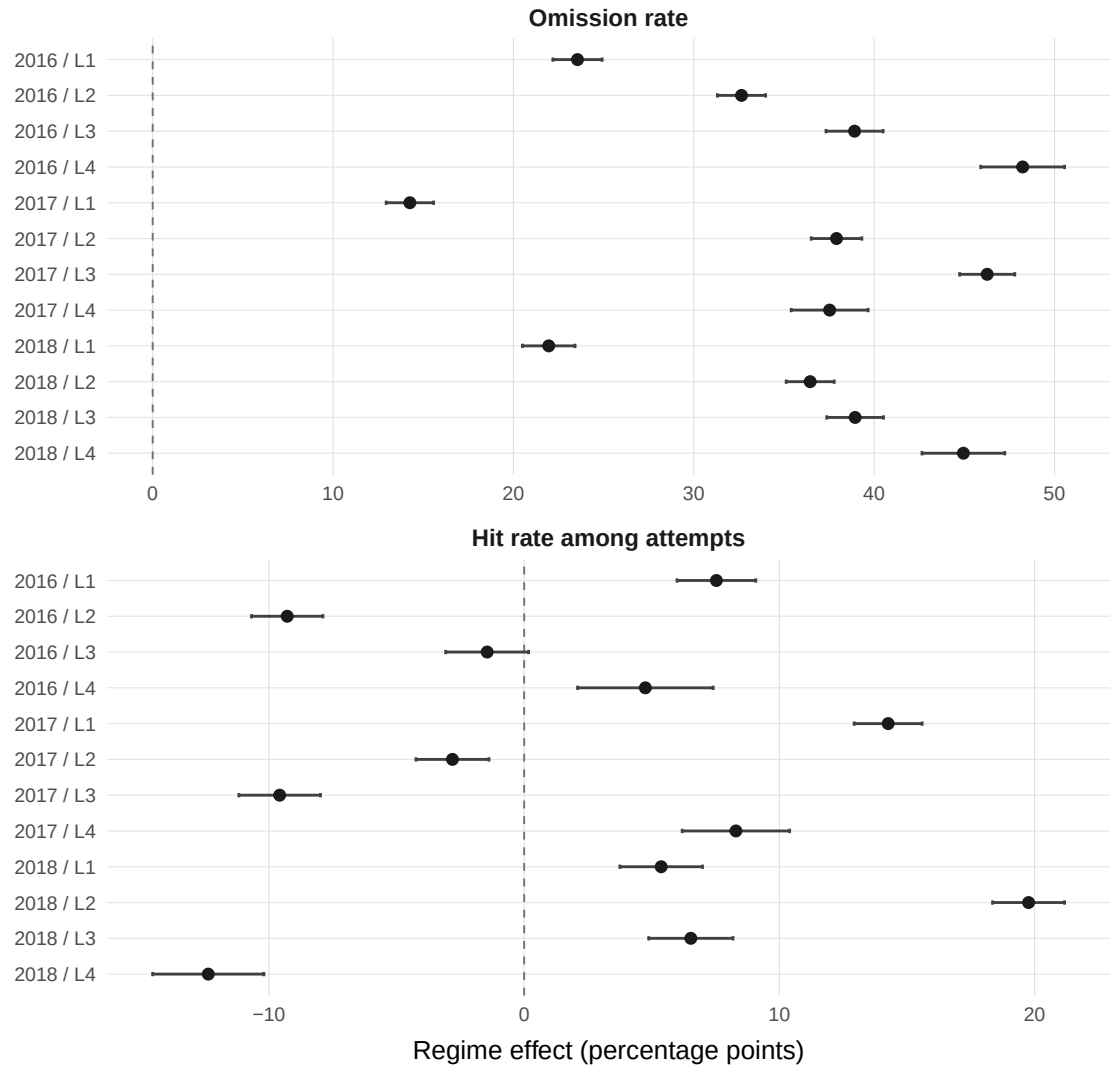
Notes: The figure reports the share of students omitting each question in the working sample, defined as students who attempt at least one question in each block. The dashed vertical line marks the scoring-rule boundary.

Figure A2: Hit rate by question among full-completer students



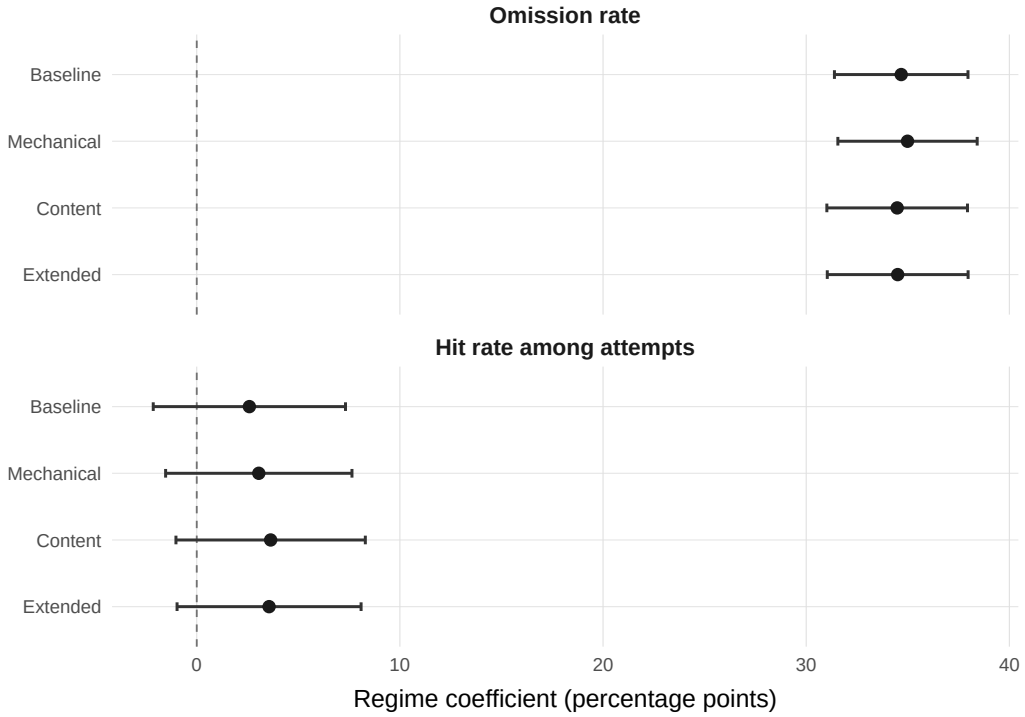
Notes: The figure reports the share of students that get each question right in among those that attempt all questions ("full-completers"). The dashed vertical line marks the scoring-rule boundary.

Figure A3: Sitting Heterogeneity in Regime Effects



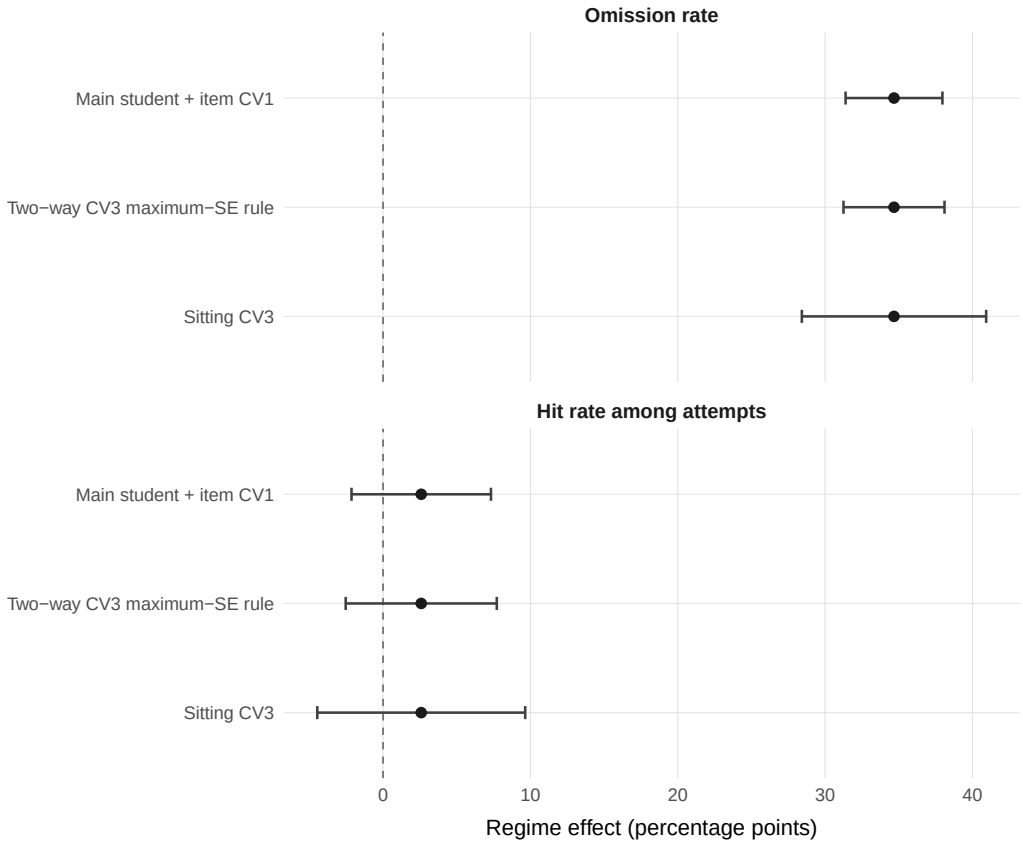
Notes: Points report sitting-specific average within-student regime effects in percentage points. Horizontal intervals are 95 percent HC3 confidence intervals across students, conditional on the realized test form.

Figure A4: Between-Regime Estimates under Item-Covariate Adjustment



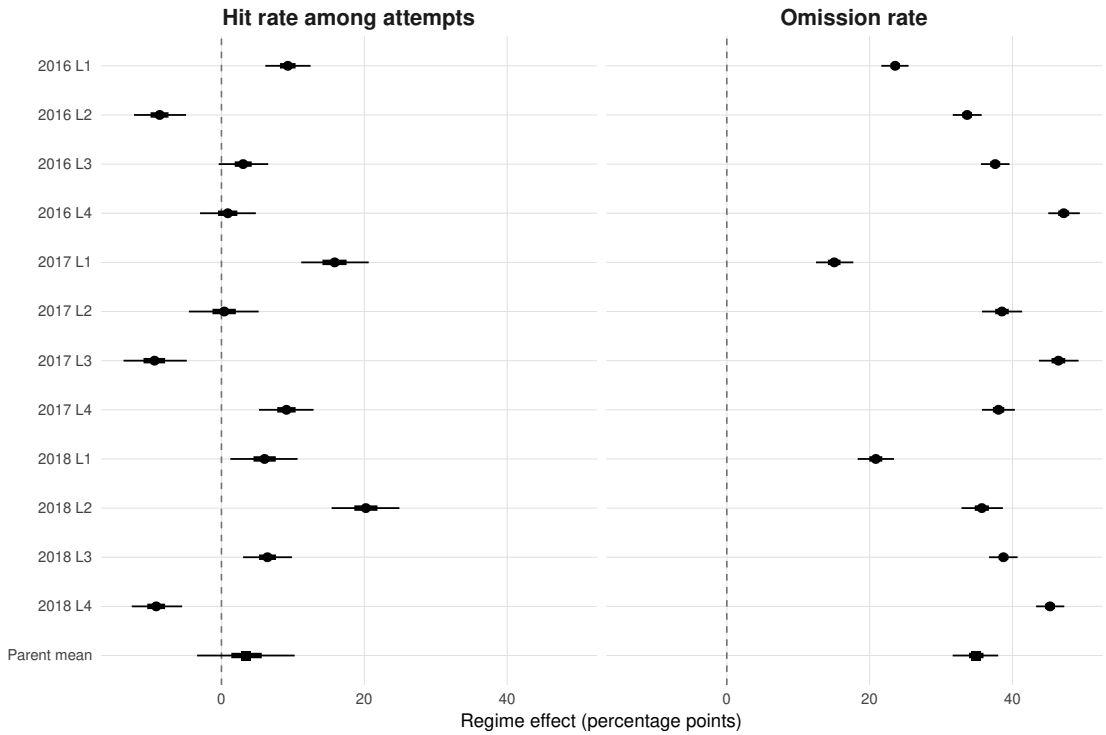
Notes: Points report weighted student-fixed-effects estimates in percentage points; horizontal intervals are 95 percent confidence intervals clustered by physical participant and form-specific item. Baseline includes no item controls. Mechanical adds stem length, mathematical-expression density, images, nonnumeric answer options, and correct-answer position. Content additionally controls for mathematical domain, contextualization, curricularity, reasoning structure, and structural devices. Extended additionally controls for near-duplicate options and official-solution quality flags.

Figure A5: Robustness of Pooled Regime-Effect Inference



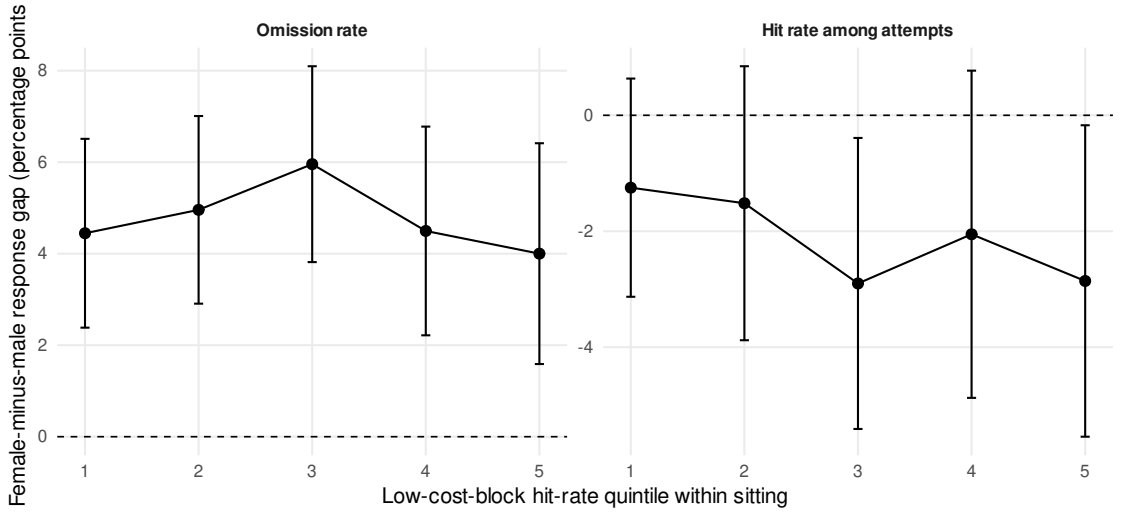
Notes: Points report the baseline weighted student-fixed-effects estimates in percentage points; only the inference method varies across rows. Main CV1 clusters by physical participant and form-specific item; two-way CV3 applies the maximum-standard-error rule along the same dimensions; sitting CV3 clusters by the 12 year-by-level test forms. Horizontal intervals are 95 percent confidence intervals.

Figure A6: Partially Pooled Sitting Effects and Parent Means



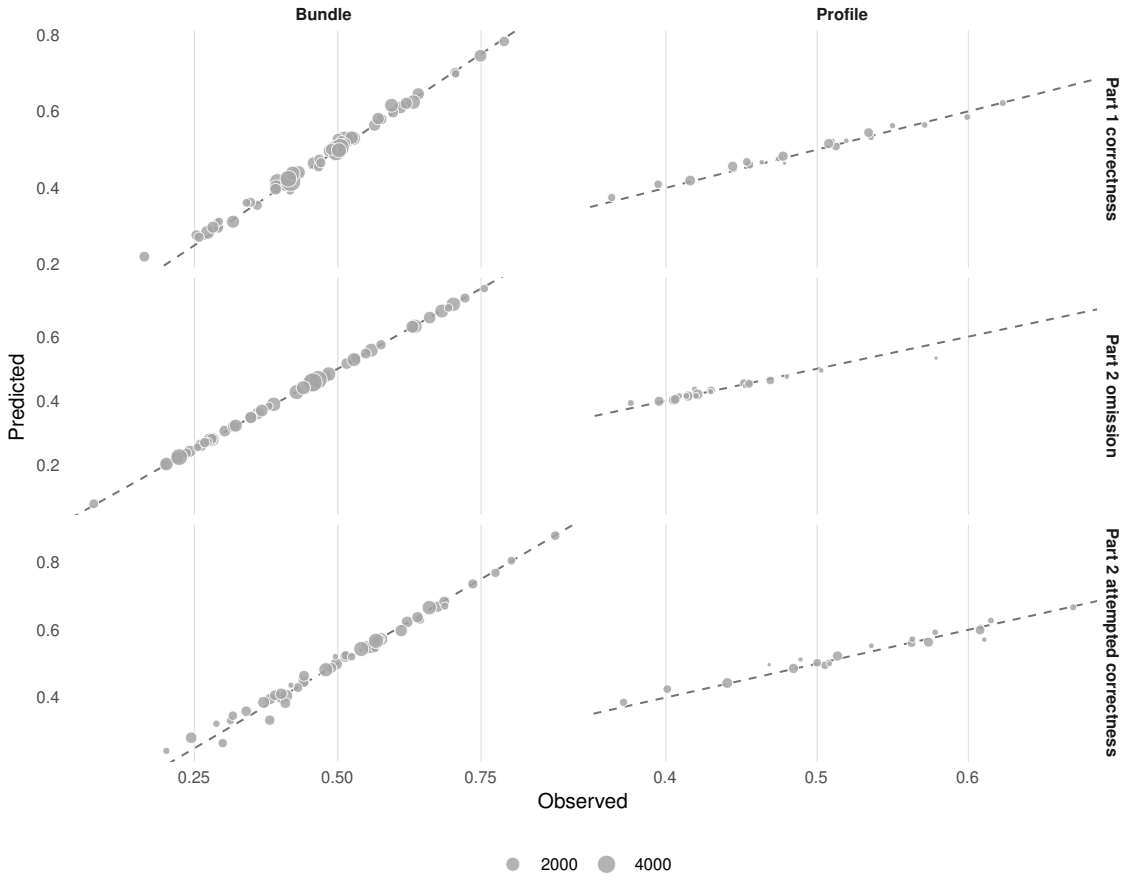
Notes: Circles report posterior means of the average probability-scale regime effect in each observed sitting, retaining its fitted sitting-specific innovation. Squares report the parent mean, obtained by giving equal weight to the 12 year-by-level cells and integrating over a fresh sitting innovation. Thick and thin lines are central 50 and 95 percent credible intervals, respectively. The parent interval describes uncertainty about the mean, not the predictive dispersion across new sittings. See Appendix Section D.1 for construction details.

Figure A7: Gender Differences in Regime Responses by Part-1 Hit-Rate Quintile



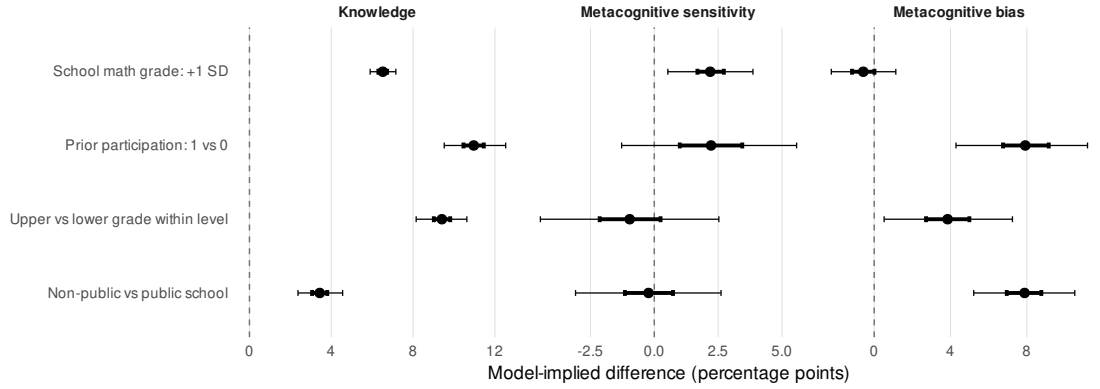
Notes: Points report female-minus-male differences in the within-student regime response, in percentage points, separately by quintile of the hit rate on Questions 1–13. Quintiles are constructed within sitting using average ranks, with ties kept together. Estimates use the weighted student-question specification with student fixed effects and sitting-specific regime slopes. Intervals are 95 percent confidence intervals clustered two ways by physical participant and form-specific item. Because the Part-1 hit rate is a component of the subsequent hit-rate change, heterogeneity in the right panel may partly reflect mechanical regression to the mean.

Figure A8: Observed-versus-Predicted Structural Model Fit by Profile



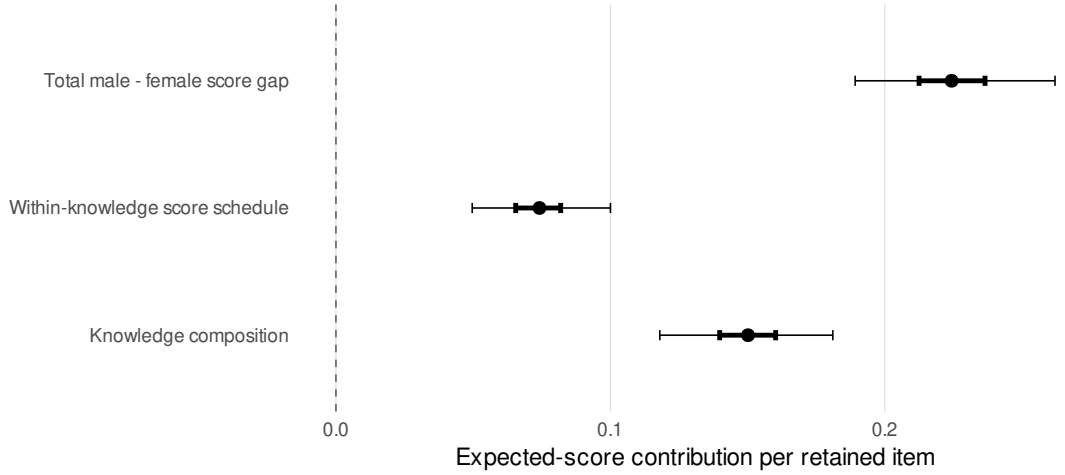
Notes: The figure compares observed rates with posterior mean fitted probabilities in the selected structural sample. Rows report Part 1 correctness, Part 2 omission, and Part 2 correctness conditional on attempt. Columns aggregate observations by item bundle and student profile. Item bundles are year-by-level sitting–domain cells; student profiles cross gender, mathematics-grade quintile, and prior-participation status. Bubble size reflects cell support, and the dashed line is the 45-degree line. Predictions use 200 posterior draws and the fitted student, sitting, and item effects. This is an in-sample calibration diagnostic rather than an out-of-sample validation or identification test.

Figure A9: Observed-Covariate Gradients in Structural Objects



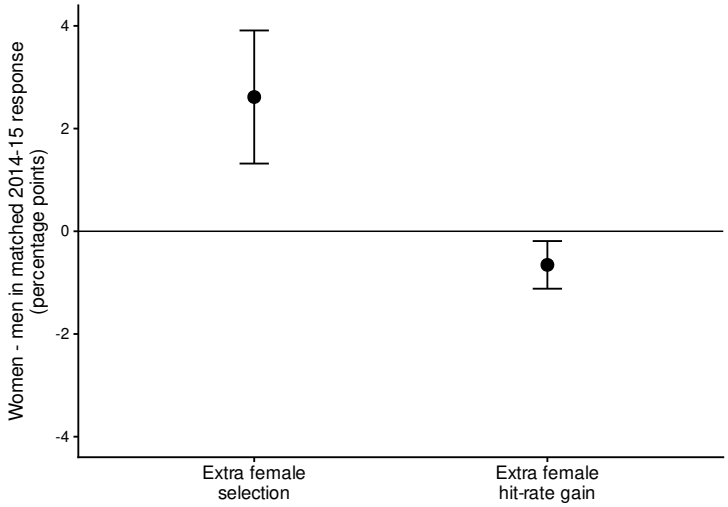
Notes: The figure reports model-implied differences in knowledge, metacognitive sensitivity, and metacognitive bias associated with observed student characteristics. The contrasts compare a one-standard-deviation increase in mathematics grade, one versus zero prior participations, the upper versus lower grade within level, and non-public versus public school. Each contrast changes only the indicated characteristic, holding all other observed characteristics and the parent residual draw fixed. Estimates integrate over the parent distribution of student heterogeneity and use the average-item convention. Points report posterior means; thick and thin intervals report 50 and 95 percent posterior credible intervals. Differences are expressed in percentage points and should be interpreted as conditional associations, not causal effects.

Figure A10: Oaxaca–Blinder Decomposition of the Gender Score Gap



Notes: The figure decomposes the model-implied male–female difference in expected Part 2 score per retained item. Within each posterior draw, female and male latent-knowledge values are pooled into five equal-frequency bins. The within-knowledge component compares the male and female expected-score schedules using the female knowledge distribution; the knowledge-composition component values gender differences in bin shares using the male score schedule. Expected scores apply the experimental 5/1/0 scoring rule and evaluate each profile against the retained Part 2 items in its sitting. Points report posterior means; thick and thin intervals report 50 and 95 percent posterior credible intervals.

Figure A11: Gender Differences in the Quantity and Quality Response in the PSU



Notes: The figure reports female-minus-male differences in bias-corrected matched responses between the 2014 penalty regime and the 2015 no-penalty regime of the mathematics PSU. Extra female selection is the gender difference in the change in the percentage of questions omitted; extra female hit-rate gain is the gender difference in the change in correctness conditional on attempting a question. Students are matched exactly on gender, high school, age, and employment status and by nearest neighbor on NEM, school rank, parental education, and household income, with replacement and calipers on NEM and school rank. Intervals are 95 percent confidence intervals with standard errors clustered by high school.

Figure A12: Selected Reduced-Form MCMC Trace Plots

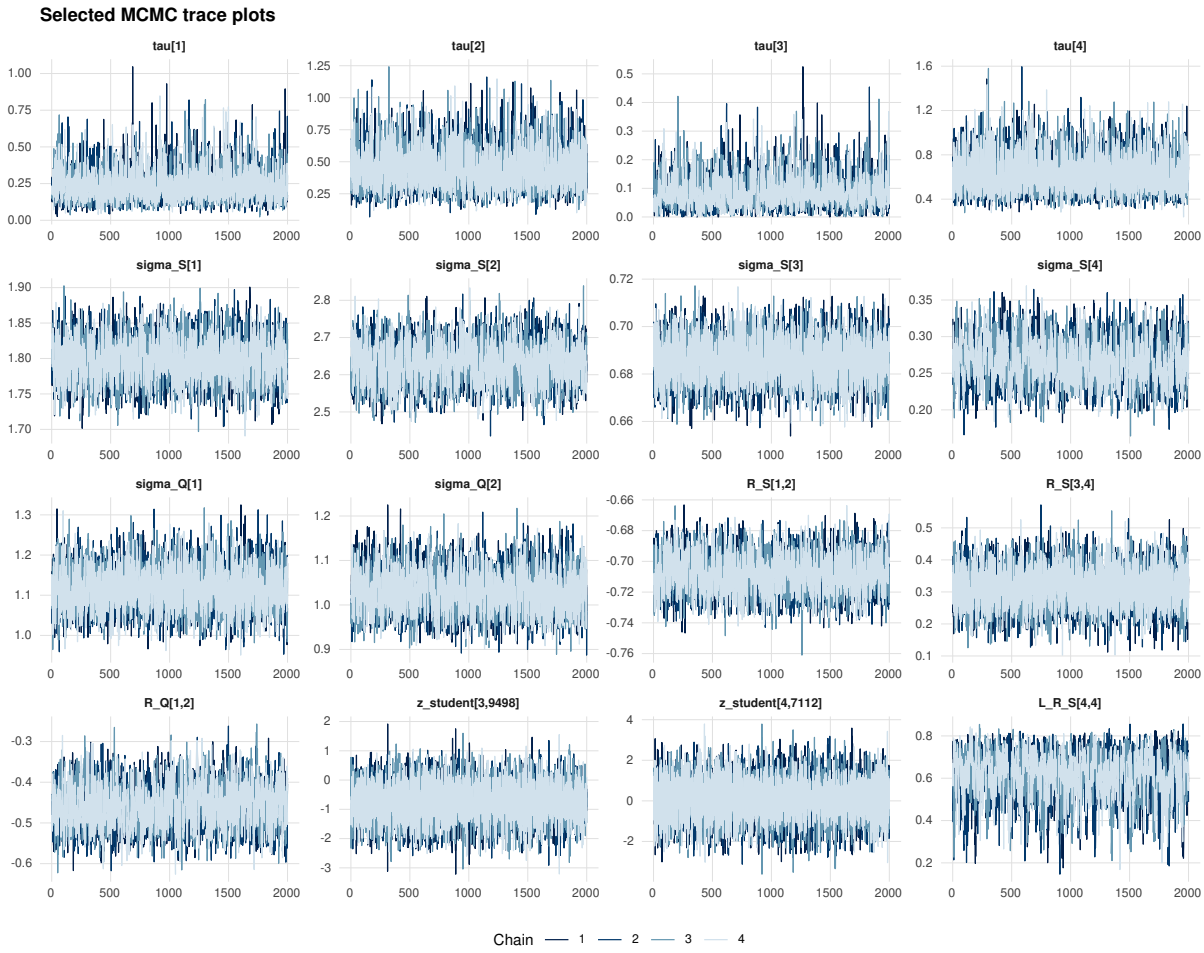


Figure A13: Selected Reduced-Form MCMC Rank Plots

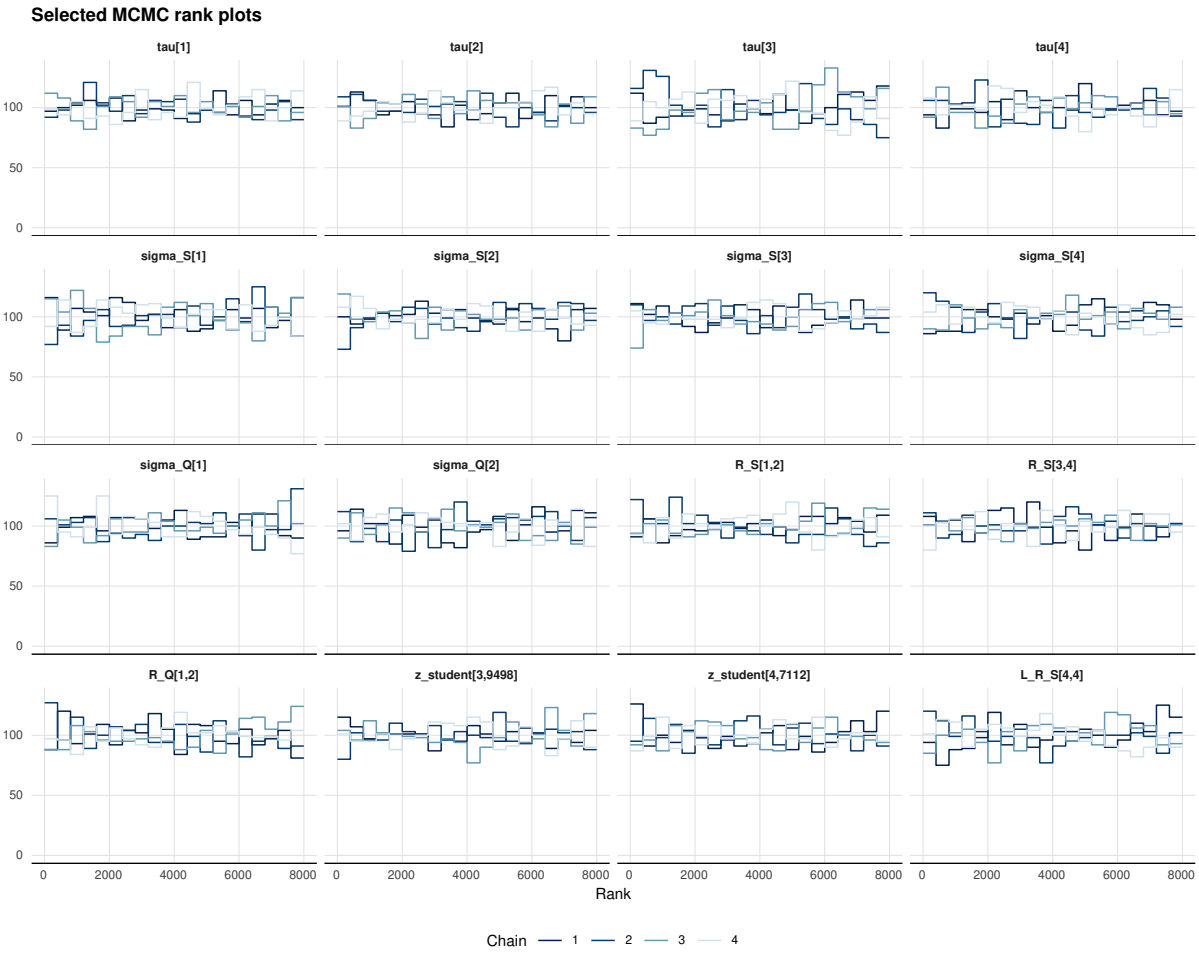


Figure A14: Posterior Predictive Check of Sitting-Regime Omission Rates

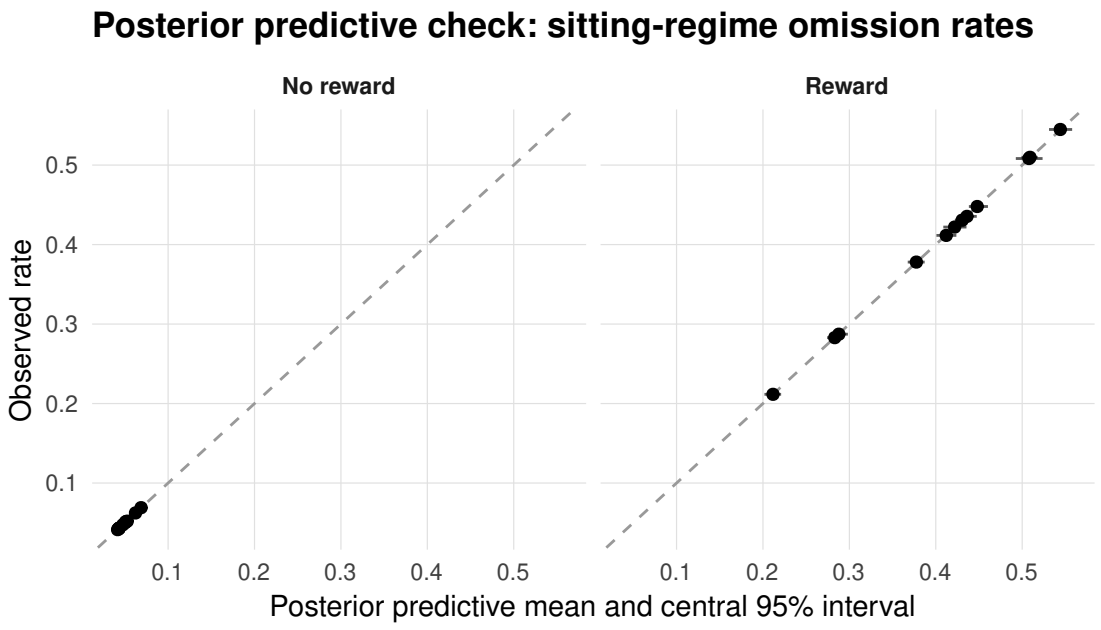


Figure A15: Posterior Predictive Check of Sitting-Regime Conditional Hit Rates

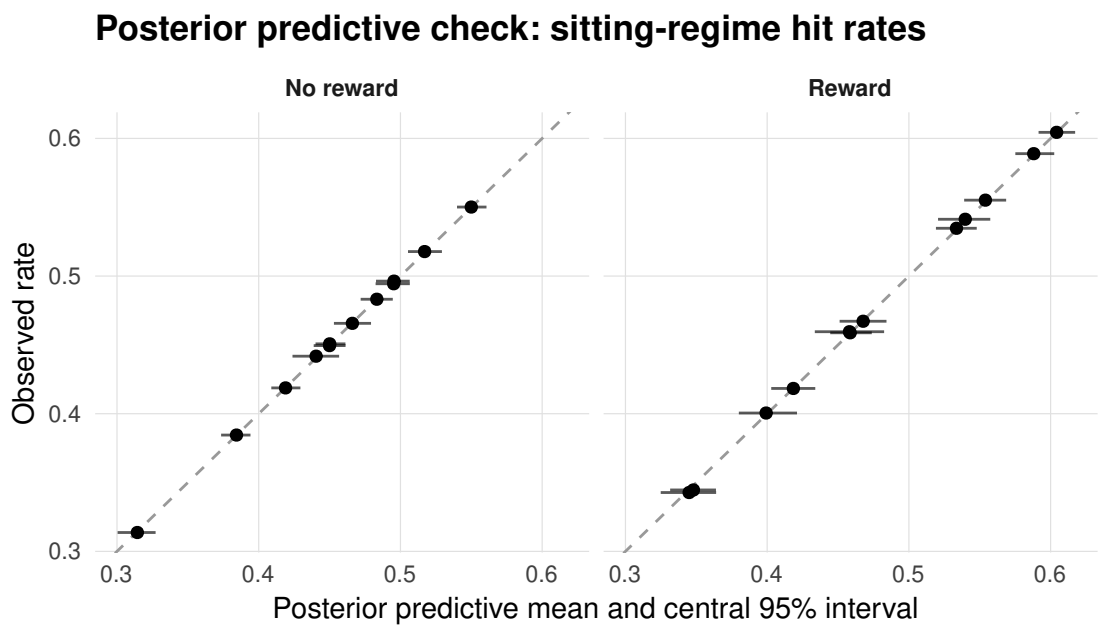


Figure A16: Posterior Predictive Student and Item Dispersion Checks

