Imperfect Metacognition in Standardized Tests

Francesco Bilotta*

March 23, 2026

Click here for the most recent version.

Abstract

This paper studies decision-making in standardized tests. I model test-taking as a multitasking problem, where students select questions guided by subjective confidence. Standard belief-formation models predict efficiency: while students may attempt too many or too few questions, they should attempt those they are objectively more likely to hit. Using data from a large field experiment, I show that multiple versions of this prediction fail. To account for the puzzle, I introduce imperfect metacognition, allowing errors in how students rank their own productivity across tasks. Structural estimation reveals substantial heterogeneity in metacognitive skill, accounting for over a third of test score dispersion. Moreover, whereas both genders are equally metacognitive, females are better at identifying harder questions, while males easier ones. This gender asymmetry implies that test score gaps should widen with productivity – a distinctive prediction confirmed in the data. My findings inform the design of scoring rules that restore fairness while rewarding both productivity and metacognitive skill.

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

*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 Martino Banchio, Roland Bénabou, Daniele Caliari, Chris Carnehl, Sylvain Chassang, Carlo Cusumano, Stefano Della Vigna, Astrid Esparza Sánchez, Ignacio Esponda, Francesco Fabbri, Katrin Gödker, Nenad Kos, Alessandro Lizzeri, Giacomo Manferdini, Paola Manzini, Marco Mariotti, Pietro Ortoleva, Filippo Palomba, Wolfgang Pesendorfer, and Collin Raymond for helpful conversations, as well as participants to Northwestern School in Economic Theory, Bocconi-CCA-Cornell Workshop in Political Economy, Princeton Micro Theory Student Lunch, Boston College PhD Conference, HEC Economics PhD Conference, Berlin Micro Theory and Behavioral Econ PhD Conference, Transatlantic Theory Workshop. The support of Premio Rotary Dott. Gabriele Corbelli is gratefully acknowledged. All errors remain my own.

> I see, Prometheus; and I want to give you the best advice, although you yourself are wily.
> Learn to know yourself and adapt yourself to new ways; for new also is the ruler among the gods.
>
> Aeschylus, *Prometheus Bound*, lines 309-312

Across time, societies have faced the challenge of reliably assessing skill and aptitude, often as a basis for allocating efficiently educational resources. In the modern economy, standardized tests have become the dominant instrument for this purpose, and test scores are used as input in a wide range of consequential decisions.

A chief example is college admission, where small differences in test performance can enable access to flagship universities, with long-run consequences for individual earnings and social mobility (Anelli, 2020; Ebenstein et al., 2016; Hoekstra, 2009; Zimmerman, 2019). Moreover, tests regulate school enrollment (Abdulkadiroğlu et al., 2014), track students into separate classes (Duflo et al., 2011), and assign them to gifted or remedial education (Bui et al., 2014; Figlio and Ozek, 2024). In turn, student scores inform value-added measures of teacher quality, thereby shaping career progression (Chetty et al., 2014a,b). Given their centrality, it is essential to understand what skills standardized tests actually measure. Part of the answer is straightforward: tests require solving problems, and productivity hinges on knowledge or domain-specific ability. At the same time, test-design often includes penalties for errors, time pressure, or examinee choice forms. These features introduce an implicit cost that compels students not only to produce answers, but also to choose which questions to tackle and which to skip, trading off quantity and quality.

Solving this trade-off efficiently may itself require a skill that students possess to varying degrees. A vast literature in the cognitive sciences documents substantial heterogeneity in individual *metacognition* – broadly defined as the ability to evaluate one's own cognitive processes and form appropriate personal-level beliefs (see Fleming, 2024, for a review). This skill has been linked to the efficient allocation of time and effort in problem solving (Ackerman and Thompson, 2017), as well as to the accuracy of performance self-assessment (Rouault et al., 2019). Importantly, metacognition appears to be distinct from fluid intelligence (Mazancieux et al., 2020; Rouault et al., 2018) and hence it may introduce an independent and potentially underappreciated source of variation in test performance.

In this paper, I develop a theoretical framework and provide empirical evidence on the role of metacognition in test-taking. I begin by showing that standard belief formation models imply (quasi-)perfect metacognition. While these models allow errors on the quantity margin of question selection (answering too much or too little), they predict efficiency on the quality margin: students should attempt those questions they are more likely to answer correctly. Using data from Iriberri and Rey-Biel (2021), I document that this prediction fails in multiple ways. I then show how introducing imperfect metacognition allows for errors in the ranking of one's own productivity across questions, thereby explaining the empirical puzzles. Structurally estimating the model further reveals systematic gender differences in metacognitive profiles: males and females have equal metacognitive skill,

but whereas females are better at recognizing low-success questions, males are better at recognizing high-success ones. This asymmetry explains why the gender gap in test scores widens with student productivity – a recurring pattern that alternative explanations struggle to account for (Akyol et al., 2016; Baldiga, 2014; Coffman and Klinowski, 2020; Conde-Ruiz et al., 2020; Díez-Rituerto et al., 2024; Iriberri and Rey-Biel, 2021). Motivated by these findings, I derive policy implications that speak to the informativeness and fairness of alternative scoring rules.

My analysis frames test-taking as a multitasking problem. Students choose a portfolio of items to attempt from a common set of questions, modeled as lotteries that yield either a correct answer (success) or an error (failure). Success probabilities vary across both students and questions, reflecting differences in ability and item difficulty. These probabilities are not observed directly. Instead, students act on noisy transformations of them – confidence levels. The mapping between objective success probability and subjective confidence encodes how accurately a student assesses their own productivity across tasks, that is, their metacognitive skill.

In this setting, test design determines whether a selection cost is present, and how it depends on the number of attempted items. Under preference monotonicity, subjectively optimal behavior takes a simple form: When selection costs are absent, students attempt all items; when costs are present, the optimal portfolio is confidence-monotone – it includes questions associated with higher confidence levels than those it omits.

Combining this decision framework with standard belief formation models yields a simple but powerful consequence, the *No Omission Without Selection (NOWS)* principle. Many benchmark assumptions – including "rational expectations", overconfidence (Moore and Healy, 2008), the hard-easy effect (Erev et al., 1994; Griffin and Tversky, 1992), or probability weighting (Gonzalez and Wu, 1999; Prelec, 1998) – imply that higher objective probabilities deterministically correspond to higher subjective ones. As a result, while students may be biased in their average confidence level, they can still correctly rank their productivity across questions, and hence attempt those they are most likely to succeed on.

The NOWS principle delivers two sharp behavioral predictions. The first concerns how choice behavior differs across test-designs with and without selection costs, holding item difficulty comparable. Given sufficient data, it should become extremely unlikely to observe students who omit questions at a positive rate when costs are present, but do not improve their share of correct answers among attempted items – *hit-rate* – compared to the no-cost case. The reason is straightforward: costs incentivize selection of higher confidence items, but since confidence is rank-preserving, higher confidence corresponds to higher success probability. This implication applies both at the individual level, as the number of questions grows, and at the aggregate level, as the number of students increases.

The second prediction requires that questions can be ordered by difficulty, so that easier

items correspond to higher success probabilities in the student population. Under this condition, NOWS implies that students should attempt easier questions than those they omit. Specifically, our multitasking problem induces a binary classification task: for each question, the student either selects it or skips it. The quality of a student's sorting can then be assessed using a standard metric for classifiers, the *area under the receiver operating characteristic curve (AUC)*. In my context, AUC is the probability that a question chosen at random from a student's portfolio has a lower difficulty score than a randomly chosen skipped one, and NOWS predicts an AUC equal to one, i.e. perfect sorting.

Does the NOWS principle hold empirically? To answer this, I leverage data from Iriberri and Rey-Biel (2021), who experimentally manipulate the scoring rule in the final stage of the *Concurso de Primavera de Matemáticas*, a large math competition held annually in Madrid. Across three years and four age groups, 11,095 students completed a multiple-choice test divided into two blocks – one with penalties for errors and one without. For each student, I observe item-level choices and outcomes under both regimes. Because selection costs vary within-subject and question difficulty is designed to be comparable across blocks, this setting offers an ideal testbed for evaluating the benchmark predictions.

The results reveal that the NOWS principle is systematically violated. I begin with the first implication, comparing omit-rates and hit-rates across regimes. At the individual level, 88% of students omit more questions under penalties, yet 45% of them fail to improve their hit-rate. At the aggregate level, students omit on average 0.35 points more when costs are active – over one-third of the test – and this effect is statistically robust; by contrast, the average increase in hit-rate is just 0.02 and becomes insignificant clustering standard errors by age group and year. Turning to the second implication, I estimate AUC at the student level using item difficulty rankings derived from the performance of students who answer all questions (treated as "graders"). The average AUC is 0.64 – closer to a random classifier than the perfect classification benchmark implied by NOWS. Taken together, these findings show that while students respond strongly on the quantity margin, they do not reliably improve the quality of their selection, failing to sort questions by difficulty.

To understand how the NOWS principle fails, I document three additional empirical facts. First, selection appears close to random for a large share of students: both the distribution of hit-rate differentials and that of AUC scores are peaked around their means. Consistent with this, the quantity and quality margins of selection appear largely decoupled – students who omit more questions do not, on average, achieve disproportionately higher hit-rates. Second, there are systematic gender differences in this regard: female students omit more than males, reflecting different responses to incentives, yet both groups are equally inaccurate in question selection. Third, the modal pattern masks substantial heterogeneity. Selection quality varies systematically across students, and the two measures of selectiveness (hit-rate gain and AUC) are positively correlated, pointing to a common

underlying skill.

These facts call for a richer account of student behavior. To move beyond the benchmark predictions, I relax the rank-preserving assumption embedded in standard models, allowing for a stochastic mapping between objective success probabilities and subjective confidence, introducing imperfect metacognition.

To understand this conceptual shift, consider a binary setting, where success probabilities and confidence levels each take two values: high or low. In this case, metacognition is described by two parameters: the true positive rate (the probability of reporting high confidence when success probability is high) and the true negative rate (the probability of reporting low confidence when success probability is low). These capture, respectively, a student's ability to recognize when they are likely to succeed and when they are not. In turn, their sum (minus one) defines a natural measure of metacognitive skill which reflects how strongly confidence co-varies with actual productivity and corresponds to the notion of *sensitivity* in the cognitive science (Fleming and Lau, 2014). Dually, the difference between the true positive and true negative rates, *bias*, captures whether the correlation between confidence and productivity is stronger at high or low levels, and thus determines the average displacement between the two dimensions. In this sense, classical notions of over- and underconfidence are recovered within the new framework, but are now separated from the ability to self-monitor.

In the binary model, metacognitive sensitivity governs the efficiency with which students resolve the quantity-quality trade-off at the core of question selection. Specifically, the sign of sensitivity determines the geometry of a frontier linking the fraction of questions attempted to the gain (or loss) in hit-rates generated by selection. When sensitivity is positive, reducing quantity yields increasing marginal gains in quality through efficient sorting. When it is negative, the geometry flips: increasing quantity "insures" against adverse selection, pushing hit-rates back toward the no-selection baseline. When sensitivity is zero, selection is as good as random, and the two margins are independent. The magnitude of these effects scales with the level of sensitivity and the dispersion of success probabilities across questions.[1] In this picture, metacognitive bias determines where a student lies along the frontier – that is, how many questions they attempt – capturing how the trade-off is resolved, while leaving total efficiency (the height of the frontier) invariant.

This geometric representation suggests that the binary model can account for the empirical findings above. To make this point quantitatively, I structurally estimate the model's parameters, leveraging the joint behavior of individual omit-rates and hit-rates across regimes, while assuming for identification that success probabilities are common in the population. In particular, I take the low state to reflect random guessing and the high

[1] Drawing a parallel with canonical selection models (such as Roy, 1951), dispersion in success probabilities acts as an intrapersonal comparative advantage, which is either exploited or squandered through sorting.

state to reflect perfect knowledge. The estimates reveal that the distribution of sensitivity is tightly peaked near zero, implying that selection is, on average, close to random – explaining the null effect of penalties on hit-rates, the average AUC, and the decoupling of quantity and quality. At the same time, meaningful dispersion in sensitivity accounts for heterogeneity in sorting ability, and hence the observed variation in selection quality measures. Finally, while males and females exhibit similar productivity and similarly low sensitivity, gender differences emerge through lower bias among females – resulting in fewer attempts without improved accuracy, consistent with the presence of a gap in omission but not in hit-rate differentials.

I next turn to the implications of metacognition for test scores, focusing on settings with penalties for errors, as in the experiment. I show that a student's score reflects not only how productive they are, but also how effectively they target their attempts. A simple decomposition makes this precise: the score equals a baseline term – what the student would get under random guessing – plus a (potentially negative) metacognitive premium. This premium has two components. The first captures the return to sensitivity and grows with the dispersion of success probabilities across questions. The second captures the return to bias. As intuitive, holding sensitivity fixed, the value of accurately monitoring high-success questions (true positive rate) relative to low-success ones (true negative rate) grows in the frequency of the former relative to the latter, generating a complementarity between productivity and bias. The asymmetry is especially sharp under *formula scoring*, where penalties are calibrated so that omitting a question yields the same expected payoff as random guessing: in this case, recognizing ignorance carries no reward.

Combining this decomposition with the structural estimates of my model yields a novel prediction: under penalties for errors, the gender gap in test scores should widen with student productivity. This pattern is borne out in the data. Average scores by gender diverge across the distribution of baseline hit-rates, with the gap more than tripling from the bottom to the top quintile. As noted above, a widening gender gap in test scores has been documented beyond the setting of Iriberri and Rey-Biel (2021), including high-stakes college admission (Coffman and Klinowski, 2020), and it is widely regarded as a puzzle, as standard preference-based explanations struggle to account for this pattern. In particular, if the gap were driven by gender differences in risk aversion, such differences would need to intensify with productivity – an assumption that is not supported by the survey measure or risk attitude in the dataset, nor by the broader literature.

Overall, my results show that metacognition plays a central role in shaping decision-making in tests and, in turn, performance. To conclude, I use counterfactual simulations to quantify its impact and draw policy implications. First, decomposing the observed score variation under penalties suggests that metacognitive heterogeneity accounts for roughly 35% of dispersion-confirming its role as a first-order driver. Second, a decomposition of the gender gap shows that metacognitive differences explain around 40% of the gap,

nearly twice the variation explained by productivity. As is natural, policy depends on objectives. If the goal is to measure productivity alone, no-penalty scoring is optimal. But in settings where metacognitive skill is itself a relevant criterion – for instance, in selecting individuals for high-stakes decision-making roles – penalties become necessary. In such cases, fairness can be restored by calibrating the omission payoff to student performance, which neutralizes the role of bias while preserving incentives for both productivity and sensitivity.[2]

In the Appendix, I develop a microfoundation for the framework and derive some general insights on the identification of metacognition from choice data. First, I show how modeling explicitly decision-making under penalties or capacity constraints give rise to instances of the multitasking problem studied in the main text, for different classes of preferences. Second, I provide a Bayesian foundation for metacognition, where confidence and success probabilities arise from a two-layer information process involving cognitive and metacognitive signals. Bayesianism imposes distinctive constraints to the stochastic mapping between success probability and confidence, formalized through implementation results. Third, I extend the geometric analysis beyond the binary case by introducing a hit-omit curve, which identifies properties of the confidence map with observable patterns in selection behavior. Finally, I generalize the complementarity between productivity and bias, showing as a corollary that overconfidence can be optimal when decision thresholds are conservative.

Roadmap The rest of this paper is organized as follows. The remaining part of the Introduction discusses some related contributions. Section 1 sets up my decision making framework. Section 2 develops the NOWS principle and Section 3 brings it to the data. Section 4 introduces the binary model of imperfect metacognition, which is estimated in Section 5. Section 6 analyzes the impact of metacognition on test scores and 7 derives policy implications. Section 8 concludes. Appendix A provides microfoundation for selection costs; Appendix B provides the Bayesian foundation for productivity and metacognition; Appendix C provides nonparametric results on metacognition. All the proofs in the paper are contained in Appendix D. Additional Tables and Figures are contained in Appendix E.

Related Literature

Confidence This paper contributes directly to a vast and influential literature on confidence in economic decision-making (Moore and Healy, 2008). A large body of work has studied overconfidence across a wide range of domains, linking it to excess entry in

[2]Still, setting higher or lower penalties may be useful to screen for differences in true positive and true negative rates – particularly in contexts where the cost of errors differs from the cost of inappropriate inaction.

competitive environments (Camerer and Lovallo, 1999; Niederle and Vesterlund, 2007), excessive trading (Barber and Odean, 2001; Odean, 1998), distorted managerial decisions (Malmendier and Tate, 2005, 2008), reduced job search (Spinnewijn, 2015), and political extremism and participation (Ortoleva and Snowberg, 2015).

In these applications, confidence is typically modeled as a distortion in levels: individuals misestimate a scalar object – such as their ability, their relative placement, or the precision of their beliefs – and make a single decision, typically whether or how intensely to act. By contrast, I model confidence as a noisy mapping from a vector of task-specific success probabilities into subjective confidence levels. This shift introduces a distinct dimension – sensitivity – which captures how accurately individuals rank their own performance across tasks (Fleming and Lau, 2014). I show that this largely neglected dimension is pivotal in a high-stakes multitasking environment: question selection in standardized tests, where it directly affects performance measurement and policy design.

Closest in spirit to my approach is Enke et al. (2023), who study the role of confidence in group decisions, showing that aggregate bias depends not on average overconfidence, but on the confidence-performance correlation at the task level. My contribution is distinct and complementary: I treat this correlation as an individual-level metacognitive feature and study its role in multitask decision-making, where performance depends on which problems individual choose to focus on.

Gender Differences in Test-Taking I also add to a growing literature that documents how the design of standardized tests may determine gender gaps in performance, focusing on the role of penalties for errors (Akyol et al., 2016; Baldiga, 2014; Coffman and Klinowski, 2020; Conde-Ruiz et al., 2020; Díez-Rituerto et al., 2024; Funk et al., 2023; Iriberri and Rey-Biel, 2021; Karle et al., 2022; Riener and Wagner, 2017), competitive pressure (Díez-Rituerto et al., 2025; Iriberri and Rey-Biel, 2019; Niederle and Vesterlund, 2010; Pekkarinen, 2015), time constraints (Galasso and Profeta, 2024; Krishna et al., 2025), as well as gendered cues in question framing (Ruiz-Tagle, 2024).

These works consistently find that such features lead women to omit more questions and obtain lower scores than equally able men, while offering mixed recommendations on whether penalties should be abolished. Such gaps are typically attributed to differences in risk preferences, competitive preferences or in overconfidence. My paper shifts the focus from the quantity to the quality margin of question selection and its dependence on metacognitive sensitivity. In doing so, it offers an explanation for a persistent puzzle – why gender gaps in test scores widen with ability – and points to a new dimension in the design of scoring rules as screening devices (e.g. Dasgupta, 2023).

Higher-Order Skills In my framework, metacognition functions as a higher-order skill: it determines how well students monitor their own productivity across questions, as

opposed to their odds of success. This connects my paper to an emerging literature on higher-order human capital and its implications for individual and social welfare (Deming, 2022; Deming and Silliman, 2024).

Specifically, Deming (2021) shows that U.S. jobs increasingly demand and reward the ability to decide which tasks to pursue, while Caplin et al. (2023) finds that cross-sectional differences in how individuals allocate attention across production factors predict income in both the U.S. and Denmark. In healthcare, Currie and MacLeod (2017) and Chan et al. (2022) conceptualize diagnostic skill as the ability to match patients to appropriate treatment (rather than executing the treatment itself) and show that heterogeneity along this margin explains variation in health-care practices, with first-order welfare consequences. Complementing this view, Chandra and Staiger (2020) show how misperceived comparative advantage determines misallocation of cases across hospitals.

Cognitive Noise Finally, the Bayesian foundation for metacognition relates to the literature on cognitive noise (Enke, 2024; Woodford, 2020), which shows how individuals exhibit attenuated responses to decision-relevant fundamentals, especially in complex tasks (e.g., Enke and Graeber, 2023; Frydman and Jin, 2022; Gabaix, 2019; Khaw et al., 2021).

In these models, attenuation arises from noisy cognitive signals about an optimal action or input, which are then weighted against a known prior. While signals are noisy, the agent is assumed to understand their structure. By contrast, my Bayesian model introduces a second layer of uncertainty: individuals are also uncertain about the informativeness of their own cognitive process. This metacognitive noise leads to misallocation across tasks, rather than attenuation within a task. Formally, this mechanism connects to Augenblick et al. (2024), who model Gaussian updating with uncertain signal strength, yielding overinference from weak signals and underinference from strong ones (see also Benjamin, 2019).

1 Decision-Making in Tests

I set up a framework to understand decision making in tests. Test-taking is modeled as a multitasking problem, where students' metacognition determines how accurately they assess their own productivity across questions. This structure yields a data generating process for observable outcomes, grounding the subsequent analysis.

1.1 Productivity and Metacognition

I consider a population of students i and questions j. Both students and questions are heterogeneous. In particular, each student has a type $s(i)$ from a finite set $\mathcal{S}$ and, similarly,

each question has a type $q(j)$ from a finite set $\mathcal{Q}$. A test consists of a sample of M students and of N questions, common across students.

Students choose which questions to attempt and which to skip. If student i attempts question j, the outcome is either a correct answer (success) or a mistake (failure). The probability of success, $p_{ij} \in [0,1]$, is drawn from a nondegenerate distribution $\pi_{s,q} \in \Delta([0,1])$ that depends both on the student i and the question j, but only through their types s and q. Draws are independent across (i,j) pairs.

The type-dependent distribution of success probabilities flexibly captures variation in *productive skill* across students and questions, allowing for fixed effects of both dimensions on performance. In particular, more productive students will draw from better success probability distributions across questions, whereas harder questions will involve worse distributions across students.

Students do not observe success probabilities but rather noisy transformations of them, called *confidence levels*. For each student-type s there is a *confidence map*

$$\psi_s : [0,1] \to \Delta([0,1])$$

such that, when the objective success probability is p_{ij} the subjective confidence level c_{ij} is drawn from the distribution $\psi_s(p_{ij})$. Students take confidence levels at face value and use them to assess their odds of success across questions.

The confidence map directly encodes the student's *metacognitive skill*, understood as the ability to monitor one's own productivity across cognitive tasks (Fleming, 2024) – in my case, test questions. Unlike standard belief formation models (Section 2), this formulation allows distortion not only in the level of success probabilities, but also in their relative ranking across items.

1.2 The Portfolio Problem

Students derive utility from the number of correct answers they obtain throughout the test. Additionally, depending on the test-design, they may face a *selection cost* that is a function of how many questions they attempt. Formally, each student chooses a portfolio of questions for answer, $A \subseteq \{1,\dots,N\}$, to solve

$$V_i^{\lambda}\left((c_{ij})_{j=1}^{N}\right) \equiv \max_{A\subseteq\{1,\dots,N\}} U_s\left(A^{+}\right) - \lambda\, k_s\left(|A|\right), \qquad \text{(Portfolio Problem)}$$

where A^+ is the random number of correct answers associated to the portfolio, $|A|$ is its size and $\lambda \in \{0,1\}$ is a parameter that indicates whether costs are present.

Both the benefit from correct answers (U_s) and the selection cost (k_s) may depend on the student's type, capturing heterogeneous preferences in the population. I maintain only one natural assumption: students prefer more correct answers to less, in a stochastic sense.

Denote by $\preceq_{FOSD}$ first-order stochastic dominance between subjective distributions of the number of correct answers associated to a portfolio.[3]

Assumption 1 (Preference Monotonicity)**.** For each student-type s,

$$A^{+} \preceq_{FOSD} A'^{+} \Longrightarrow U_s(A^{+}) \leq U_s(A'^{+}),$$

with strict inequality if the two distributions differ.

A direct consequence of the assumption is that the optimal portfolio has a *confidence-monotone* structure: it always includes questions whose confidence levels are (weakly) higher than those it omits, and it includes all questions in the absence of costs. For cost regime $\lambda \in \{0, 1\}$, let A_i^{λ} be a solution for student-i's Portfolio Problem (existence is guaranteed since the domain of optimization is finite).

Proposition 1. *Assume Preference Monotonicity. When costs are active ($\lambda = 1$), the optimal portfolio A_i^1 satisfies*

$$j \in A_i^1 \text{ and } k \in (A_i^1)^{\mathbf{c}} \Longrightarrow c_{ij} \geq c_{ik}.$$

In the absence of costs ($\lambda = 0$), the optimal portfolio is $A_i^0 = \{1, \ldots, N\}$.

To understand the logic of the result, observe that portfolio maximization can be split in two steps. First, conditional on a portfolio size, the student chooses *which* questions to answer; second, they optimally select the size itself. Since portfolios of the same size incur the same cost, Preference Monotonicity implies that the student prefers to swap lower-confidence items with higher-confidence ones, yielding the first point. Moreover larger portfolios allow for more successes with positive probability, hence they are preferred to smaller ones under no costs, yielding the second point.

1.3 Microfoundation

Whereas my framework is presented in reduced form, in the Appendix I provide a microfoundation. I conclude this Section presenting briefly its components.

Examples of Portfolio Problems In Appendix A, I show how the Portfolio Problem captures two sources of selection costs that are most common in practice. The first example is penalties for errors, assuming that students maximize expected utility over total score levels for either risk-neutral or CARA preferences. In these cases, moreover, the optimal portfolio is characterized by a *threshold rule* on confidence: there exists a fixed

[3]From a subjective perspective the distribution of the number of correct answers, depends on the vector of realized confidence levels $(c_{ij})_{j \in A}$ whereas, from an objective perspective, it depends on the realized success probabilities, $(p_{ij})_{j \in A}$.

cutoff $\hat{c}_s$ such that a question is attempted if and only if confidence exceeds the cutoff. In the following I will sometimes consider threshold rules and assume them in Section 4. A second example is when the test features capacity constraints, so that students can attempt only a fixed fraction of questions either by design, as in examinee choice formats (Wainer and Thissen, 1994), or due to time pressure.

Bayesian Metacognition In Appendix B, I show how success probabilities and confidence levels may jointly arise from a Bayesian model of information processing, in the spirit of cognitive noise (Enke, 2024; Woodford, 2020). The main innovation is that students face two layers of uncertainty. They form beliefs about the correct answer to a question through a cognitive experiment, and beliefs about the informativeness of that experiment through a second-order one, which models metacognition. Confidence then becomes the student's subjective forecast of success in a question, whereas success probability its long-run frequency. Importantly, Bayesian updating imposes that confidence levels and success probabilities must form a martingale coupling (Proposition A6). Hence, confidence matches the frequency of success on average, while the informativeness of the metacognitive experiment determines the strength of correlation between the two.

2 No Omission Without Selection

Leveraging my framework, I first ask what standard models of belief formation imply for decision-making in tests. I show that, whereas these models allow for bias on the quantity margin – students may answer too many or too few questions – they predict efficiency on the quality margin – students should attempt the questions they are most likely to hit. I refer to this principle as *No Omission Without Selection (NOWS)* and formalize two testable versions of it, which I bring to the data in the next Section.

2.1 Rank-Preserving Confidence

To understand the logic behind the NOWS principle, note that standard models typically assume that higher objective probabilities map into higher subjective ones. This is immediate for correct beliefs, but also holds for several common probabilistic biases. For example, underconfidence and overconfidence are typically modeled as constant shifts to probability levels (Moore and Healy, 2008). Even in non-linear models, distortions typically involve monotone inverse S-shapes, as in probability weighting (Gonzalez and Wu, 1999; Prelec, 1998) or in the hard-easy effect (Erev et al., 1994; Griffin and Tversky, 1992). To capture these and related cases, I impose the following condition.

Assumption 2 (Rank-Preserving Confidence)**.** For each student-type s, the confidence map ψ_s is (i) deterministic, (ii) non-decreasing, and (iii) non-constant.

In the context of tests, an immediate implication of Rank-Preserving Confidence is that the monotone structure of the optimal portfolio (Proposition 1) carries over from confidence levels to underlying success probabilities

$$j \in A_i^1 \quad \text{and} \quad k \in (A_i^1)^{\mathbf{c}} \quad \Longrightarrow \quad p_{ij} \geq p_{ik}. \tag{1}$$

This simple, but crucial, observation captures the core logic of the NOWS principle and underlies the two formalizations that follow.

2.2 Comparison between Tests

The first version of the NOWS principle compares student behavior across two tests: one with selection costs ($\lambda = 1$) and one without ($\lambda = 0$). Proposition 1 implies that, in the absence of costs, each student should answer all questions. When costs are active, by contrast, students may omit some items. If they do, Equation (1) implies that the omitted questions must have lower success probabilities than those selected. Hence, given sufficient data, the share of correct answers among attempted questions should weakly increase relative to the no-cost case.

For this logic to hold, the two test forms must involve comparable content, so that the same student draws success probabilities from the same type-dependent distribution across regimes.

Assumption 3 (Parallel Testing)**.** Question types are i.i.d. from the population and equal in both test regimes.

This assumption is minimal: it imposes no restriction on realized success probabilities and allows for arbitrary correlation across regimes. In particular, it nests both the counterfactual case where success probabilities are identical – e.g., the same test taken twice under different scoring rules – and the case where they are independent draws – e.g., two tests that involve different questions of comparable difficulty.

Towards a statement, define for each student i and each cost regime $\lambda \in \{0, 1\}$ the *omit-rate* $O_{i,N}^{\lambda}$ as the fraction of questions left unanswered and the *hit-rate* $H_{i,N}^{\lambda}$ as the fraction of correct answers among attempted questions. Moreover, denote their *sample averages*, respectively, by $O_{M,N}^{\lambda}$ and $H_{M,N}^{\lambda}$.[4]

Next, observe that there are two ways the amount of data may increase. First, fixed a student i, the test-length N may grow. Second, fixed the test-length, the student sample size M may. In either limit, I show that it becomes increasingly unlikely to observe a strictly positive rate of omission under selection costs, but no corresponding improvement in hit-rate compared to the no-cost regime.

[4]Formally, $O_{i,N}^{\lambda} = 1 - |A_i^{\lambda}|/N$ and $H_{i,N}^{\lambda} = \frac{1}{|A_i^{\lambda}|} \sum_{j \in A_i^{\lambda}} \mathbf{1}\{\text{success on } j\}$. Note that $H_{i,N}^{\lambda}$ is undefined if $|A_i^{\lambda}| = 0$, hence the sample averages are taken on the subsample for which $|A_i^{\lambda}| \neq 0$.

Proposition 2. *Assume Preference Monotonicity, Rank-Preserving Confidence and Parallel Testing. Fix any $\varepsilon \in (0, 1/2)$. The following hold.*

(i) For any student i,

$$\lim_{N\to\infty} \mathbb{P}\Big(\Big\{O^1_{i,N} \in [\varepsilon, 1-\varepsilon]\Big\} \cap \Big\{H^1_{i,N} - H^0_{i,N} \leq 0\Big\}\Big) = 0.$$

(ii) For any test length $N \geq 2$,

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

Both parts of the result follow from Law of Large Numbers arguments. To get the gist, consider a student that adopts a threshold rule. Under Rank-Preserving Confidence, the confidence cutoff induces a corresponding cutoff on success probabilities. Then, as test-length grows, the hit rate under selection costs converges to the expected success probability above this cutoff, while the no-cost hit rate converges to the unconditional mean. If the student omits a positive share of questions, by the law of total expectations, the conditional mean must strictly exceed the unconditional one, implying a higher hit rate under selection costs. This yields the individual level statement. The logic for sample averages is analogous, but convergence is driven by averaging across students rather than across items. The key point is that the population is only limitedly heterogeneous: students belong to a finite set of types and are ex-ante identical within each type.[5]

2.3 Area Under the Curve

The second version of the NOWS principle exploits question fixed-effects on performance. Suppose that question types can be ranked by difficulty, so that higher-ranked questions are systematically easier, for all student-types. Formally, consider the following restriction on the support of success probabilities.

Assumption 4 (Strictly Ordered Supports)**.** There exists an order $\prec$ on $\mathcal{Q}$ such that $q \prec q'$ if and only if, for all $s \in \mathcal{S}$,

$$p \in \operatorname{supp} \pi_{s,q} \text{ and } p' \in \operatorname{supp} \pi_{s,q'} \implies p < p'.$$

Under this assumption, Equation (1) implies that, when facing selection costs, students should systematically attempt the easier questions.

Proposition 3. *Assume Preference Monotonicity, Rank-Preserving Confidence, and Strictly Ordered Supports. Then, for any student i,*

$$\mathbb{P}(q(k) \prec q(j)) = 1,$$

[5]Moreover, the proof shows that Parallel Testing can be weakened in complementary ways across the two parts of the result. For the individual-level result, it suffices that the two question-type vectors are drawn independently from the same distribution. For the population-level result, the vectors must be identical across regimes but can be chosen arbitrarily, not necessarily i.i.d.

where j is drawn uniformly from the set of attempted questions A_i^1, and k from the set of omitted questions $(A_i^1)^c$.

The probability in the statement can be interpreted as the *area under the receiver operating characteristic curve (AUC)*, a standard metric in binary classification. To see the connection, regard the student as a binary classifier that assigns a positive label to questions they choose to answer, and a negative one to those they skip. In this sense, the question's difficulty is the latent target variable and correct classification means attempting easier questions and omitting harder ones. An AUC equal to 1 (as in the Proposition) corresponds to perfect classification; if questions are selected at random, instead, the AUC falls to 0.5; values below this threshold are diagnostic of adverse selection. In this sense, beyond offering a test for Rank-Preserving Confidence, the Proposition defines a natural measure of selection quality, gauging the deviation from this benchmark.

3 Evidence from a Large Field Experiment

I now bring to the data the NOWS principle, exploiting a large-scale field experiment on multiple-choice tests that features within-subject variation in penalties for errors. My results systematically challenge the benchmark predictions. While students do omit more questions when facing penalties, their choices resemble near-random selection, on average. At the same time, there is meaningful heterogeneity across students, suggesting that selecting questions requires skill. In this respect, I document distinctive gender differences: female students omit more than males, but do not achieve higher selection quality.

3.1 The Experiment (Iriberri and Rey-Biel, 2021)

Iriberri and Rey-Biel (2021) conduct a large-scale field experiment in collaboration with the organizers of the *Concurso de Primavera de Matemáticas*, a regional math competition held annually in Madrid.

Students face a test that features *within-subject* variation in the scoring rule: 13 questions have no penalty for errors relative to omission, whereas 12 questions do. Each question has 5 candidate answers. In the first block, correct answers earn 5 points while both errors and omissions earn 0. In the second block, scoring remains the same for correct and incorrect answers, but omissions now earn 1 point. Aside from the scoring rule, the two parts are designed to be comparable across all dimensions, especially question difficulty.

I observe question-level performance on the test (i.e., whether the question is answered and whether the answer is correct) for approximately 11,095 students, over three editions (2016, 2017, 2018) of the *Concurso*. In each edition, a different set of questions is administered to each of four age groups (spanning ages 10–17): to control for test form I define a *sitting* as a (year, level) pair. In addition to performance data, the dataset records demographics

– gender, math grade at school, school type (public/mixed/private) – and how many times the student participated in the *Concurso*. For participants in the 2017 and 2018 editions, I also observe two survey measures, elicited at the end of the test (i) risk aversion, rated on a decreasing 1–5 scale[6]; (ii) the student's forecast of the number of correct answers in the whole test.

Given my focus on students' selection behavior, I restrict attention to those students who attempt at least one question in both parts of the test, yielding a subsample of 9,807 students. Among them, 79% answer all questions in the no-penalty block, just as Preference Monotonicity predicts. Of the remaining 21%, 90% still omit (weakly) more questions under penalties than without.[7] This suggests that students understand the incentive scheme, and that omissions in the no-penalty part may reflect trembles rather than non-monotone preferences. I therefore retain all students in the sample, but report robustness checks throughout to show that excluding this subgroup does not affect the results.

Within my framework, the variation in the scoring rule is an exogenous shift in the regime parameter, from $\lambda = 0$ (no selection costs) to $\lambda = 1$ (costs). I analyze treatment effects of this variation on omit-rates and hit-rates, both at the individual and at the average level. Two considerations support causal identification. First, placebo evidence from earlier editions helps rule out confounders related to test timing (e.g., fatigue), supporting temporal stability. In the 2013-2015 editions – where the scoring rule was constant across both test halves – Iriberri and Rey-Biel (2021, Table A2) find no statistical differences in omit-rates, hit-rates, or scores between halves.[8] Second, students were unaware they were part of a study, reducing within-test anticipation effects and across-edition behavioral spillovers. This supports the assumption of causal transience.

3.2 Testing the NOWS principle

I next test the NOWS principle, bringing to the data Propositions 2 and 3.

Treatment Effect Analysis

First, I examine individual treatment effects, defined as the differences in omit-rates and hit-rates between the penalty and no-penalty regimes. Figure 1a plots their joint distribution across students, with omit-rate differentials on the x-axis and hit-rate differentials on the y-axis. The first part of Proposition 2 predicts that students who omit more under

[6]The prompt was: *When omitting a question was worth 1 point, I answered the question – (1) only if absolutely sure, (2) if almost sure, (3) if uncertain between 2 answers, (4) if uncertain between 3 answers, (5) always.*

[7]In total, 97% of the sample does.

[8]Indeed, students are free to answer questions in any order, so presentation order need not match completion order.

penalties should increase their hit-rate. Graphically, we should see no mass in the southeast quadrant. This prediction fails: while most students (88%) omit more questions when penalties are present, a sizable fraction of them (45%) fails to simultaneously improve their hit-rate.

Because each student answers only 13 and 12 questions per regime, individual treatment effects are subject to non-negligible finite-sample noise. Still, by the second part of Proposition 2 the sign pattern should also hold for sample averages. Hence, I turn to average treatment effects, which can be estimated precisely given the large number of students observed. To this end, I run a within-subject fixed effects specification

$$Y_{it\lambda} = \mu_{it} + \beta \cdot \lambda + \varepsilon_{is\lambda}, \tag{2}$$

where $Y_{it\lambda}$ is the outcome (omit-rate or hit-rate) for student i in sitting t under regime $\lambda \in \{0, 1\}$, and μ_{it} is a student-by-sitting fixed effect. Following Abadie et al. (2023), I report both standard errors clustered at the student and the student-sitting level.[9]

Results are shown in Table A1 and plotted in Figure 1b. Penalties have a large effect on omit-rates: on average, students omit 35 percentage points more when costs are active – over one-third of the test. This effect is also robust, as it remains highly significant under both clustering strategies. In contrast, the average effect on hit-rates is negligible. The point estimate is close to zero (0.02) and only significant under the less conservative student-level clustering. Hence, whereas students consistently respond to incentives by omitting more, they do not reliably improve the quality of their selection.[10]

Figure A1 replicates the findings both for individual and for average treatment effects, excluding those students who omit questions absent penalties.

AUC Analysis

I next examine the distribution of the area under the curve (AUC) metric across students. Estimating AUC requires ranking questions by difficulty. Under Strictly Ordered Supports, such a ranking can be inferred from performance data: if one could make a sufficiently large random sample of graders answer all questions, the fraction of correct responses per item would proxy its difficulty.

While no such sample is directly available in my data, I construct an analogue by identifying as graders those students who self-select into answering all questions in both parts of the test. Since graders do not select among items, variation in their success rates should

[9]The first clustering choice treats the sample as fixed and yields valid inference for the observed data. The second clustering choice is conservative with respect to treatment heterogeneity across-sitting, treating sittings as draws from a super-population; in this sense, it has a meta-analytic interpretation.

[10]This picture is also confirmed by a heterogeneity analysis across sittings (Figures A2a and A2b). The treatment effect on the omit-rate is consistently large and significant, while the effect on the hit-rate fluctuates around zero, with most estimates falling in the interval $[-0.1, 0.1]$.

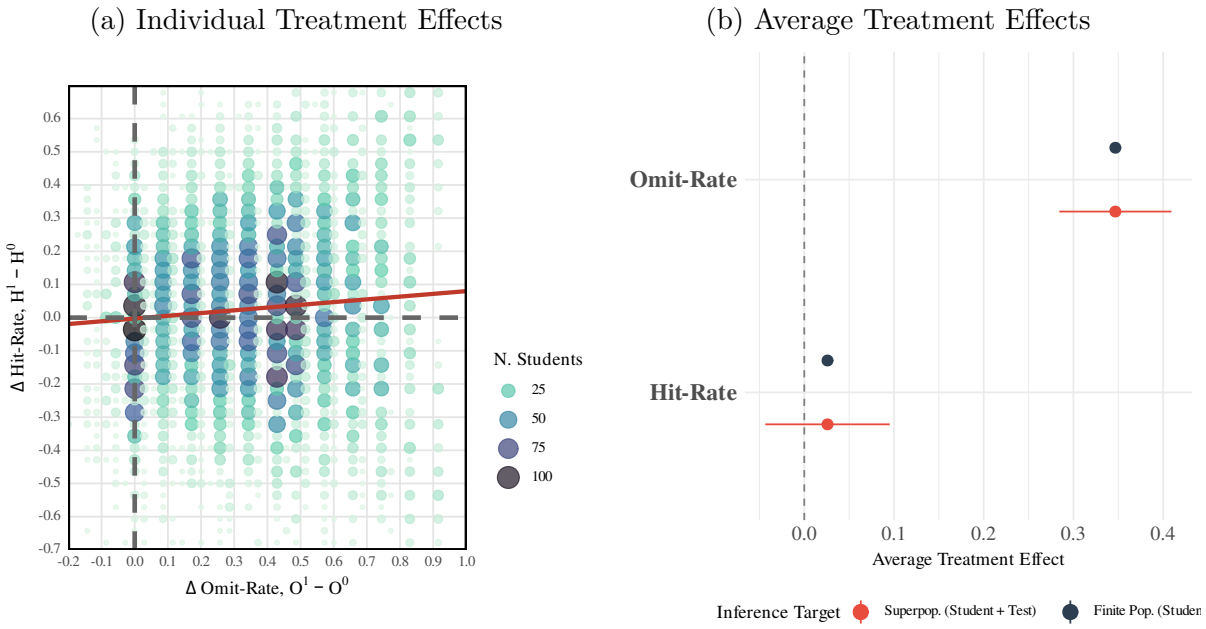


Figure 1: Treatment Effects of Penalty for Errors

reflects only differences in question difficulty. Hence, I define each item's difficulty score as the share of graders who answer it correctly.

Several facts support the credibility of this proxy. First, the number of graders is large enough to ensure statistical precision: each test sitting includes between 35 and 150 graders, with an average of 75 (Table A2). Second, graders are broadly representative of the main population: their hit-rates in the no-penalty block differ from the population average by at most 0.1 and by only 0.03 on average. Third, the resulting difficulty scores vary substantially across items (Figure A3) and do not appear to be driven by fatigue or test position – items near the end of the test do not show systematically lower success rates.

Given the difficulty ranking, I estimate each student's AUC using the Mann-Whitney U statistic. Let A denote the set of attempted questions, and define $X = \{x_1, \ldots, x_{n_1}\}$ as the difficulty scores of attempted items, and $Y = \{y_1, \ldots, y_{n_0}\}$ those of omitted items. The AUC estimator is

$$\widehat{\text{AUC}} = \frac{1}{n_1 n_0} \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} \left[\mathbf{1}\{x_i > y_j\} + \frac{1}{2} \mathbf{1}\{x_i = y_j\} \right].$$

Figure 2 shows the distribution of estimated AUC scores across students. Under the benchmark (Proposition 3), the distribution should concentrate near one. Instead, the data deviate sharply from this prediction. The average AUC is 0.64 – closer to random selection than to perfect sorting. Moreover, 24% of students have AUC ≤ 0.5, meaning they are more likely to attempt harder questions than those they skip, as measured by

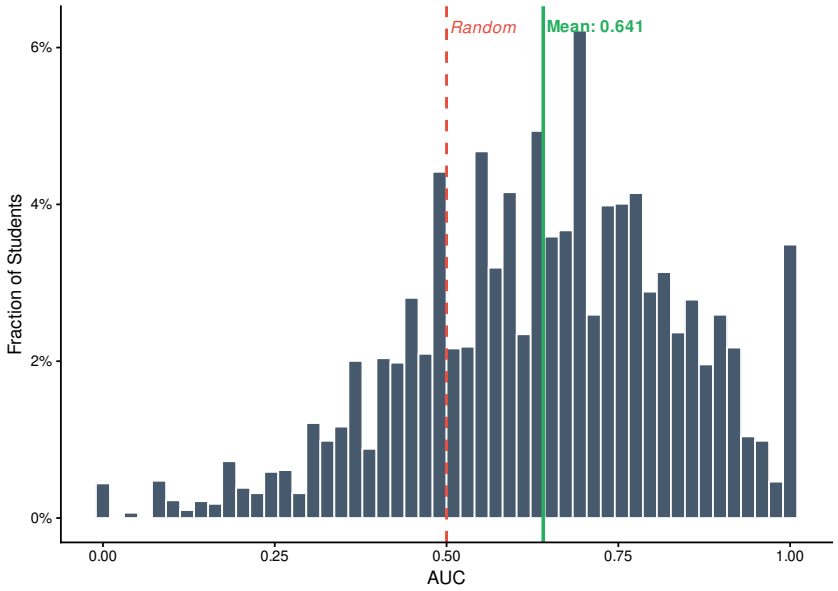


Figure 2: Population Distribution of AUC

grader difficulty.

Again, Figure A4 replicates the analysis restricting to students who do not omit any questions in the no-penalty block.

3.3 Three Facts on Question Selection

Beyond rejecting the NOWS principle, the experimental evidence also sheds light on the structure of student selection behavior. Three observations are informative.

First, both the distributions of the hit-rate differential and of the AUC score concentrate around their mean values (Figures A5 and 2). Hence, selection is close to random for a large share of students, not just on average. Moreover, consistent with students selecting questions at random, the data reveal that students who omit more questions do not improve their hit-rates disproportionately. A regression of hit-rate differentials on omit-rate differentials reveals only a weak relationship between the two margins, with a slope of just 0.08 (plotted as a red line in Figure 1a).

Second, despite this modal pattern, there is meaningful heterogeneity in selection quality across students. Hit-rate differentials span from -0.5 to 0.5, with a raw standard deviation of 0.25 (Figures 1a and A5). Dispersion is confirmed by an Empirical Bayes model (Walters, 2024), which adjusts for limited observations per student. The implied mixing distribution (Figure A6) concentrates in [-0.3, 0.3], and retains a meaningful standard deviation of 0.1. AUC scores show similar variation, ranging from 0.25 to 1 with a standard deviation of 0.2 (Figure 2). Moreover, the two measures are connected: students with higher AUCs exhibit larger treatment effects on hit-rates. Interacting the penalty treatment with standardized AUC, shows that a one standard deviation increase in AUC raises the hit-rate differential

by 0.04, and by 0.06 when controlling for baseline accuracy (Table A3). These effects are large relative to the average treatment effect (respectively around 1 and 1.5 of the ATE) and statistically significant under both clustering strategies.

Third, there are systematic gender differences in selection behavior. Consistent with Iriberri and Rey-Biel (2021) and the broader literature on gender in test-taking, female students omit more questions under penalties – on average 0.87 more than males (Table A4). This reflects both higher baseline omission (+0.31) and a stronger response to penalties (+0.56). These patterns persist after controlling for risk aversion, overconfidence, and school grades (Figure A7a). However, the additional omission does not yield a higher hit-rate: the interaction effect of gender on hit-rate is statistically indistinguishable from zero, and this is robust to adding demographics (Figure A7b). Thus, while males and females are equally (in)accurate on average, they resolve the quantity-quality tradeoff differently. As explained below, this is a novel channel through which penalties for errors reduce female performance.

4 A Model of Imperfect Metacognition

Taken together, the results of the previous Section show that contrary to the NOWS principle efficiency in question selection cannot be taken for granted, but rather reflects a specific skill. I now write down a parametric model that captures how imperfect metacognition shapes the relationship between the quantity and quality margins of selection, offering an explanation for the empirical puzzles documented above.

4.1 Binary Metacognition

Consider a minimal version of the general framework, where success probabilities take only two values: high $\overline{p}$ and low $\underline{p}$, common across the population. In a multiple-choice setting, $\underline{p}$ may capture the probability of guessing correctly when clueless, while $\overline{p}$ may reflect an attempt informed by knowledge of the subject matter. For each student-type, a single parameter $\pi_s \in (0,1)$ characterizes the success probability distribution, denoting the fraction of high-success questions they face.

Departing from Rank-Preserving Confidence, I now allow for a stochastic confidence map. In particular, assume two confidence levels, again high $\overline{c}$ and low $\underline{c}$. In this binary setting, for each student-type, the confidence map is characterized by two parameters, $\psi_s^-, \psi_s^+ \in (0,1)$, such that

$$\psi_s = \begin{array}{c|cc} & \underline{c} & \overline{c} \\ \hline \underline{p} & \psi_s^- & 1-\psi_s^- \\ \overline{p} & 1-\psi_s^+ & \psi_s^+ \end{array}$$

The parameter ψ_s^+ is the probability of feeling high confidence when facing a high-success item (true positive rate), while ψ_s^- is the probability of feeling low confidence on a low-success item (true negative rate). An interpretation is that ψ_s^+ captures the skill of knowing what one knows, whereas ψ_s^- the skill of knowing what ones does not know. Together, they summarize the student's metacognitive skill.[11]

Motivated by literature on the measurement of metacognition (Fleming and Lau, 2014), I associate two metrics to the confidence map: *sensitivity* and *bias*. Sensitivity is defined subtracting one to the sum of the true positive and true negative rates,

$$\psi_s^\Sigma = -1 + (\psi_s^+ + \psi_s^-).$$

This statistic captures the overall association between confidence and underlying success probability, which is known as Youden's J in the context of binary classification. When confidence is statistically independent of success probability, we have $\psi_s^+ = \psi_s^- = 1/2$, implying $\psi_s^\Sigma = 0$. Taking this as a reference, sensitivity can be *good* ($\psi_s^\Sigma > 0$, better than random) or *bad* ($\psi_s^\Sigma < 0$, worse than random). By contrast, in standard models, Rank-Preserving Confidence implies $\psi_s^\Sigma = 1$, a perfect co-movement. Sensitivity, therefore, unfolds a dimension that is entirely suppressed at the benchmark.

Metacognitive bias is defined, dually, as the difference between the true positive and true negative rates,

$$\psi_s^\Delta = \psi_s^+ - \psi_s^-.$$

This parameter captures the student's average displacement between confidence and success at a given level of accuracy (higher values imply greater average confidence). In particular, ψ_s^Δ recovers over- and under-confidence within the new framework. Crucially, however, these notions are now disentangled from the ability to discriminate between high and low success levels.

For technical convenience, I assume that students follow a threshold rule. This implies that, under penalties, students fall into two groups, those who attempt questions regardless of the confidence level and those who answer only when confidence is high, capturing difference in risk attitude.

4.2 The Geometry of Selection

My first result shows that sensitivity determines the efficiency with which a student resolves the tradeoff between quantity and quality at the core of question selection. Specifically, whenever selection is non-trivial, the product of the hit-rate differential and the fraction of questions answered approximates sensitivity in the long-test limit, up to a scaling factor

[11] This echoes the notion of positive and negative introspection in modal logic and computer science (e.g. Fagin et al., 2004)

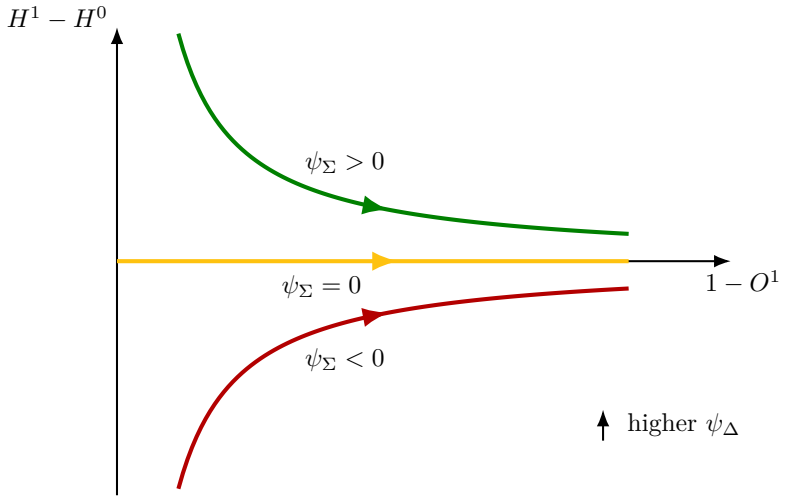


Figure 3: The Metacognition Frontier

which gauges the volatility of success probabilities.

Proposition 4. *Assume Parallel Testing and a threshold rule. In the binary model, for each student-type s, either $O_N^1 \xrightarrow{N\to\infty} 0$ or*

$$\underbrace{(H_N^1 - H_N^0)}_{\textit{quality}}\underbrace{(1 - O_N^1)}_{\textit{quantity}} \xrightarrow{N\to\infty} (\overline{p} - \underline{p})\mathbb{V}(\pi_s)\psi_s^\Sigma. \tag{3}$$

where $\mathbb{V}(\pi_s) = \pi_s(1 - \pi_s)$.

The Proposition follows by direct computation. Its interest lies rather in interpreting Equation (3) as a technological frontier for the selection problem. To see this, read the left-hand side as the realized surplus from selection, decomposed into a quality markup (or markdown) applied on the quantity of questions answered. The equation says that this quantity is bounded by a potential surplus that, as in canonical selection models (e.g., Roy, 1951), arises from variation in productivity – here captured by the variance $\mathbb{V}(\pi_s)$ times the spread between high and low success probability, $\overline{p} - \underline{p}$. On this potential, the sensitivity parameter ψ_s^Σ, acts as a signed conversion rate, which determines both whether selection is beneficial or harmful and the marginal gain or loss per question omitted.
Geometrically, when sensitivity is good, $\psi_s^\Sigma > 0$, the frontier is a convex decreasing hyperbola. Reducing the fraction of answered questions marginally improves selection quality with increasing returns. When sensitivity is bad $\psi_s^\Sigma < 0$, the geometry flips: increasing quantity "insures" against adverse selection, pushing the hit-rate towards the baseline. Finally, when $\psi_s^\Sigma = 0$ omitting questions has no impact on quality and selection is as good as random: the geometry of selection is "flat".
In this picture, metacognitive bias, ψ_s^Δ, pins down the student's position on the frontier: higher bias means feeling high confidence more frequently and hence more questions

answered, sliding the student to the right. Hence, conditional on sensitivity, bias changes the composition of the surplus in terms of quantity vs. quality, but leaves total efficiency (the height of the curve) unchanged.

In this sense, the binary model provides a natural lens through which to interpret the empirical puzzles in Section 3. Suppose sensitivity varies across students but is centered near zero. Then, in the geometry of Figure 3, most students would lie along the flat branch, with some dispersion around it. This would explain why penalties strongly affect omit-rates but leave hit-rates unchanged, and why the two margins appear largely uncorrelated, beyond directly accounting for shape of the AUC score distribution. Moreover, gender differences fit naturally into this framework. Higher omission rates among women without corresponding higher hit-rates are consistent with similar sensitivity but lower metacognitive bias relative to men. Below, I make these points quantitatively, via structural estimation.

The Hit-Omit Curve I conclude this Section by briefly describing how the insights of Proposition 4 extend beyond the binary case. Appendix C.1 studies what can be learned about metacognition for a general nonparametric confidence map.

When confidence takes more than two levels, what counts as "high enough" to answer depends on the selection threshold induced by costs. Varying the cost intensity then traces out a *hit-omit curve*, which links the hit-rate differential to the omit-rate across decision environments. For each point in the curve, the area of the bottom-right rectangle equals the selection surplus, directly generalizing Equation 3.

I show that the geometry of this curve characterizes metacognition: increasingly strong notions of stochastic monotonicity of the confidence map correspond to increasingly strong guarantees on surplus – from non-negativity on average, to pointwise non-negativity, to monotone quality gains as omission increases (Proposition A8). Moreover, I show that the returns to metacognition are amplified when success probabilities become more polarized, formalized through a notion of spread for their distribution (Proposition A9). Together, these results pin down what features of metacognition are identified from observable selection behavior in the general case.

5 Structural Estimation

I next structurally estimate the binary metacognition model, quantifying the qualitative picture developed above.

5.1 Statistical Model

I restrict to students who omit no questions in the absence of penalties, to maximize adherence to the theory. This yields a subsample of $N = 7{,}673$ individuals. For each student i in this subsample, the binary model predicts four counts: h_{i0}, the number of correct answers under no penalties; h_{i1}, the number of correct answers under penalties; o_{i1}, the number of omitted answers under penalties; and $w_{i1} = 12 - h_{i1} - o_{i1}$, the number of wrong answers under penalties. Given individual parameters $(\pi_i, \psi_i^+, \psi_i^-)$ and fixed the low and high success probability $(\underline{p}, \overline{p})$ as primitives, the likelihood for the no-penalty block is just

$$h_{i0} \sim \text{Binomial}(13, p_{i0}) \quad \text{where} \quad p_{i0} = \pi_i \overline{p} + (1 - \pi_i)\underline{p}.$$

For the penalty block, there are two cases, depending on whether we observe omissions. In the first case, $o_{i1} > 0$, the model implies that the student must be omitting questions at low confidence levels, hence a trinomial likelihood

$$(h_{i1}, o_{i1}, w_{i1}) \sim \text{Multinomial}(12, p_{i,h}, p_{i,o}, p_{i,w})$$
$$\text{where} \quad \begin{cases} p_{i,h} = \pi_i \psi_i^+ \overline{p} + (1 - \pi_i)(1 - \psi_i^-)\underline{p}, \\ p_{i,o} = (1 - \pi_i)\psi_i^- + \pi_i(1 - \psi_i^+), \\ p_{i,w} = 1 - p_{i,h} - p_{i,o}. \end{cases}$$

In the second case, $o_{i1} = 0$, there are two observationally consistent models, corresponding to the student's risk attitude. One possibility is that the student follows the same selective rule as above, but draws high confidence on all 12 questions, by chance. In this case, we still have the same trinomial likelihood. The other possibility is that the threshold is sufficiently low that student attempts all questions regardless of confidence (e.g., because of lower aversion to risk). In this case the penalty-block likelihood is analogous to the no penalty block

$$h_{i1} \sim \text{Binomial}(12, p_{i0}) \qquad \text{where} \qquad p_{i0} = \pi_i \overline{p} + (1 - \pi_i)\underline{p}.$$

To handle this ambiguity, I use a mixture likelihood for the penalty block.
The test is short ($n_0 = 13$, $n_1 = 12$), so individual-level likelihoods are noisy. I therefore impose a hierarchical prior that (i) regularizes individual parameters via partial pooling and (ii) allows parameters to be correlated in the population. I first map parameters to $\mathbb{R}$ via the logit transform,

$$\eta_i \equiv \Big(\text{logit}(\pi_i),\, \text{logit}(\psi_i^+),\, \text{logit}(\psi_i^-)\Big).$$

To capture gender heterogeneity, I let $g_i \in \{0, 1\}$ indicate whether student i is female and assume a random-effects structure

$$\eta_i \mid g_i \;\sim\; \mathcal{N}(\mu_0 + \delta \cdot g_i,\; \Sigma)\,,$$

where $\mu_0 \in \mathbb{R}^3$ is the male mean on the logit scale, $\delta \in \mathbb{R}^3$ is the female–male shift, and Σ is an unrestricted 3×3 covariance matrix. I also allow the mixture weight to differ by gender, and re-parameterize it on the logit scale obtaining a vector $(\alpha_0, \alpha_1) \in \mathbb{R}^2$. I place weakly informative normal priors on all model hyperparameters.[12]

To estimate the model, I set the low success probability to capture random guessing, ($\underline{p} = 0.2$ since each question has five candidate answers in the experiment) and the high success probability to capture perfect knowledge of the answer ($\overline{p} = 1$). This choice is intuitive, as it corresponds to the boundaries of the interval of admissible success probabilities. Moreover, it is conservative, as Equation 3 suggests that setting $\overline{p} - \underline{p}$ at its maximum minimizes the variation in sensitivity needed to explain a given pattern of behavior.

5.2 Results

Figure 4 plots the estimated means of model parameters by gender, with 95% posterior credible intervals. Females and males seem to have comparable productivity and sensitivity in the test. For both genders, the distribution of success probabilities is right-skewed and averages around 0.3 (0.33 for males, 0.29 for females), whereas sensitivity follows a bell-shaped distribution centered just above zero, again with modest differences in the mean (0.12 for males, 0.08 for females).

The main gender gap lies in the structure of metacognition. Males exhibit a higher true positive rate (0.66 vs. 0.59), suggesting a greater ability to recognize questions they know and answer them. Females, in contrast, show a higher true negative rate (0.49 vs. 0.46), indicating stronger awareness of what they do not know. This asymmetry results in a pronounced gap in metacognitive bias: males are more positively biased than females (0.21 vs. 0.11) – the largest gender difference in the model.

Figures A8 and A9 plot the full distributions for the model parameters. All of them show meaningful dispersion, with overall standard deviations of 0.27 for sensitivity, of 0.38 for bias ans of 0.17 for productivity. Moreover, the three parameters show moderate positive correlations. In particular, productivity correlates at 0.16 with bias and 0.35 with sensitivity, showing that more productive students tend to be also better at sorting questions, in particular identifying questions they can hit (Table A5). All these figures

[12] Priors: $\mu_0 \sim \mathcal{N}(0, 4I_3)$, $\delta \sim \mathcal{N}(0, 4I_3)$, and $\Sigma = \mathrm{diag}(\tau)\Omega\mathrm{diag}(\tau)$ with $\tau_k \sim \mathcal{N}^+(0, 1)$ and $L_\Omega \sim$ LKJ(2). Estimation uses the non-centered parameterization $\eta_i = \mu_0 + \delta \cdot g_i + \mathrm{diag}(\tau) L_\Omega z_i$ with $z_i \sim \mathcal{N}(0, I_3)$ i.i.d. For the mixture weights, $\alpha_0 \sim \mathcal{N}(0, 4)$ and $\alpha_1 \sim \mathcal{N}(0, 4)$.



Figure 4: Means of Model Parameters by Gender

are similar between genders.[13]

These findings confirm the qualitative picture developed above and quantify how the model accounts for the empirical patterns. To validate the interpretation of the metacognitive parameters, I next show that they not only fit observed behavior, but also predict variation in outcomes not used in estimation.

First, I validate the role of sensitivity as selection skill by linking it to AUC, an independent measure of the same feature. Regressing normalized AUC scores on the normalized posterior means of productivity, sensitivity, and bias I find that a one standard deviation increase in sensitivity predicts a 0.235 standard deviation increase in AUC and the effect is highly significant (Table A6). In contrast, the coefficients on productivity and bias are small and not significant. This confirms that sensitivity captures the student's ability to discriminate between easier and harder items, independent of productivity or response rate.

Second, I turn to a survey-based measure of overconfidence, defined as the gap between the student's forecasted and actual number of correct answers, normalized in the population. Regressing this measure on the same set of parameters, I find that higher productivity and sensitivity are associated with less overconfidence (-0.405 and -0.151), while higher bias predicts more (+0.153). These patterns are again in line with theoretical predictions.

[13]Posterior means and densities of the individual-level parameters (Figures A10, A11, and A12) confirm this pattern.

6 Metacognition and Test Scores

I next explore the impact of metacognition on test scores. Penalties for errors generate a complementarity between productivity and positive metacognitive bias: the more frequent high-success questions are, the more valuable it becomes to recognize them compared to low-success ones. This asymmetry is even sharper under the common "formula scoring" design that implies no reward for the true negative rate. Combined with my structural estimates, this structure predicts a gender gap in scores that should widen in productivity. The data confirm this prediction and rule out competing explanations.

6.1 Returns to Metacognition

Generalizing from the experimental setting, consider a scoring rule that awards $h > 0$ points for correct answers, 0 for incorrect ones, and $o \in (0, h)$ for omissions. I next decompose the student's average score per question, W_N^1, in terms of the primitives of the binary model, focusing on the nontrivial case where the student omits a positive fraction of questions. Towards the statement, define $g(p)$ as the expected change in score from answering high-probability questions and omitting low-probability ones, i.e. $g(\overline{p}) = h\overline{p} - o$ and $g(\underline{p}) = o - h\underline{p}$.

Proposition 5. *In the binary model, for each student-type s, either $O_N^1 \xrightarrow{N\to\infty} 0$, or*

$$W_N^1 \xrightarrow{N\to\infty} K(\pi_s) + \frac{1}{2}\left[A(\pi_s)\psi_s^{\Sigma} + B(\pi_s)\psi_s^{\Delta}\right] \tag{4}$$

where $A(\pi_s) = \mathbb{E}_{\pi_s}[g(p)]$ and $B(\pi) = h\mathbb{E}_{\pi_s}[p] - o$.

The expression in Proposition 4 admits a transparent interpretation. The term $K(\pi_s)$ is the baseline score a student would earn if they chose at random whether to answer each question – i.e., under purely noisy metacognition ($\psi_s^+ = \psi_s^- = 1/2$). The remaining term is a *metacognitive premium*, composed of two parts. In the first, $A(\pi_s)$ captures the marginal return to metacognitive sensitivity. If $g(p) \geq 0$, so that omitting low-success questions and answering high-success ones improves scores, this term reflects the student's expected value of making the "right call." In this case, the return grows with the spread of success probabilities, echoing the role of variance in selection gains (cf. Equation 3). The second term, $B(\pi_s)$, captures the return to metacognitive bias. When attempting questions is, on average, more rewarding than skipping, a positively biased student earns a higher score. This return increases with the expected success probability, highlighting a complementarity between productivity and bias.

In practice, even more may be true. Many scoring rules – including the one used in the experiment – follow *formula scoring* (Thurstone, 1919), where penalties are calibrated so

that the expected payoff from omitting a question equals that from random guessing.[14] As in my structural estimation, random guessing provides a natural benchmark for the low success probability. Intuitively, then, under formula scoring, correctly identifying low-success questions and skipping them does not improve scores. Formally, the marginal return to the true negative rate,

$$\frac{\partial W_N^1}{\partial \psi_s^-} = \pi_s(o - h\underline{p}), \tag{5}$$

equals zero in this case. In fact, the return may turn negative if students guess better than random – for instance, by ruling out implausible answers. Under formula scoring, then, test scores reward productivity and the ability to identify questions one knows, while systematically neglecting – or even penalizing – the ability to recognize ignorance.

Complementarity beyond Binary In Appendix C.2, I extend the complementarity between metacognitive bias and productivity to the nonbinary setting, formalizing positive bias via an increasing-differences condition on the CDFs of two confidence maps. In particular, Proposition A10 shows that for sufficiently productive students, greater positive bias yields larger score gains at higher productivity levels. The proof of this result also shows that a perfectly calibrated confidence map (no bias) need not maximize expected score. In particular, when thresholds are overly conservative (e.g. as in the case of risk aversion – Appendix A) any maximal profile involves a degree of overconfidence (Corollary 1).

6.2 The Widening Gender Gap

I next use the model to explain gender differences in test scores. As shown in Section 5.2, males and females exhibit similar distributions of productivity and sensitivity, but differ in metacognitive bias: males are more positively biased, a difference driven by higher true positive rates and lower true negative rates. Combined with the complementarity between productivity and bias in Equation 4 (and the fact that formula scoring neglects the ability to recognize low-success questions) the model predicts a gender gap in scores that should widen with productivity in the test.

This prediction is confirmed in the data. Figure 5 plots average scores per question by quintile of no-penalty hit-rate (used as a proxy for productivity), separately by gender. The gender gap more than triples from the bottom to the top quintile, increasing from 4% to 12% of the population mean score per question.

By contrast, as also noted by Iriberri and Rey-Biel (2021), this pattern is difficult to reconcile with preference-based explanations such as gender differences in risk aversion.

[14] With five answer options, this holds when $h = 5$ and $o = 1$, as in the experiment.

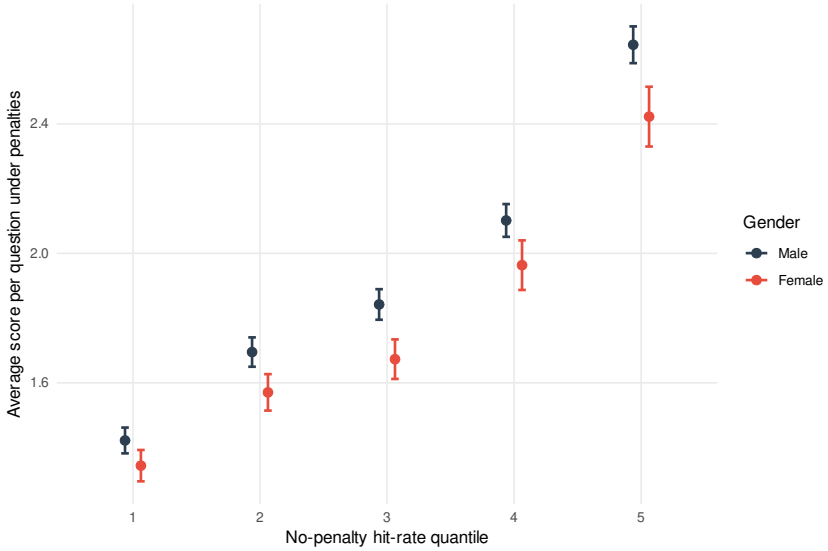


Figure 5: The Widening Gender Gap in Test Scores

Explaining the widening gap would require not only that females are more risk averse – a debated claim (e.g., Croson and Gneezy, 2009; Filippin and Crosetto, 2016) – but also that risk aversion is positively correlated with productivity in the test. Yet the data show no such relationship: the survey measure of risk aversion and no-penalty hit-rate are essentially uncorrelated for both males (correlation: -0.04) and females (-0.06).[15]

7 Policy Implications

In this Section, I use the estimated model to simulate counterfactual scenarios that quantify how variation in metacognition shapes both overall score dispersion and the gender gap. Both exercises reveal a first-order role of metacognition for outcomes. Motivated by this finding, I comment on the fairness and informativeness of alternative scoring rules.

7.1 Decomposing the Observed Variation

I first use Equation 4, together with my model estimates, to assess the relative contribution of productivity and metacognition to score variation under penalties.

To this end, I compute the reduction in the variance of average scores by "eliminating heterogeneity" along the two dimensions of the model. I begin by calculating the baseline score variance using posterior draws from the estimated model. I then simulate three counterfactual scenarios: (i) fixing each student's productivity parameter π_i at its sample mean while allowing metacognition to vary; (ii) fixing metacognitive parameters (ψ_i^+, ψ_i^-) at their sample means while allowing productivity to vary; and (iii) fixing all three

[15] More broadly, the literature finds risk aversion negatively associated with cognitive skill (e.g., Benjamin et al., 2013; Dohmen et al., 2010).

parameters at their respective means. Because scores depend nonlinearly on the primitives, I decompose the total variance reduction using Shapley shares, i.e. averaging each component's marginal contribution across the two possible elimination orders.

The results show that while productivity is the dominant driver of score dispersion, heterogeneity in metacognition also plays a first-order role. Eliminating heterogeneity in π alone reduces variance by 76%; eliminating heterogeneity in (ψ^+, ψ^-) alone reduces it by 48%; eliminating both reduces it by 97%. Hence, the Shapley decomposition attributes roughly 62% of baseline score variance to productivity and 35% to metacognition.

7.2 Decomposing the Gender Gap

Second, I conduct a posterior-predictive decomposition of the gender gap using draws from the estimated model. For each posterior draw, I map students' latent parameters into expected scores using Equation 4, and compute the model-implied gender gap by comparing average predicted scores for males and females. As shown in Figure A13, the model closely replicates the observed gap across quintiles of no-penalty hit-rate. I then perform a sequence of counterfactual "women-to-men" replacements to decompose the gap. Holding male parameters fixed, I sequentially replace women's values of π_i, ψ_i^+, and ψ_i^- with the corresponding male group means, recomputing the implied gender gap in each case and averaging across draws. The resulting change isolates the contribution of each component-productivity and the two metacognitive channels-to the overall gap.

The decomposition reveals that differences in productivity explain about 23% of the gap, while differences in metacognition account for roughly 41%. Recall that, under my parametrization, the latter are driven entirely by the true positive rate. Hence the role of bias is potentially understated: it would be higher, for values of the low success probability higher than random guessing. This confirms that metacognitive differences a major contributor to the gender gap in scores, suggesting an important role for training policies that act on female confidence.

7.3 Calibrated Penalties

My results show that variation in metacognitive skill, is a first-order driver of score dispersion, and that differences along this dimension (particularly bias) are the main driver of the widening gender gap in performance. This points to clear policy implications for the design of multiple-choice scoring rules, both in terms of informativeness and fairness.

From an informational perspective, no-penalty scoring is optimal when the goal is to measure productivity. In this regime, scores increase monotonically with average success probability, making the metric both interpretable and equitable. Introducing penalties, by contrast, activates a confounding effect of metacognitive skill, which my estimates reveal

to be sizable.

In some settings, however, metacognitive sensitivity may itself be of interest: think of medicine, law, or other high-stakes decision-making domains. In those cases, penalties are necessary to measure a student's capacity to judge when to act and when to abstain. Once penalties are introduced, their magnitude becomes central. As shown in Equation 4, low penalties reward students with positive imbalance and high productivity, while high penalties advantage those with negative imbalance and low productivity. While this structure may improve informativeness, it raises fairness concerns whenever metacognitive bias is correlated with group identity, as for gender in my setting.

A natural question, then, is whether one can design scoring rules that reward both knowledge and metacognitive accuracy, while remaining neutral to bias. Proposition 5 suggests a simple solution: calibrate the omission payoff to eliminate returns to imbalance. One implementation is a two-stage test. In the first stage, students complete a no-penalty test. In the second, they face penalties, with the omission score set equal to their average score per question in stage one, so that $o = h\mathbb{E}[p]$ and $B = 0$.

This adjustment ensures that the incentive to omit scales with student ability, neutralizing the mechanical advantage or disadvantage from bias. While implementation may be imperfect, especially for students at the extremes, the conceptual insight remains: aligning fairness and informativeness requires tailoring the value of skipping questions to individual skill.

8 Conclusion

In this paper, I argued that under penalties for errors, time pressure or more general selection costs, standardized tests measure not only raw productivity but also metacognition – the higher-order skill of monitoring one's own capabilities. In particular, metacognition determines efficiency in question selection and gender differences along this dimension help explain why the gender gap in test performance widens with ability. I also proposed a Bayesian foundation for metacognitive imperfections, linking them to cognitive noise.

While standardized tests provide a clear illustration of how metacognition influences performance, the implications of my framework extend beyond educational assessments. First, with minor modifications, the model can be applied to task selection in the workplace. More broadly, its core insight – that imperfect metacognition leads to a misperception of one's own ability schedule across tasks – may be relevant in a variety of labor settings. For instance, it may distort the allocation of effort across activities, thereby shaping allocative skill (Caplin et al., 2023). Similarly, the realization of comparative advantage in collaborative environments may depend not only on social skills that reduce frictions while trading tasks (Deming, 2017), but also, upstream, on individuals' ability to accurately assess their own competencies and identify which tasks to exchange. Finally, in

organizations structured around knowledge hierarchies (Garicano, 2000), recognizing the limits of one's expertise may be essential for efficiency, as it triggers the upward referral of problems.

References

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2023). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics 138*(1), 1–35.

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

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.

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*.

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.

Benjamin, D. J. (2019). Errors in probabilistic reasoning and judgment biases. *Handbook of Behavioral Economics: Applications and Foundations 1 2*, 69–186.

Benjamin, D. J., S. A. Brown, and J. M. Shapiro (2013). Who is behavioral? cognitive ability and anomalous preferences. *Journal of the European Economic Association 11*(6), 1231–1255.

Blackwell, D. (1953). Equivalent comparisons of experiments. *The annals of mathematical statistics*, 265–272.

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.

Croson, R. and U. Gneezy (2009). Gender differences in preferences. *Journal of Economic literature 47*(2), 448–474.

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.

Cusumano, C. M. and F. Pieroth (2024). Due diligence in common value auctions.

Dasgupta, S. (2023). Screening knowledge. *Mimeo*.

Deming, D. J. (2017). The growing importance of social skills in the labor market. *The Quarterly Journal of Economics 132*(4), 1593–1640.

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.

Dohmen, T., A. Falk, D. Huffman, and U. Sunde (2010). Are risk aversion and impatience related to cognitive ability? *American Economic Review 100*(3), 1238–1260.

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.

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.

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., 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.

Fagin, R., J. Y. Halpern, Y. Moses, and M. Vardi (2004). *Reasoning about knowledge*. MIT press.

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

Filippin, A. and P. Crosetto (2016). A reconsideration of gender differences in risk attitudes. *Management Science 62*(11), 3138–3160.

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*.

Garicano, L. (2000). Hierarchies and the organization of knowledge in production. *Journal of political economy 108*(5), 874–904.

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.

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.

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.

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.

Kolotilin, A. (2018). Optimal information disclosure: A linear programming approach. *Theoretical Economics 13*(2), 607–635.

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

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.

Milgrom, P. R. and R. J. Weber (1982). A theory of auctions and competitive bidding. *Econometrica: Journal of the Econometric Society*, 1089–1122.

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.

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.

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

Read, D., G. Loewenstein, M. Rabin, G. Keren, and D. Laibson (2000). Choice bracketing. *Elicitation of preferences*, 171–202.

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.

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.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford economic papers 3*(2), 135–146.

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

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.

Wainer, H. and D. Thissen (1994). On examinee choice in educational testing. *ETS Research Report Series 1994*(2), 159–195.

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 Examples of Portfolio Problems

In this Section of the Appendix, I show how the Portfolio Problem captures decision making under two important sources of selection costs: penalties for errors and capacity constraint.

A.1 Penalties for Errors

Consider a scoring rule that awards h points for correct answers, m for errors and $o \in (m, h)$ for skipped questions. A natural assumption is that the student maximizes subjective expected utility over total score levels. Fix an arbitrary vector of confidence levels $(c_j)_{j=1}^N$. Observe that, given this vector, when the student selects portfolio A they believe total score is the random variable

$$W(A) = o(N - |A|) + m\left(|A| - \sum_{j \in A} B_j\right) + h \sum_{j \in A} B_j \tag{6}$$

where B_j is a Bernoulli of parameter c_j. Hence, for a utility u they solve

$$V((c_j)_{j=1}^N) \max_{A \subseteq \{1,\dots,N\}} \mathbb{E}[u(W(A))] \tag{7}$$

I next show that when such utility risk neutral or CARA the problem of the student can be rewritten as an instance of the Portfolio Problem. Moreover, in both cases the optimal decision rule is a threshold rule.

Risk Neutrality In this case, we can just take u to be the identity. Hence, the conclusion follows immediately, taking expectations in Equation 6. In particular, define $A^+ = \sum_{j \in A} B_j$. Then the objective rewrites, dropping constants

$$W(A) = \underbrace{(h-m)\mathbb{E}[A^+]}_{U(A^+)} - \underbrace{(o-m)|A|}_{k(|A|)}$$

which is an instance of the Portfolio Problem. To derive the threshold on confidence that defines the optimal portfolio, we can impose that the student is indifferent between the portfolio associated to the highest J confidence levels and the one associated to the highest $J+1$. Then, we obtain that the student answers a question iff the associated confidence is higher than

$$\hat{c} = \frac{o-m}{h-m}$$

which is simply the ratio of the incentives to skip and to hit a question relative to committing an error.

CARA In this case $u = 1 - e^{-\alpha W}$ for a risk aversion coefficient $\alpha > 0$. Then, using the moment generating function of the Bernoulli distribution, since outcomes across questions are believed to be independent, we obtain that the objective rewrites

$$\mathbb{E}\left[1 - e^{-\alpha W(A)}\right] = 1 - \mathbb{E}\left[e^{-\alpha(aA^+ + b)}\right] = 1 - e^{-\alpha b}\mathbb{E}\left[e^{-\alpha a A^+}\right]$$

where $a = (h - m)$ and $b = No + |A|(m - o)$ to shorten notation. In turn, applying to the last line the increasing function $-\log(1 - x)$ and dropping constants, we obtain that maximizing Problem 7 is equivalent to maximizing

$$\underbrace{\mathbb{E}[-e^{-\alpha(h-m)A^+}]}_{U(A^+)} - \underbrace{\alpha(o - m)|A|}_{k(|A|)}$$

which is again an instance of the Portfolio Problem. Similarly to before, we obtain a threshold rule for the confidence cutoff

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

Note that the cutoff level is monotone in the risk aversion coefficient α.

Other Examples Some preference specifications fall outside the Portfolio Problem class but still imply the monotonicity result of Proposition 1, so that the core results go through. One such case is when students maximize the expected value of a strictly increasing utility function, even if they are not risk-neutral or CARA. More generally, the conclusion holds whenever the value function over portfolios is non-separable but remains monotone with respect to first-order stochastic dominance (FOSD), conditional on portfolio size. Formally:

$$|A| = |A'| \quad \text{and} \quad A^+ \preceq_{\text{FOSD}} A'^+ \quad \Rightarrow \quad V(A) \leq V(A').$$

Beyond this, monotonicity arises in models where students evaluate each question in isolation as in narrow bracketing (Read et al., 2000). In such cases, each question i is treated as a separate decision problem

$$V_i\left((c_j)_{j=1}^N\right) = \max_{a_j \in \{0,1\}} a_j \cdot [c_j u(h) + (1 - c_j)u(m)] + (1 - a_j)u(o),$$

where $a_j \in \{0, 1\}$ represents the binary choice of attempting or omitting the item. This again leads to threshold behavior, with a uniform threshold given by $\frac{u(h)-u(m)}{u(o)-u(m)}$, and thus preserves the monotone structure of selection.

A.2 Capacity Constraint

Consider a test administered under severe time pressure or with an examinee choice design. In these settings, students are permitted to attempt only a fixed fraction κ of the total number of questions. This can be modeled as a degenerate instance of the Portfolio Problem, with the following cost function

$$k(|A|) = \begin{cases} 0 & \text{if } |A| \leq \kappa N, \\ \infty & \text{if } |A| > \kappa N. \end{cases}$$

Under this constraint, students select the top κN questions according to their confidence levels. While the selection rule is still monotone in confidence, it no longer takes the form of a fixed threshold: the cutoff varies with the realized distribution of confidence across questions. In this sense, the decision rule is quantile-based rather than threshold-based.

B Bayesian Metacognition

In this Section of the Appendix, I propose a Bayesian microfoundation for the distribution of success probabilities and for the confidence map.

B.1 Statistical Decision Problem

I model the test as a decision problem, in the sense of Blackwell (1953). For each question $j = 1, \dots, N$ the state of the world ω_j represents the correct answer. The state is drawn independently across questions from a finite set Ω, according to a full-support distribution $d_0 \in \Delta(\Omega)$.[16] Actions $\mathcal{A} = \Omega \cup \{0\}$ correspond to candidate answers, plus the option to skip the question. An outcome function $\rho : \mathcal{A} \times \Omega \to \mathbb{R}$ determines the score obtained in each question,

$$\rho(a, \omega) = \begin{cases} h & \text{if } a = \omega \\ o & \text{if } a = 0 \\ m & \text{otherwise} \end{cases} \qquad \text{with} \quad h > o \geq m.$$

Payoffs are defined over streams of outcomes, $u\left((\rho(a_i, \omega_i))_{j=1}^N\right)$. The payoff function u is such that, when the agent's belief about the correct answer is given by a distribution $d \in \Delta(\Omega)$, the *subjectively optimal action* $\hat{a}(d)$ selects the mode of d if confidence is sufficiently high, and the outside option otherwise,

$$\hat{a}(d) = \text{mode}(d) \cdot \mathbf{1}\{d(\{\text{mode}(d)\}) > \hat{c}\}, \quad \text{where} \quad \hat{c} \in [0, 1]. \tag{8}$$

(examples of preferences that satisfy this property are collected in Appendix A). Hence, overall, payoffs describe a (generalized) hypothesis testing problem, enriched with an outside option.

B.2 Cognitive and Metacognitive Signals

To model metacognition in this setting, I distinguish between two sources of information: *first-* and *second-order*. Before taking an action, the decision maker observes a cognitive signal s_ω about the state of the world ω, drawn from a finite set S_Ω. This signal is generated by a Blackwell experiment whose informativeness is indexed by a question-specific *type* $\theta \in \Theta$, capturing heterogeneity in the quality of the decision maker's reasoning across

[16] For instance, the distribution may represent the frequency of correct answer labels within a question database.

questions. Formally, I consider a map

$$\sigma : \Omega \times \Theta \to \Delta(S_\Omega),$$

and posit that types are drawn from a full-support distribution $t_0 \in \Delta(\Theta)$ independently from each other and from the realization of the state of the world.[17]

Crucially, I posit that the student does not observe the question-specific realization of θ. Therefore they are uncertain about the experiment that generated the observed cognitive signal.

I model imperfect metacognition, as the partial resolution of such second-order uncertainty. In particular, I assume that the decision maker receives a noisy metacognitive signal s_θ about the realized informativeness type θ, drawn from a *meta-experiment*

$$\mu : \Theta \to \Delta(S_\Theta).$$

Given the cognitive and the metacognitive signal, the student updates their belief about the correct answer using Bayes' rule. In particular, the experiments σ and μ, together with the prior $d_0 \otimes t_0$, induce the probability measure

$$\mathbb{P}(\omega, \theta, s_\omega, s_\theta) = d_0(\omega) t_0(\theta) \sigma(s_\omega \mid \omega, \theta) \mu(s_\theta \mid \theta) \tag{9}$$

and the student's belief about ω after observing the signal pair (s_ω, s_θ) is just the (marginalized) Bayesian posterior $d_{s\omega}^{s_\theta} \equiv \mathbb{P}(\omega | s_\omega, s_\theta)$.

B.3 Success Probabilities and Confidence Levels

This setting allows for a Bayesian definition of success probabilities and confidence levels. To this end, the first ingredient is an *ex-interim* notion of "success" which applies to the stage after updating beliefs, but before action is taken. Given beliefs $d \in \Delta(\Omega)$, define the *forced-optimal action* $a^\circ(d)$ as the action the decision maker would take, were the outside option made unavailable – so that they maximized over $\mathcal{A} \setminus \{0\}$. Also, define the *ex-post optimal action* $a^*(\omega)$ as the action which would be taken if the state of the world was observed. Then, *success* is the event where the posterior forced-optimal action and the ex-post optimal action coincide,

$$E = \left\{ (\omega, \theta, s_\omega, s_\theta) \mid a^*(\omega) = a^\circ \left(d_{s_\theta}^{s_\omega} \right) \right\}.$$

[17] Independence reflects the idea that correct answer to a question should not systematically relate to the quality of the decision maker's cognitive processing for that question. For instance, answer labels can be randomized across questions, effectively severing any statistical dependence between the correct option and the student's ability to solve the problem.

For the sake of exposition, I refer to the realization of the informativeness type and the signals in a question as its *sort*. Fix a question of sort $(\theta, s_\omega, s_\theta)$. I define the *confidence level* in the question as the subjective posterior probability that the decision maker assigns to the success event E,

$$c(\theta, s_\omega, s_\theta) = \mathbb{P}(E \mid s_\omega, s_\theta).$$

I define the *success probability* in the question as the limit frequency of success across all questions of the same sort when the number of questions N diverges[18]

$$p(\theta, s_\omega, s_\theta) = \operatorname*{plim}_{N\to\infty} \frac{\sum_{i=1}^{N} \mathbf{1}\{(\theta^i, s_\omega^i, s_\theta^i) \in E\} \cdot \mathbf{1}\{(\theta^i, s_\omega^i, s_\theta^i) = (\theta, s_\omega, s_\theta)\}}{\sum_{i=1}^{N} \mathbf{1}\{(\theta^i, s_\omega^i, s_\theta^i) = (\theta, s_\omega, s_\theta)\}}.$$

Given these definitions, we can map each instance of the information processing model into a pair of probability distribution and confidence map (π, ψ), recovering the primitive of Section 1. To this end, denote by $d^\theta_{s_\omega} = \mathrm{BU}(d_0, \sigma_\theta, s_\omega)$ the posterior of a (virtual) observer accessing the perfectly informative meta-experiment, $\mu^* : \theta \mapsto \delta_\theta$. By the Strong Law of Large Numbers we have that

$$p(\theta, s_\omega, s_\theta) = d^\theta_{s_\omega}(\mathrm{mode}(d^{s_\theta}_{s_\omega})) \quad \text{and} \quad c(\theta, s_\omega, s_\theta) = d^{s_\theta}_{s_\omega}(\mathrm{mode}(d^{s_\theta}_{s_\omega})). \tag{10}$$

Hence, the joint distribution for success probabilities and confidence levels can be obtained pushing forward the joint law in Equation 9 via these two functions,

$$(p, c)_\#[\mathbb{P}(\theta, s_\omega, s_\theta)] \in \Delta([0, 1]^2).$$

As a consequence, the distribution of success probabilities, π, is just the marginal of the pushforward on p, whereas the confidence map, ψ is the set of conditional distributions of c given p.

This construction clarifies the role of first-order and second-order information in my framework. Note that confidence levels depend only on the student's information set (s_ω, s_θ). Conditional on such a pair, different informativeness types θ may be realized for the first-order signal, driving variation in success probability. Hence, second-order information, which regulates the relation between θ and s_θ, introduces a "correlation wedge" between these two dimensions.

As a limit case, when the student observes perfect second-order information, confidence and success probability show perfect co-variation. This recovers within my framework models where agents observe cognitive signals of heterogeneous quality, as in cognitive uncertainty (e.g. Enke and Graeber, 2023).

[18] Note that each question sort realizes infinitely often, since $\mathbb{P}$ has finite full support.

B.4 Implementation Results

An immediate consequence of the previous construction is a necessary condition on the pair (π, ψ) to be compatible with Bayesian updating: confidence and success probability should form a martingale coupling.

Proposition A6. *Consider an instance (d_0, t_0, σ, μ) of the information processing model. The success probability distribution π and the confidence map ψ are such that*

$$\mathbb{E}_{(\pi,\psi)}[p \mid c] = c$$

i.e. (π, ψ) is a martingale coupling.

The result has two key implications. First, by the tower property, the confidence map exhibits no over- or under-confidence ($\mathbb{E}[p] = \mathbb{E}[c]$). Hence, the informativeness of the meta-experiments regulates correlation strength between confidence and performance, without affecting their average displacement. Second, success probabilities are more dispersed that confidence levels. Hence, the model displays *overinference* from weak cognitive signals and *underinference* from strong ones.
Then, a natural question is whether, in the converse direction, any martingale coupling of success probability and confidence level can be implemented in an information processing model. I show that this is the case assuming that (i) success probabilities are higher than random guessing and (ii) each action is ex-post optimal with the same frequency.

Proposition A7. *Suppose that d_0 is a discrete uniform distribution. Let (π, ψ) be a martingale coupling such that supp$\pi \subseteq [1/|\Omega|, 1]$. Then there exists a distribution of informativeness types t_0 a family of cognitive experiments σ and a meta-experiment μ such that*

$$(\pi, \psi) = (p, c)_{\#}[\mathbb{P}(\theta, s_\omega, s_\theta)].$$

Key to the construction is the notion of *symmetric cognitive experiment*. Such experiment have signal corresponding to states of the world, $S_\Omega = \Omega$, and are defined by

$$\forall \omega \neq \omega' \in \Omega \qquad \sigma(\omega \mid \omega) = \theta \quad \text{and} \quad \sigma(\omega' | \omega) = \frac{1-\theta}{|\Omega| - 1}$$

for some $\theta \geq 1/|\Omega|$. Hence, signal is diagnostic of the corresponding state but preserve prior likelihood ratios among the other states.
By independence of states and informativeness types, Bayesian updating can be performed hierarchically (first updating beliefs about the informativeness type via the meta-experiment μ, and then updating beliefs about the state ω using the expected experiment implied by the posterior on θ). Hence, symmetry implies that the probability of success

coincides with the informativeness type of the latent experiment, whereas confidence with the posterior mean about this parameter, given the meta-signal. The result then follows from standard techniques for the implementation of distributions of posterior means (e.g. Kolotilin, 2018).

C Nonparametric Results on Metacognition

In this Section of the Appendix, I derive nonparametric results on metacognition. First, I generalizes Proposition 4, characterizing what can be learned about the relationship between confidence and success probabilities from observed behavior. Second, I extend Proposition 5, showing a general form of the complementarity between productivity and metacognitive bias.

C.1 The Hit-Omit Curve

In the binary model, Proposition 4 establishes an identification between metacognition sensitivity and the shape of the technological frontier linking the hit-rate differential and the omit-rate under penalties. This is possible because, since confidence is binary, the model allows only two behaviors: either a student answers all questions or they skip questions in the low confidence state.

When there are more than two confidence levels, what counts as "high enough" confidence for answering depends on the structure of the Portfolio Problem – in particular, on the threshold implied by selection costs. For instance, as penalties grow, or pressure increases, students may require higher confidence to attempt a question. This suggests that to learn about metacognition, we must observe how selection behavior changes across different environments.

Imagine we could gradually tighten selection costs. As the selection threshold sweeps the $[0,1]$ interval, we would trace out a curve linking the hit-rate differential to the omit-rate at different levels of selectivity. I define the *hit-omit curve* as the large-N limit of this object. Formally, for a distribution of success probabilities $\pi \in \Delta([0,1])$ and a confidence map ψ, define

$$\Gamma(\pi, \psi) = \lim_{N\to\infty} \left(O_N^1(k), H_N^1(k) - H_N^0(k)\right)_{k\in[0,1]}.$$

Note that the area of the rectangle bottom-right of each point in the curve is just the selection surplus at a threshold k,

$$\gamma(k; \pi, \psi) = (1 - O_N^1(k))(H^1(k) - H^0(k)),$$

showing the connection to Equation (3).

My next results shows that the hit-omit curve Γ and the selection surplus function γ embed information about the structure of ψ. In particular, I show an equivalence between the shape of these objects and different notion of monotonicity for the confidence map ψ. Say that the ψ is, respectively, *(i) expectation monotone; (ii) stochastically monotone and (iii) hazard rate monotone* if $p \leq p'$ implies, respectively, *(i)* $\mathbb{E}_{\psi(p)}[C] \leq \mathbb{E}_{\psi(p')}[C]$; *(ii)* $\psi(p) \preceq_{FOSD} \psi(p')$ and *(iii)* $\psi(p) \preceq_{HRO} \psi(p')$.

Proposition A8. *The following set of equivalences hold:*

(i) ψ *is expectation monotone* $\iff$ *for all* π, $\int_0^1 \gamma(k;\pi,\psi)dk$ *is non-negative.*

(ii) ψ *is stochastically monotone* $\iff$ *for all* π, $\Gamma(\pi,\psi)$ *is non-negative.*

(iii) ψ *is hazard-rate monotone* $\iff$ *for all* π, $\Gamma(\pi,\psi)$ *is non-negative and non-decreasing.*

These three notions form a strict hierarchy: (iii) implies (ii), which implies (i). The stronger the structure on ψ, the stronger the guarantees on selection behavior. Expectation monotonicity ensures that surplus is positive on average across environments. Stochastic monotonicity ensures that surplus is never negative. Hazard-rate monotonicity implies in addition that quality gains grow with omission, making the hit-omit curve monotonic.

Proposition A8 concerns changes in the hit-rate differential fixed a success probability distribution. We can dually ask what happens to Γ_π if we fix payoffs and vary the composition of the test in terms of question difficulty. Intuitively, when the distribution becomes "more bimodal", metacognitive ability should earn larger gains from selection, generalizing the role of variance and spread in Equation 3. To formalize this, given two distributions π^A and π^B over success probabilities, say that π^B is a *centrifugal spread* of π^A, written $\pi^B \succeq_{CS} \pi^A$, if there exists $\beta > 1$ such that

$$\pi^B = l_\#(\pi^A) \quad \text{for} \quad l(p) = \mathbb{E}_\pi[p] + \beta(p - \mathbb{E}_\pi[p]). \tag{11}$$

The affine transformation l pushes mass away from the barycenter of the distribution, its mean, towards the extremes – much like the surface of a fluid in a spinning centrifuge. The parameter β captures the intensity of the dispersion. Since expectation is linear, centrifugal spreads preserve the mean.[19]

Proposition A9. *If* ψ *is stochastically monotone,*

$$\pi^B \succeq_{CS} \pi^B \Rightarrow \Gamma_{\pi^B} \geq \Gamma_{\pi^B}$$

C.2 A General Result on Complementarity

In the binary model, Proposition 5 reveals a complementarity between productivity and metacognitive bias. I now show that the same force is at play more generally. For arbitrary profiles, however, comparing distributions based on the relative strength of correlation at high versus low success probabilities is not straightforward, since notions such as true positive rates and true negative rates are not well-defined.

To overcome this difficulty observe that, in the binary model, when two students have the same metacognitive sensitivity, but one exhibits stronger bias, their confidence distributions

[19] In the context of common value auctions, Cusumano and Pieroth (2024) independently introduce the notion of *mean-preserving k-scaling*. A centrifugal spread is a mean preserving scaling where $k > 1$.

dominate the first students' ones at all probability levels. I generalize this property, asking on top that the difference between confidence distributions is increasing. Say that ψ^B is *more positively biased* than ψ^A if

$$\delta(c \mid p) \equiv \overline{F}_{\psi^B(p)}(c) - \overline{F}_{\psi^A(p)}(c) \geq 0 \quad \text{and increasing in } p. \tag{12}$$

where $\overline{F}_{\psi}(p)$ is the survival function of $\psi(p)$.
For a general scoring rule, denote by $f(p) = m + (h - m)p$ the expected score from attempting a question with success probability p. For a statistic (scores, omit-rates or hit-rates) denote by ∇ the difference between B's and A's value.

Proposition A10. *Consider two students A and B that follow the same uniform threshold rule $\hat{c}$ and have the same distribution of success probabilities π. Then*

$$\nabla W_N \xrightarrow{N\to\infty} \mathbb{E}\left[(f(p) - o)\ \delta(\hat{c} \mid p)\right]. \tag{13}$$

In particular, if ψ^B is more positively imbalanced than ψ^A, then

$$\mathbb{P}\left(\nabla W_N \geq -\left(\mathbb{E}_\pi\left[f(p)\right] - o\right)\nabla O_N\right) \xrightarrow{N\to\infty} 1, \tag{14}$$

where the omission gap ∇O_N is negative and FOSD-decreasing in π. Moreover, there exists a threshold $\tilde{\pi}$ such that for all $\pi' \succeq^{FOSD} \pi \succeq^{FOSD} \tilde{\pi}$,

$$\mathbb{P}\left(\nabla W_N(\pi') \geq \nabla W_N(\pi)\right) \xrightarrow{N\to\infty} 1.$$

Equation 13 compactly expresses how the impact on scores of a shift in metacognition, holding everything else constant, depends on the expectation of the product of two terms. The first, $f(p) - o$, captures the relative payoff from answering versus omitting a question as a function of its success probability p (as in Section **??**). The second, $\delta(\hat{c} \mid p)$, represents the change in the probability of answering a question with success probability p, induced by the metacognitive shift. Inequality 14 follows because, under condition 12, both of these terms are increasing functions of performance. As such, they are affiliated (Milgrom and Weber, 1982), which implies a positive covariance between them. This re-establishes the core intuition from Section 6: when students are sufficiently able – specifically, when $\mathbb{E}_\pi[p] \geq (o - m)/(h - m)$, just as in the binary case – the lower bound on the welfare gap increases under FOSD improvements in knowledge. This is the key force driving the complementarity result, ensuring monotonicity of the gap "eventually."
Another immediate implication of Equation 13 is that a degree of overconfidence can help compensate for an excessively conservative threshold rule. This nuance was obscured in the binary model, where only two confidence levels preclude sufficient nuance to capture the effect. To bring this aspect into focus, I characterize the structure of *maximal profiles*

– that is, those maximizing average test score among students sharing the same knowledge π and decision threshold $\hat{c}$.

Corollary 1. *Among students with knowledge distribution π and uniform threshold $\hat{c}$, the average score is maximized, in the limit, if and only if*

$$\begin{cases} p \leq p^* & \Rightarrow supp(\psi(p)) \subseteq [0, \hat{c}] \\ p > p^* & \Rightarrow supp(\psi(p)) \subseteq [\hat{c}, 1] \end{cases} \qquad where \quad p^* = \frac{o-m}{h-m}.$$

As a consequence, unless $\hat{c} = p^*$, perfect calibration $\psi^* = id$ is not score-maximizing.

To see how maximal types may involve overconfidence, consider the CARA model discussed in Appendix A, where risk aversion implies $\hat{c} > p^*$. In such cases, maximality requires overstating success probabilities "near the decision threshold", i.e. in the interval $[p^*, \hat{c}]$.[20]

[20] By contrast, under risk neutrality, perfect metacognition is always maximal, $\hat{c} = p^*$.

D Proofs

D.1 Main Results

Proof of Proposition 1 Fix a student i of type s and omit indices for brevity. First, I consider the case $\lambda = 1$. Let A^1 be an optimal portfolio. Assume by contradiction there exists an included question $j \in A^1$ and an excluded question $l \notin A^1$ such that $c_l > c_j$. Consider the alternative portfolio $A' = (A^1 \setminus \{j\}) \cup \{l\}$. Since $|A'| = |A^1|$, the cost term is identical: $k(|A'|) = k(|A^1|)$. Let $S = \sum_{k \in A^1 \setminus \{j\}} B_k$ be the random number of successes from the common questions, where B_k is a Bernoulli of parameter c_k. Since questions are independent, S is independent of B_j and B_l. Since $c_l > c_j$, the variable $B_l \succ_{FOSD} B_j$. Hence, in turn $A'^+ = S + B_l \succ_{FOSD} S + B_j = (A^1)^+$. We obtain contradiction with Preference Monotonicity.

Second, consider $A^0 = \{1, \ldots, N\}$. Consider any proper subset $A \subset A^0$. Let $l \in A^0 \setminus A$. Define $A' = A \cup \{l\}$. The score corresponds to $A'^+ = A^+ + B_l$. Since $c_l \in [0, 1]$, adding a non-negative random variable implies $(A')^+ \succeq_{\text{FOSD}} A^+$. By Preference Monotonicity, $U(A'^+) \geq U(A^+)$. Since costs are zero, the student (weakly) benefits from expanding the portfolio.

□

Proof of Proposition 2 *Part (i).* Fix a student i of type s and some $\varepsilon \in (0, 1/2)$. For convenience, suppress the index i. By Parallel Testing, for each test length N, the data generating process yields random vectors

$$\left(p_j^0, B_j^0, p_j^1, B_j^1\right)_{j=1}^N$$

such that $(p_j^0, p_j^1)_{j=1}^N$ are i.i.d. draws from a joint distribution whose marginals equal the mixture $\pi_s = \sum_{q \in \mathcal{Q}} w_q \pi_{s,q}$ and, for $\lambda \in \{0, 1\}$, $B_j^\lambda | p_j^\lambda \sim \text{Bernoulli}(p_j^\lambda)$. For brevity, define the event $E_N = \{O_N^1 \in [\varepsilon, 1 - \varepsilon]\}$. Consider the hit-rate differential. Note that we can rewrite

$$H_N^1 - H_N^0 = \underbrace{\left(\frac{1}{|A_N^1|} \sum_{j \in A_N^1} p_j^1 - \frac{1}{N} \sum_{j=1}^N p_j^0 \right)}_{\Delta} + \underbrace{\left(\frac{1}{|A_N^1|} \sum_{j \in A_N^1} (B_j^1 - p_j^1) - \frac{1}{N} \sum_{j=1}^N (B_j^0 - p_j^0) \right)}_{\nu}.$$

where Δ is a selection term and ν a Bernoulli noise term. I first work on the noise term, showing that, for any $\delta > 0$,

$$\lim_{N \to \infty} \mathbb{P}(E_N \cap \{|\nu| \geq \delta\}) = 0.$$

To see this, observe that ν is a difference of a regime-1 and a regime-0 term. For the regime-0 term it holds

$$\frac{1}{N}\sum_{j=1}^{N}(B_j^0 - p_j^0) \xrightarrow{a.s.} 0$$

by the Strong Law of Large Numbers, since the sequence $(B_j^0 - p_j^0)_{j=1}^{\infty}$ is bounded, independent, and mean zero (indeed, $\mathbb{E}[B_j^0] = \mathbb{E}[\mathbb{E}[B_j^0|p_j^0]] = \mathbb{E}[p_j^0]$). The regime-1 term needs care because A_N^1 is random. To start, rewrite it as

$$\frac{1}{|A_N^1|}\sum_{j\in A^1}(B_j^1 - p_j^1) = \frac{N}{|A_N^1|}\left(\frac{1}{N}\sum_{j=1}^{N}(B_j^1 - p_j^1)\mathbf{1}\{j \in A_N^1\}\right).$$

Note that on E_N, the fraction $N/|A_N^1| \leq 1/\varepsilon$, so this scalar is almost surely finite. To conclude, then, it is sufficient to show that the sum in parentheses converges to zero. To see this, let $\mathcal{G} = \sigma((p_j^1)_{j=1}^N)$ be the σ-algebra generated by the vector of success probabilities. Thanks to Rank-Preserving Confidence, the selection indicator $\mathbf{1}\{j \in A_N^1\}$ is $\mathcal{G}$-measurable because A_N^1 is determined by $(c_j)_{j=1}^N$, which is a deterministic transformation of $(p_j^1)_{j=1}^N$. Hence, by the law of iterated expectations, the summands are centered,

$$\begin{aligned}\mathbb{E}\left[(B_j^1 - p_j^1)\mathbf{1}\{j \in A_N^1\}\right] &= \mathbb{E}\left[\mathbb{E}\left[(B_j^1 - p_j^1)\mathbf{1}\{j \in A_N^1\} \mid \mathcal{G}\right]\right] \\ &= \mathbb{E}\left[\mathbf{1}\{j \in A_N^1\}\underbrace{\mathbb{E}\left[B_j^1 - p_j^1 \mid \mathcal{G}\right]}_{=0}\right] = 0.\end{aligned}$$

Furthermore, conditional on $\mathcal{G}$, the terms in the sum are independent and have variances bounded by 1/4. Therefore, by the Strong Law of Large Numbers, the sample average converges to zero almost surely.

Next, I work on the selection term, showing that there exists $\gamma > 0$ such that

$$\lim_{N\to\infty}\mathbb{P}\left(E_N \cap \{\Delta \geq \gamma\}\right) = 1.$$

Start with the regime-0 term. The Strong Law of Large Numbers directly implies that $\frac{1}{N}\sum_{j=1}^{N} p_j^0 \xrightarrow{a.s.} \mathbb{E}_{\pi_s}[p]$, by definition of the data generating process. Consider the regime-1 term. Combining Rank-Preserving Confidence with portfolio monotonicity we get that if the realized omit rate is α, the average success probability in A^1 is bounded below by the average of the top $\lfloor N\alpha \rfloor$ order statistics of the sequence $(p_j^1)_{j=1}^N$. Now, by the Glivenko-Cantelli, this quantity converges uniformly to the expected value of p conditional on being in the top α-quantile of the distribution F_{π_s}. Define $g(\alpha) = \mathbb{E}_{\pi_s}\left[p \mid p \geq F_{\pi_s}^{-1}(1-\alpha)\right]$. For

any $\eta > 0$, we have

$$\lim_{N\to\infty} \mathbb{P}\left(\sup_{\alpha\in[\varepsilon,1-\varepsilon]} \left| \frac{1}{\lfloor N\alpha \rfloor} \sum_{k=N-\lfloor N\alpha\rfloor+1}^{N} p^1_{(k)} - g(\alpha) \right| < \eta \right) = 1,$$

where $p^1_{(k)}$ is the k-th biggest element of $(p^1_j)_{j=1}^N$. Since the distribution of success probability is non-degenerate and by Rank-Preserving Confidence the confidence map is non-constant we have $g(\alpha) - \mathbb{E}_{\pi_s}[p] > 0$ for all $\alpha \in [\varepsilon, 1-\varepsilon]$. Since the domain is compact and $g(\alpha)$ is continuous,[21]

$$\min_{\alpha\in[\varepsilon,1-\varepsilon]} (g(\alpha) - \mathbb{E}_{\pi_s}[p]) > 0.$$

Pick γ equal to this value. The statement follows since on the event E_N, the random selection rate $\alpha_N = |A^1|/N$ is almost surely bounded into $[\varepsilon, 1-\varepsilon]$.
To conclude the proof, fix any arbitrarily small $\eta > 0$. Fix γ as in the second step. Set $\delta = \gamma/2$. From the first step there exists N_1 such that for all $N \geq N_1$

$$\mathbb{P}(E_N \cap \{|\nu| \geq \gamma/2\}) < \eta/2.$$

From the second step, there exists N_2 such that for all $N \geq N_2$:

$$\mathbb{P}(E_N \cap \{\Delta < \gamma\}) < \eta/2.$$

Recall that $H^N_1 - H^0_N = \Delta$. Applying the union bound on the intersection with E_N:

$$\begin{aligned}\mathbb{P}(E_N \cap \{H^1_N - H^0_N\}) &\leq \mathbb{P}(E_N \cap \{\Delta < \gamma\}) + \mathbb{P}(E_N \cap \{|\nu| \geq \gamma/2\}) \\ &< \eta/2 + \eta/2 = \eta.\end{aligned}$$

Since this holds for any $\eta > 0$ for sufficiently large N, it holds

$$\lim_{N\to\infty} \mathbb{P}\left(\{O^1_N \in [\varepsilon, 1-\varepsilon]\} \cap \{H^1_N - H^0_N \leq 0\}\right) = 0.$$

Part (ii). Fix $N \geq 2$ and $\varepsilon \in (0, 1/2)$. We prove the statement conditional on a vector of question types $\mathbf{q} = (q_j)_{j=1}^N$, which implies the result. Define $D_i = \mathbf{1}\{|A^1_i| \geq 1\}$, $\tilde{O}^1_i = D_i O^1_{i,N}$ and $Z_i = D_i(H^1_{i,N} - H^0_{i,N})$ (with the convention that $0 \cdot NA = 0$). By Parallel Testing, for any M the data generating process yields random vectors

$$(\tilde{O}^1_i, D_i, Z_i)_{i=1}^M$$

which are bounded and i.i.d. with a law determined by the mixture $\pi_{\mathbf{q}} = \sum_{s\in\mathcal{S}} w_s \pi(\cdot \mid s, q)$. In particular, population expectations are convex combinations of type subpopulation

[21] Recall: $g(\alpha) = \frac{1}{\alpha}\inf_{1-\alpha}^1 F^{-1}_{\pi_s}(u)du$.

expectations. Hence, define

$$\omega = \mathbb{E}_{\pi_{\mathbf{q}}}[\tilde{O}_i^1] = \sum_{s \in \mathcal{S}} w_s \underbrace{\mathbb{E}_{\pi_{\mathbf{q}}}[\tilde{O}_i^1 \mid s(i)]}_{\omega_s} \tag{15}$$

$$\delta = \mathbb{E}_{\pi_{\mathbf{q}}}[D_i] = \sum_{s \in \mathcal{S}} w_s \underbrace{\mathbb{E}_{\pi_{\mathbf{q}}}[D_i \mid s(i)]}_{\delta_s} \tag{16}$$

$$\zeta = \mathbb{E}_{\pi_{\mathbf{q}}}[Z_i] = \sum_{s \in \mathcal{S}} w_s \underbrace{\mathbb{E}_{\pi_{\mathbf{q}}}[Z_i \mid s(i)]}_{\zeta_s} \tag{17}$$

Now, by the Strong Law of Large Numbers,

$$\frac{1}{M}\sum_{i=1}^{M} \tilde{O}_i^1 \xrightarrow{a.s.} \omega, \quad \frac{1}{M}\sum_{i=1}^{M} D_i \xrightarrow{a.s.} \delta, \quad \frac{1}{M}\sum_{i=1}^{M} Z_i \xrightarrow{a.s.} \zeta.$$

Define our event of interest, $E_M = \{O_{M,N}^1 \in [\varepsilon, 1-\varepsilon]\} \cap \{H_{M,N}^1 - H_{M,N}^0 \leq 0\}$. We show it has zero probability proceeding by branching cases.

First, suppose $\delta = 0$. This can be the case if and only if almost surely, for each i, $D_i = 0$. In turn, $\mathbb{P}(E_M) \leq \mathbb{P}(\{O_{M,N}^1 \in [\varepsilon, 1-\varepsilon]\}) \to 0$. Hence, assume $\delta \in (0, 1]$. In this case, by the Continuous Mapping Theorem,

$$O_{M,N}^1 \xrightarrow{a.s.} \frac{\omega}{\delta}.$$

Hence, if $\omega \notin [\delta\varepsilon, \delta(1-\varepsilon)]$, we have $\mathbb{P}(E_M) \to 0$.

We can then restrict to $\omega \in [\delta\varepsilon, \delta(1-\varepsilon)]$ for $\delta \in (0, 1]$. In particular, $\omega > 0$. Hence, by Equation (15), there exists a student-type s^* such that $\omega_{s^*} > 0$. Hence, for any student i of type s^*, it must be

$$\mathbb{P}(\{1 \leq |A_i^1| < N\}) > 0.$$

Define $S = \{1 \leq |A_i^1| < N\}$. By the law of total expectations, for any type s

$$\zeta_s = \mathbb{P}(S)\mathbb{E}[Z_i \mid S] + \mathbb{P}(S^C)\mathbb{E}[Z_i \mid S^C].$$

Observe that on the complement set S^C, either $|A_i^1| = 0$ (and $D_i = 0$) or $|A_i^1| = N$ (implies $\mathbb{E}[H_i^1 \mid |A_i^1| = N] = \mathbb{E}[H_i^0 \mid |A_i^1| = N]$). Hence $\mathbb{E}[Z_i \mid S^C] = 0$, so the second term vanishes. For the first term, since under Rank-Preserving Confidence and Preference Monotonicity the portfolio is monotone in success probabilities, we have $\mathbb{E}[Z_i|S] \geq 0$, and thus $\zeta_s \geq 0$ for all s. Moreover, for type s^*, the inequality is strict by the variation ensured by non-degeneracy of the success probability distribution and the non-constancy of the confidence map (Rank-Preserving Confidence). Since $\mathbb{P}(\{1 \leq |A_i^1| < N\}) > 0$ for this type, we have $\zeta_{s^*} > 0$. Altogether, we get by Equation (17) that $\zeta > 0$. To conclude,

observe that by the Continuous Mapping Theorem we have

$$H^1_{M,N} - H^0_{M,N} \xrightarrow{a.s.} \frac{\zeta}{\delta} > 0,$$

and hence we have that $\mathbb{P}(E_M) \to 0$.

□

Proof of Proposition 3 The proof is immediate. Note that under Preference Monotonicity and Rank-Preserving Confidence the optimal portfolio ins monotone in success probabilities, i.e.

$$j \in A^1_i \text{ and } k \in (A^1_i)^C \Rightarrow p_{ij} \geq p_{ik}$$

Hence, under Strictly Ordered Supports,

$$j \in A^1_i \text{ and } k \in (A^1_i)^C \Rightarrow q_{ij} > q_{ik}$$

this implies the statement.

□

Proof of Proposition 4 To prove the result first observe that, when the student follows a threshold rule $\hat{c}_i$, it holds

$$\begin{cases} O^1_N & \xrightarrow{N\to\infty} \mathbb{P}(\{c \leq \hat{c}_i\}), \\ H^0_N & \xrightarrow{N\to\infty} \mathbb{E}[p] \\ H^1_N & \xrightarrow{N\to\infty} \mathbb{E}[p \mid c > \hat{c}_i]. \end{cases}$$

Indeed, $(p_i, c_i)_{i=1}^N$ are i.i.d. draws from the joint same joint distribution determined, by (π, ψ). By the Strong Law of Large Numbers, we have

$$O^1_N = \frac{1}{N}\sum_{j=1}^N \mathbf{1}\{c_j \leq \hat{c}_i\} \xrightarrow{N\to\infty} \mathbb{P}(\{c \leq \hat{c}_i\})$$
$$E^1_N \equiv \frac{1}{N}\sum_{j=1}^N p_j \mathbf{1}\{c_j > \hat{c}_i\} \xrightarrow{N\to\infty} \mathbb{E}[p\mathbf{1}\{c > \hat{c}_i\}].$$

This shows the point for the omit-rate. Moreover, combining the two equations via the continuous mapping theorem, we obtain

$$D^1_N \equiv \frac{\sum_{j=1}^N p_j \mathbf{1}\{c_j > \hat{c}_i\}}{\sum_{j=1}^N \mathbf{1}\{c_j > \hat{c}_i\}} = \frac{E^1_N}{1 - O^1_N} \xrightarrow{N\to\infty} \mathbb{E}[p \mid c > \hat{c}_i].$$

Now, note that, applying again the SSLN

$$H_N^1 - D_N^1 = \underbrace{\frac{N}{\sum_{j=1}^N \mathbf{1}\{c_j > \hat{c}_i\}}}_{\xrightarrow{N\to\infty}\mathbb{P}_\tau(\{c>\hat{c}_i\})^{-1}} \underbrace{\frac{1}{N}\sum_{j=1}^N (Z(p_j) - p_j)\mathbf{1}\{c_j > \hat{c}_i\}}_{\xrightarrow{N\to\infty}0} \xrightarrow{N\to\infty} 0.$$

Hence, $H_N^1 \xrightarrow{N\to\infty} \mathbb{E}_\tau[p \mid c > \hat{c}_i]$. Similarly it follows $H_N^0 \xrightarrow{N\to\infty} \mathbb{E}[p]$. Given this observation, the proof follows by the Continuous Mapping Theorem, since, after rearranging

$$H_N^1 - H_N^0 \xrightarrow{N\to\infty} \mathbb{E}[p|c = \overline{c}] - \mathbb{E}[p] = (\overline{p} - \underline{p})(1-\pi)\pi\left[\frac{\psi^\Sigma}{1 - \mathbb{P}_\tau(\{\underline{c}\})}\right].$$

And hence

$$(1 - O_N^1)H_N^1 - H_N^0 \xrightarrow{N\to\infty} \mathbb{E}[p|c = \overline{c}] - \mathbb{E}[p] = (\overline{p} - \underline{p})(1-\pi)\pi\psi^\Sigma.$$

□

Proof of Proposition 5 Observe that we can write

$$W_N^1 = oO_N^1 + h(1 - O_N^1)H_N^1$$

hence, using the limit expressions for O_N^1 and H_N^1 we obtain

$$W_N^1 \xrightarrow{N\to\infty} o(\pi(1-\psi^+) + (1-\pi)\psi^-) + h(\pi\psi^+\overline{p} + (1-\pi)(1-\psi^-)\underline{p})$$

that can be rearranged to expression 4.

□

D.2 Appendix Results

Proof of Proposition A6 The result follows from the construction of success probabilities and confidence levels, via law of conditional expectations. In particular, consider the joint probability measure defined in Equation 9. By Equation 10 we have

$$c(\theta, s_\omega, s_\theta) = \mathbb{P}(E \mid s_\omega, s_\theta) = \mathbb{E}[\mathbf{1}\{E\} \mid s_\omega, s_\theta]$$
$$p(\theta, s_\omega, s_\theta) = \mathbb{P}(E \mid \theta, s_\omega, s_\theta) = \mathbb{E}[\mathbf{1}\{E\} \mid \theta, s_\omega, s_\theta]$$

Hence, by the tower property we have

$$\mathbb{E}[p \mid c] = \mathbb{E}[\mathbb{E}[p \mid s_\omega, s_\theta] \mid c]$$

To conclude it is then sufficient to show $\mathbb{E}[p \mid s_\omega, s_\theta] = c$. But this is the case since

$$\mathbb{E}[p \mid s_\omega, s_\theta] = \mathbb{E}[\mathbb{E}[\mathbf{1}\{E\} \mid \theta, s_\omega, s_\theta] \mid s_\omega, s_\theta] = \mathbb{E}[\mathbf{1}\{E\} \mid s_\omega, s_\theta]$$

where the first equality follows substituting the expression for the success probability and the second by the tower law.

□

Proof of Proposition A7 To show the result, I first introduce the following simple fact, which relies on the independence of ω and θ (Equation 9).

Lemma A1. *For any information set (s_ω, s_θ), the Bayesian posterior about ω can be achieved updating beliefs through the following hierarchical procedure:*

1. *Bayes-update the prior beliefs t_0 on θ, based on the meta-experiment μ and the signal s_θ to obtain interim beliefs t_{s_θ}.*
2. *Bayes-update the prior beliefs d_0 on ω, based on the expected experiment $\mathbb{E}_{t_{s_\theta}}[\sigma_\theta] : \Omega \to \Delta(S_\Omega)$ and the signal s_ω.*

Proof of the Lemma Observe that

$$\begin{aligned}\mathbb{P}(\omega|s_\omega, s_\theta) &= \frac{d_0(\omega)\sum_{\theta\in\Theta} t_0(\theta)\mu(s_\theta|\theta)\sigma(s_\omega|\omega,\theta)}{\sum_{\omega'\in\Omega} d_0(\omega')\sum_{\theta\in\Theta} t_0(\theta)\mu(s_\theta|\theta)\sigma(s_\omega|\omega',\theta)} \\ &= \frac{d_0(\omega)\sum_{\theta\in\Theta} \frac{t_0(\theta)\mu(s_\theta|\theta)}{\sum_{\theta'\in\Theta}\mu(s_\theta|\theta')t_0(\theta')}\sigma(s_\omega|\omega,\theta)}{\sum_{\omega'\in\Omega} d_0(\omega')\sum_{\theta\in\Theta} \frac{t_0(\theta)\mu(s_\theta|\theta)}{\sum_{\theta'\in\Theta}\mu(s_\theta|\theta')t_0(\theta')}\sigma(s_\omega|\omega',\theta)}\end{aligned}$$

Observe that $t_{s_\theta}(\theta) = \frac{t_0(\theta)\mu(s_\theta|\theta)}{\sum_{\theta'\in\Theta}\mu(s_\theta|\theta')t_0(\theta')}$ is the belief about informativeness type θ obtained updating t_0 after s_θ through the experiment μ. Hence, we have

$$\mathbb{P}(\omega|s_\omega, s_\theta) = \frac{d_0(\omega)\mathbb{E}_{t_{s_\theta}}[\sigma_\theta](s_\omega|\omega)}{\sum_{\omega'\in\Omega} d_0(\omega')\mathbb{E}_{t_{s_\theta}}[\sigma_\theta](s_\omega|\omega')}$$

which shows the point.

The second fact needed to show the result collects two properties of symmetric diagonally dominant experiments.

Lemma A2. *Let $\sigma : \Omega \times \Theta \to \Delta(S_\Omega)$ be a family of symmetric experiments. The two following facts hold.*

1. *Given a probability measure $t \in \Delta(\Theta)$, the expected experiment $\mathbb{E}_t[\sigma_\theta]$ is itself symmetric, with parameter $\mathbb{E}_t[\theta]$.*

2. *Let $d_0 \in \Delta(\Omega)$ be the uniform prior. Consider a symmetric σ_θ and a signal s_ω. Let d_{s_ω} the associated posterior. It holds $mode(d_{s_\omega}) = s_\omega$ and $d_{s_\omega}(mode(d_{s_\omega})) = \theta$*

Proof of the Lemma To see the first point, observe that the expected experiment satisfies

$$\mathbb{E}_t[\sigma_\theta](\omega'|\omega) = \sum_{\theta \in \Theta} t(\theta)\sigma_\theta(\omega'|\omega).$$

Hence substituting $\sigma_\theta(\omega \mid \omega) = \theta$ and $\sigma_\theta(\omega' \mid \omega) = \frac{1-\theta}{|\Omega|-1}$ we obtain that $\mathbb{E}_t[\sigma_\theta](\omega|\omega) = \mathbb{E}_t[\theta]$ and $\mathbb{E}_t[\sigma_\theta](\omega'|\omega) = \frac{1-\mathbb{E}_t[\theta]}{|\Omega|-1}$. To see the second point, observe that, given that the prior is uniform $d_{s_\omega}(\omega) = \sigma(\omega|s_\omega)$. This is maximized by $\omega = s_\omega$, since $\theta \geq 1/|\Omega|$. By the same token, we see that the mass on the mode is θ.

The main result follows combining the two Lemmata. First note that we have $p(\theta, s_\omega, s_\theta) = d^\theta_{s_\omega}(s_\omega) = \theta$ and $c(\theta, s_\omega, s_\theta) = d^{s_\theta}_{s_\omega}(s_\omega) = \mathbb{E}_{t_{s_\theta}}[\theta]$. Suppose that (π, ψ) is a martingale coupling such that $supp\pi \subset [1/|\Omega|, 1]$. Then, take π as a distribution over symmetric experiments $\sigma : \Omega \times P \to \Delta(S_\Omega)$. Since $p(\theta, s_\omega, s_\theta) = \theta$, we get that the distribution of success probabilities induced by the information processing model is precisely π. As a meta-experiment, take $\mu \equiv \psi$. To show that this indeed works, observe that updating prior π by experiment ψ after a signal/confidence level c we have a posterior mean

$$\mathbb{E}_{\pi_c}[p] = \frac{\sum_{p \in P} p\pi(p)\psi(c|p)}{\sum_{p \in P} \pi(p)\psi(c|p)} = \mathbb{E}_{(\pi,\psi)}[p|c] = c$$

where the last equality follows since (π, ψ), by hypothesis, is a martingale coupling.

□

Proof of Proposition A8 To prove the result, I introduce some auxiliary Lemmata

Lemma A3. *Let $\pi \in \Delta([0,1])$ and $f, g : [0,1] \to \mathbb{R}$. Then,*

$$Cov[f(p), g(c)] = Cov[f(p), \mathbb{E}_{\psi(p)}[g(c)]]$$

Proof. This follows from the law of total covariance

$$Cov(x, y) = \mathbb{E}[Cov(x, y|z)] + Cov[\mathbb{E}[x|z], \mathbb{E}[y|z]]$$

choosing $x = f(p)$, $y = g(c)$ and $z = p$, since $f(p)$ is p-measurable and hence (a) the first term is zero and (b) the first conditional expectation in the second term is deterministic.

Lemma A4 (Chebyshev's Association Inequality)**.** *Let $\pi \in \Delta([0,1])$ and $f, g : [0,1] \to \mathbb{R}$*

non-decreasing functions. Then,

$$Cov_\pi[f(p), g(p)] \geq 0$$

Proof To see the fact, observe that $Cov[f(p), g(p)] = \frac{1}{2}\mathbb{E}[(f(p) - f(p'))(g(p) - g(p'))]$ where p and p' are two independent copies of the same variable. If both f and g are non-decreasing, the two terms inside the expectation have the same sign, state-wise. Hence, the expectation is nonnegative.

Lemma A5. *Let $h : [0,1] \to \mathbb{R}$. The following are equivalent.*

1. *h is non-decreasing;*
2. *For all $\pi \in \Delta([0,1])$, $Cov_\pi[p, h(p)] \geq 0$.*

Proof 1 implies 2 follows from Lemma A4. To see that 2 implies 1 suppose h is not non-decreasing. Then, there exists $p_1, p_2 \in [0,1]$ such that $p_2 > p_1$ but $h(p_1) > h(p_2)$. Consider the distribution $\pi = \frac{1}{2}\delta(p_1) + \frac{1}{2}\delta(p_2)$ (where δ is Dirac's delta). Then, $Cov_\pi[P, h(p)] = \frac{1}{4}(p_2 - p_1)(h(p_2) - h(p_1)) < 0$.

Point (i). From left to right, use Lemma A3 to rewrite $Cov[p, c]$ as $Cov[p, \mathbb{E}_{\psi(p)}[c]]$. Observe that expectation monotonicity ensures that $\mathbb{E}_{\psi(p)}[c]$ is a non-decreasing function of p. Hence, we have expressed $Cov[p, c]$ as the covariance of two non-decreasing transformations of the same random variable. By Lemma A4, this is a non-negative quantity.
From right to left, use again Lemma A3 to rewrite $Cov[p, c]$ as $Cov[p, \mathbb{E}_{\psi(p)}[c]]$. By Lemma A5, 2 implies that $\mathbb{E}_{\psi(p)}[c]$ is a non-decreasing function. But this is precisely 1.

Point (ii). Observe that it holds

$$Cov[p, \mathbf{1}\{c > k\}] = (1 - \mathbb{P}(\{c < k\}))(\mathbb{E}[p \mid c \geq k] - \mathbb{E}[p])$$

Hence, from left to right, it is sufficient to prove that the left hand side implies

$$\text{for all } \pi \in \Delta([0,1]) \text{ and all } f, g : [0,1] \to [0,1] \text{ nondecreasing } Cov[f(p), g(c)] \geq 0.$$

In turn, this imply the right-hand side choosing $f = id$ and $g = \mathbf{1}\{\bullet > k\}$, for any $k \in (0,1)$. Hence, fix two non-decreasing functions f, g and a distribution π for P. Observe that $\mathbb{E}_{\psi(p)}[g(c)]$ is an increasing function of p (by the definition of First Order Stochastic Dominance). Then, since f is non-decreasing, we can use the identity in Lemma A3 to expresses $Cov[f(p), g(c)]$ as the covariance of two non-decreasing transformations of the same random variable. By Lemma A4, this is a non-negative quantity.
From right to left, note that by Lemma A3, $Cov[p, \mathbf{1}\{c \geq k\}] = Cov[p, \mathbb{E}_{\psi(p)}[\mathbf{1}\{c \geq k\}]]$.

Then, applying Lemma A5, we obtain that 3 implies that, for every $k \in (0,1)$, $\mathbb{E}_{\psi(p)}[\mathbb{1}\{c \geq k\}]$ is non-decreasing in p. But since $\mathbb{E}_{\psi(p)}[\mathbb{1}\{c \geq k\}]$ is just the survival function of $\psi(p)$ evaluated at k, this implies 1.

Point (iii). From left to right, fix any π and $k_2 > k_1$. Denote by π_k the law of P conditional on $c > k$. Define the hazard rate $\rho_{k_1,k_2}(p) = \frac{\overline{F}_\psi(k_2|p)}{\overline{F}_\psi(k_1|p)}$. By hypothesis, ρ_{k_1,k_2} is increasing in p. Hence, by Lemma A4, we have

$$Cov_{\pi_{k_1}}[p, \rho_{k_1,k_2}(p)] \geq 0 \quad \text{or, equivalently,} \quad \frac{\mathbb{E}_{\pi_{k_1}}[p\rho_{k_1,k_2}(p)]}{\mathbb{E}_{\pi_{k_1}}[\rho_{k_1,k_2}(p)]} \geq \mathbb{E}_{\pi_{k_1}}[p].$$

To conclude, observe that

$$\mathbb{E}_{\pi_{k_2}}[p] = \frac{\int_0^1 p\overline{F}_\psi(k_2 \mid p)d\pi(p)}{\int_0^1 \overline{F}_\psi(k_2 \mid p)d\pi(p)} = \frac{\int_0^1 p\rho_{k_1,k_2}(p)\overline{F}_\psi(k_1 \mid p)d\pi(p)}{\int_0^1 \rho_{k_1,k_2}(p)\overline{F}_\psi(k_1 \mid p)d\pi(p)} = \frac{\mathbb{E}_{\pi_{k_1}}[p\rho_{k_1,k_2}(p)]}{\mathbb{E}_{\pi_{k_1}}[\rho_{k_1,k_2}(p)]}.$$

From right to left, let $\pi \equiv \pi\delta(p_1) + (1-\pi)\delta(p_2)$ for $p_1 < p_2$ and $\pi \in (0,1)$. By hypothesis,

$$\mathbb{E}[p \mid c > k] = \frac{\pi\overline{F}_\psi(k|p_1)p_1 + (1-\pi)\overline{F}_\psi(k|p_2)p_2}{\pi\overline{F}_\psi(k|p_1) + (1-\pi)\overline{F}_\psi(k|p_2)} = \frac{\pi p_1 + (1-\pi)\sigma_{p_1,p_2}(k)p_2}{\pi + (1-\pi)\sigma_{p_1,p_2}(k)},$$

where $\sigma_{p_1,p_2}(k) = \frac{\overline{F}_\psi(k|p_2)}{\overline{F}_\psi(k|p_1)}$, is non-decreasing in k. In turn, since the function

$$\sigma \mapsto \frac{\pi p_1 + (1-\pi)\sigma p_2}{\pi + (1-\pi)\sigma}$$

is increasing in σ, then $\sigma_{p_1,p_2}(k)$ must be non-decreasing in k. Since $p_1 < p_2$ are chosen arbitrarily, this shows that

$$\forall p_1 < p_2,\ k_1 < k_2 \quad \overline{F}_\psi(k_1|p_2)\overline{F}_\psi(k_2|p_1) \leq \overline{F}_\psi(k_1|p_1)\overline{F}_\psi(k_2|p_2)$$

or equivalently ψ is hazard-rate monotone.

□

Proof of Proposition A9 Since $\pi^B \succeq_{CS} \pi^A$ implies $\mathbb{E}_{\pi^B}[p] = \mathbb{E}_{\pi^A}[p]$, it is sufficient to show that for any k $\mathbb{E}_{(\pi^B,\psi)}[p \mid c > k] \geq \mathbb{E}_{(\pi^A,\psi)}[p \mid c > k]$. Observe that by equivalence, for any k and any increasing bijection f is holds $\overline{F}_{\psi^A(p)}(k) = \overline{F}_{\psi^B(f(p))}(k)$. Hence, choosing $f = l$, $\mathbb{P}_{(\pi^A,\psi)}(\{c \leq k\}) = 1 - \mathbb{E}_{\pi^A}[\overline{F}_{\psi^A(p)}(k)] = 1 - \mathbb{E}_{\pi^A}[\overline{F}_{\psi^B(l(p))}(k)] = \mathbb{P}_{(\pi^B,\psi)}(\{c \leq k\})$ where the first equality follows since $\overline{F}_{\psi(p)}(k)$ is the probability of performing a task at performance level p; the second by equivalence and the last by the definition of pushforward.

If $\mathbb{P}_\tau(\{c \leq k\}) \in \{0,1\}$, the claim follows trivially. Consider $\mathbb{P}_\tau(\{c \leq k\}) \in (0,1)$. It holds

$$H_N^B \xrightarrow{N\to\infty} \frac{\mathbb{E}_{\pi^B}[p\overline{F}_{\psi^B(p)}(k)]}{\mathbb{E}_{\pi^B}[\overline{F}_{\psi^B(p)}(k)]} = \frac{\mathbb{E}_{\pi^A}[l(p)\overline{F}_{\psi^B(l(p))}(k)]}{\mathbb{E}_{\pi^A}[\overline{F}_{\psi^B(l(p))}(k)]} = \frac{\mathbb{E}_{\pi^A}[l(p)\overline{F}_{\psi^A(p)}(k)]}{\mathbb{E}_{\pi^A}[\overline{F}_{\psi^A(p)}(k)]}$$

where the first equality follows by definition of pushforward; the second ψ by equivalence so that $\overline{F}_{\psi^A(p)} = \overline{F}_{\psi^B(l(p))}$. Then,

$$H_N^B - H_N^A \xrightarrow{N\to\infty} \frac{\mathbb{E}_{\pi^A}[(l(p)-p)\overline{F}_{\psi(p)}(k)]}{\mathbb{E}_{\pi^A}[\overline{F}_{\psi^A(p)}(k)]} = (\beta-1)\frac{\mathbb{E}_{\pi^A}[(p-\mathbb{E}_{\pi^A}[p])\overline{F}_{\psi^A(p)}(k)]}{\mathbb{E}_{\pi^A}[\overline{F}_{\psi^A(p)}(k)]}$$

Observe that the difference has the same sign of the fraction's numerator, as the denominator is positive and $\beta > 1$. To see the numerator is nonnegative, observe that since ψ^A is positive regression dependent, $\overline{F}_{\psi^A(p))}(k)$ is an increasing transformation of the random variable p, just as $(p - \mathbb{E}_\pi[p])$. Then,

$$\begin{aligned}\mathbb{E}_{\pi^A}[(p-\mathbb{E}_{\pi^A}[p])\overline{F}_{\psi^A(p)}(k)] - \mathbb{E}_{\pi^A}[(p-\mathbb{E}_{\pi^A}[p])]\mathbb{E}_{\pi^A}[\overline{F}_{\psi^A(p)}(k)] \\ = Cov((p-\mathbb{E}_\pi^A[p]), \overline{F}_{\psi^A(p))}(k)) \geq 0\end{aligned}$$

And the claim follows because $\mathbb{E}_{\pi^A}[(p - \mathbb{E}_{\pi^A}[p])] = 0$.

□

Proof of Proposition A10 Let $r^h = h - m$, $r^o = o - m$ and observe that $f_{\rho^1}(p) - o = pr^h - r^o$. Equation 13 from the expression for the total score, observing that $O_N \xrightarrow{N\to\infty} 1 - \mathbb{E}_\pi[\overline{F}_{\psi(p)}]$ and $(1-O_N)H_N \xrightarrow{N\to\infty} \mathbb{E}_\pi[p\overline{F}_{\psi(p)}]$. To obtain 14 from 13, note that $\delta(\hat{c}|P)$ is an increasing transformation of P, since ψ' is a positive shift of ψ. Hence, since also $(pr^h - r^o)$ is increasing we have

$$\mathbb{E}_\pi[(pr^h - r^o)\delta(\hat{c}|p)] - (r^h\mathbb{E}_\pi[p] - r^o)\underbrace{\mathbb{E}_\pi[\delta(\hat{c}|p)]}_{\overline{O}-\overline{O}'} = Cov(pr^h - r^o, \delta(\hat{c}|p)) \geq 0$$

Now, by definition of positive shift $\overline{O} - \overline{O}' \geq 0$ (since $\delta(\hat{c}|p) \geq 0$) and FOSD-increasing in π (since $\delta(\hat{c}|p)$ is increasing in p). For the last point consider the Dirac's delta $\tilde{\pi} = \delta_{\frac{r^o}{r^h}}$. Observe that $\pi \succeq_{FOSD} \tilde{\pi}$ implies $supp(\pi) \subseteq [\frac{r^o}{r^h}, 1]$. The conclusion follows because, both $\delta(\hat{c}|p)$ and $pr^h - r^o$ are positive and increasing in this interval. Hence their product is (positive) and increasing.

□

Proof of Corollary 1 Fix ψ as in the statement and any ψ'. From 13 it follows that

$$\delta\overline{W}(\psi,\psi'|\pi,\alpha) = \mathbb{E}_\pi\left[(pr^h - r^o)(\mathbf{1}_{p>p^*} - \overline{F}_{\psi'(p)})\right]$$

Observe that $sign(pr^h - r^o) = -sign(\mathbf{1}_{p>p^*} - \overline{F}_{\psi'(p)})$. Hence, the argument of the expectation is non-positive. Hence the result.

□

E Figures and Tables

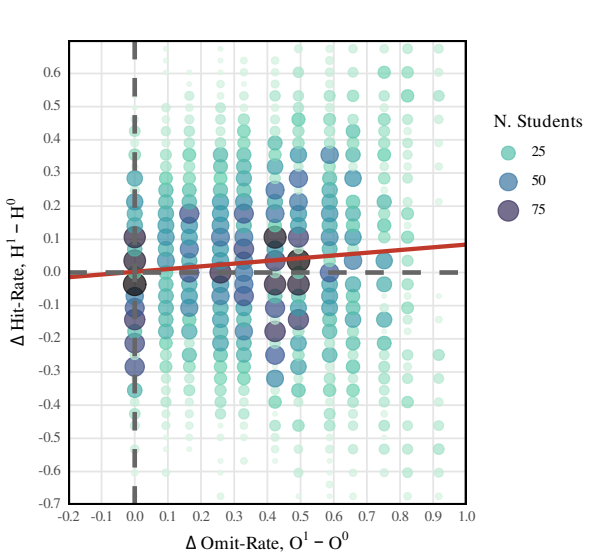
Table A1: Treatment Effects

	Omit-Rate	Omit-Rate	Hit-Rate	Hit-Rate
Penalty	0.347***	0.347***	0.026***	0.026
	(0.003)	(0.028)	(0.003)	(0.032)
Observations	19,614	19,614	19,614	19,614
R^2	0.770	0.770	0.697	0.697
Clustering	Student	Student + Sit.	Student	Student + Sit.

* p <0.1, ** p <0.05, *** p <0.01. Student-sitting FE.

Figure A1: Treatment Effects of Penalties for Errors (Preference Monotonicity)

(a) Individual Treatment Effects



(b) Average Treatment Effects

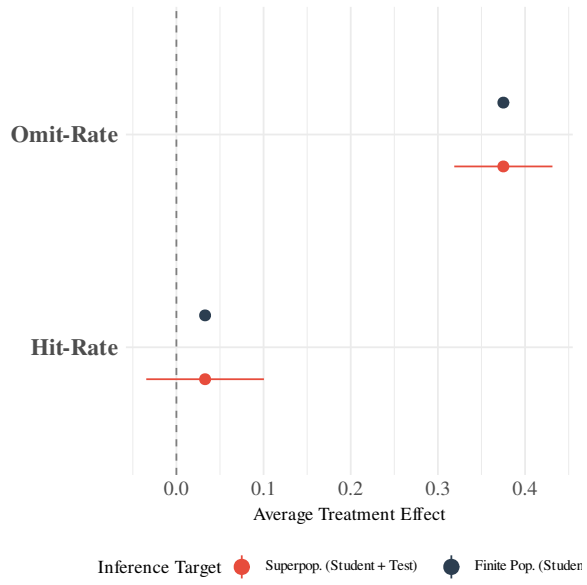


Figure A2: Treatment Effects Heterogeneity across Sittings

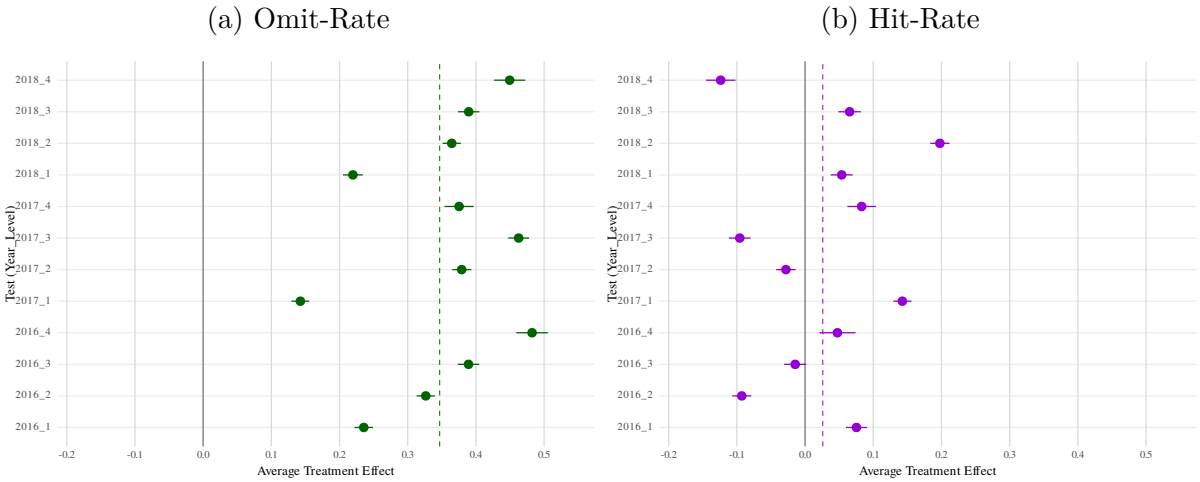


Table A2: Graders' Statistics

Sitting	Graders	Mean H_0 (G)	Non-Graders	Mean H_0 (NG)	H_0 Difference
2016_1	88	0.544	458	0.510	0.03363
2016_2	93	0.548	734	0.535	0.01317
2016_3	71	0.545	644	0.472	0.07279
2016_4	35	0.473	347	0.425	0.04712
2017_1	156	0.467	446	0.362	0.10477
2017_2	65	0.438	746	0.443	-0.00470
2017_3	48	0.476	722	0.443	0.03254
2017_4	51	0.367	381	0.299	0.06730
2018_1	78	0.515	461	0.486	0.02889
2018_2	63	0.447	790	0.415	0.03179
2018_3	90	0.501	655	0.493	0.00819
2018_4	65	0.472	386	0.460	0.01264
POOLED	903	0.488	6770	0.451	0.03697

Figure A3: Item Difficulty Estimates

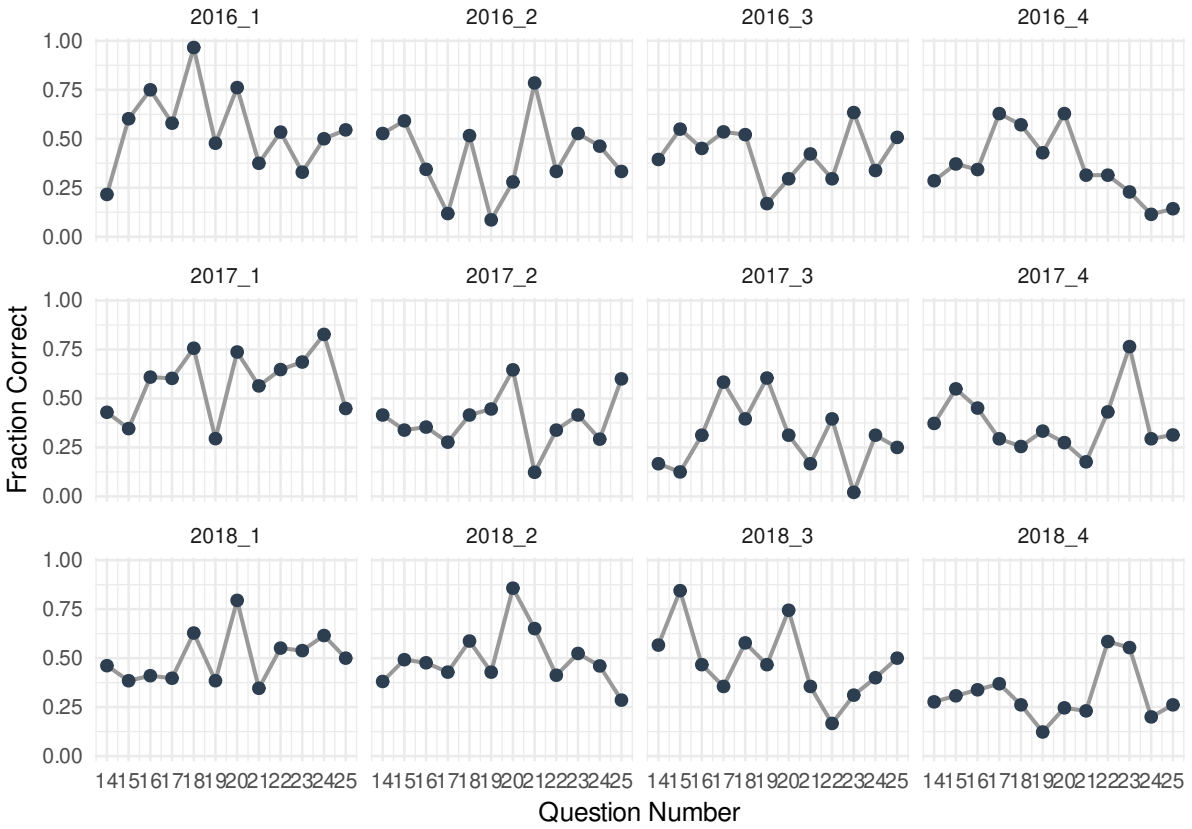


Figure A4: Population distribution of AUC (Preference Monotonicity)

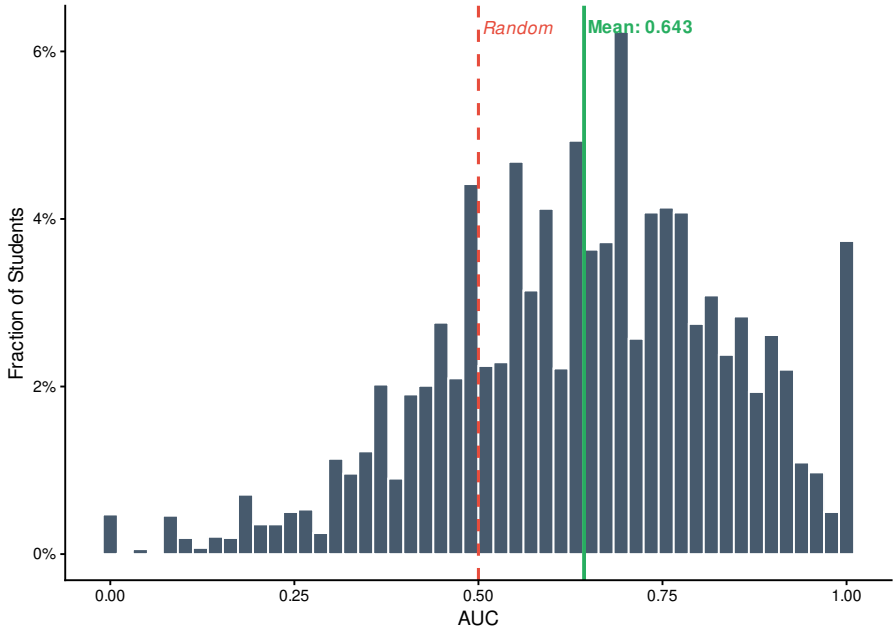


Figure A5: Marginal Distribution of the Hit-Rate Differential

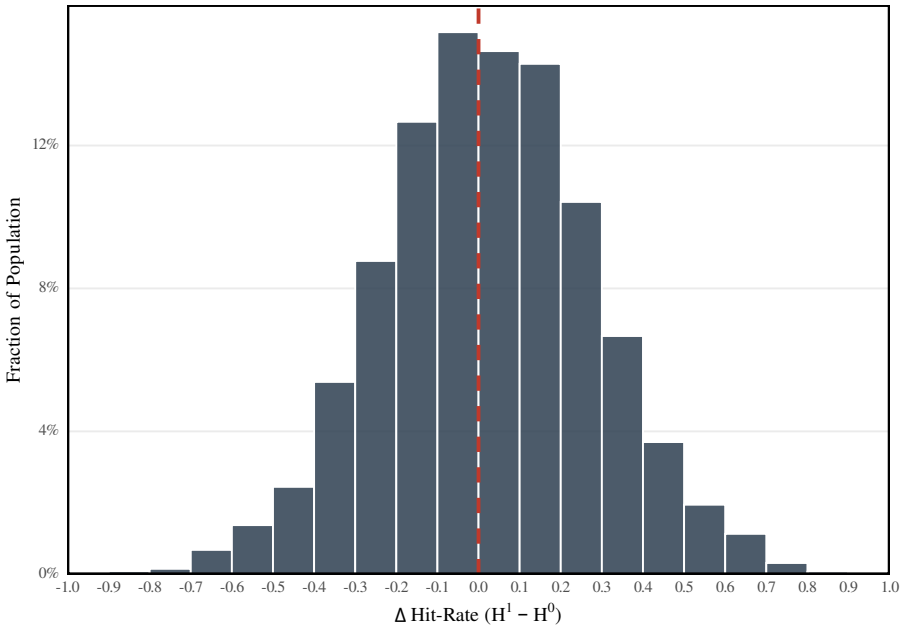


Figure A6: Empirical Bayes Estimate for ITEs

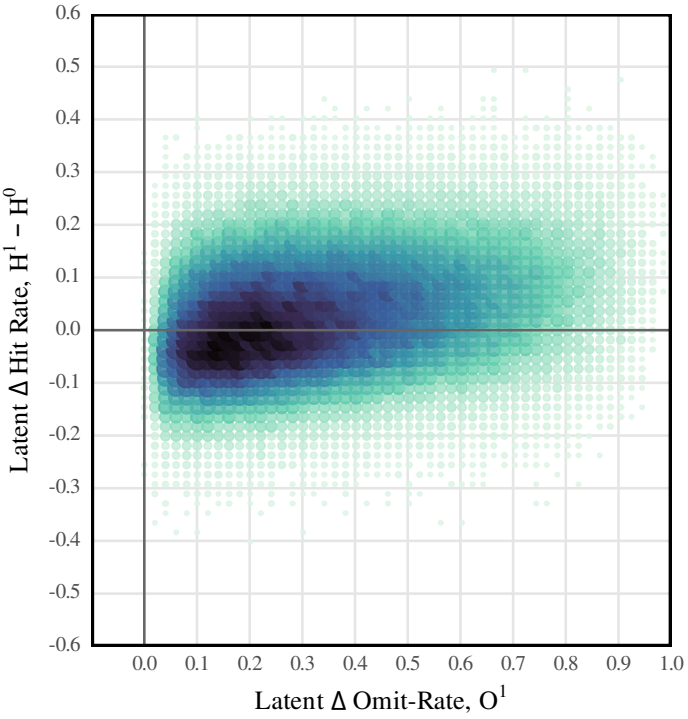


Table A4: Gender Heterogeneity Omit Rate and Hit Rate

	Omit Rate	Hit Rate
Penalty	0.330,7***	0.027,7***
	(0.003,1)	(0.003,2)
Female	0.026,2***	−0.037,2***
	(0.003,0)	(0.004,2)
Penalty × Female	0.047,3***	−0.005,2
	(0.005,3)	(0.005,6)
Observations	19,614	19,614
R2	0.485,32	0.060,80
Clustering	Student	Student

*** p <0.001, ** p <0.01, * p <0.05, . p <0.1. Student-sitting FE

Table A5: Standard Deviations and Correlations of Productivity, Sensitivity and Bias

	Female	Male	Overall
sd(π)	0.164,7	0.176,8	0.174,3
sd(ψ_Σ)	0.275,1	0.262,0	0.267,8
sd(ψ_Δ)	0.386,2	0.369,6	0.378,9
cor(π, ψ_Δ)	0.150,6	0.138,1	0.157,3
cor(π, ψ_Σ)	0.345,8	0.337,7	0.348,7
cor(ψ_Σ, ψ_Δ)	0.293,5	0.224,3	0.259,6

Table A3: Relation between AUC and the Hit-Rate Differential

	Hit-Rate	Hit-Rate	Hit-Rate	Hit-Rate
Penalty	0.041***	0.041	0.041***	0.041
	(0.003)	(0.031)	(0.003)	(0.027)
Penalty x zAUC	0.048***	0.048***	0.061***	0.061***
	(0.003)	(0.011)	(0.003)	(0.011)
Penalty x zH_0			−0.088***	−0.088***
			(0.003)	(0.008)
Observations	13,540	13,540	13,540	13,540
R^2	0.709	0.709	0.743	0.743
Clustering	Student	Student + Sit.	Student	Student + Sit.

* p <0.1, ** p <0.05, *** p <0.01. Student-sitting FE.

Figure A7: Demographic Heterogeneity in Treatment Effects

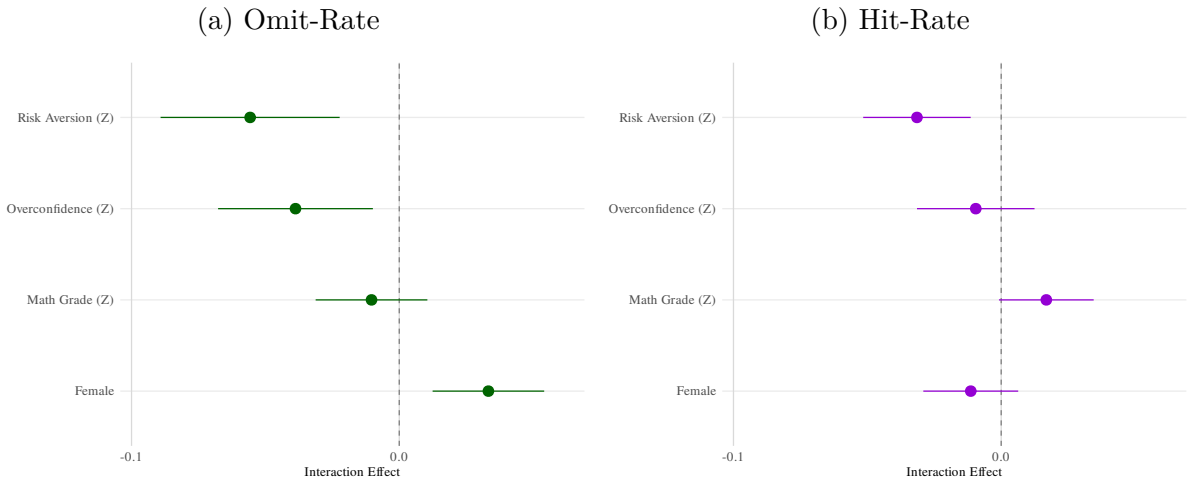


Figure A8: Density of Model Parameters by Gender (i)

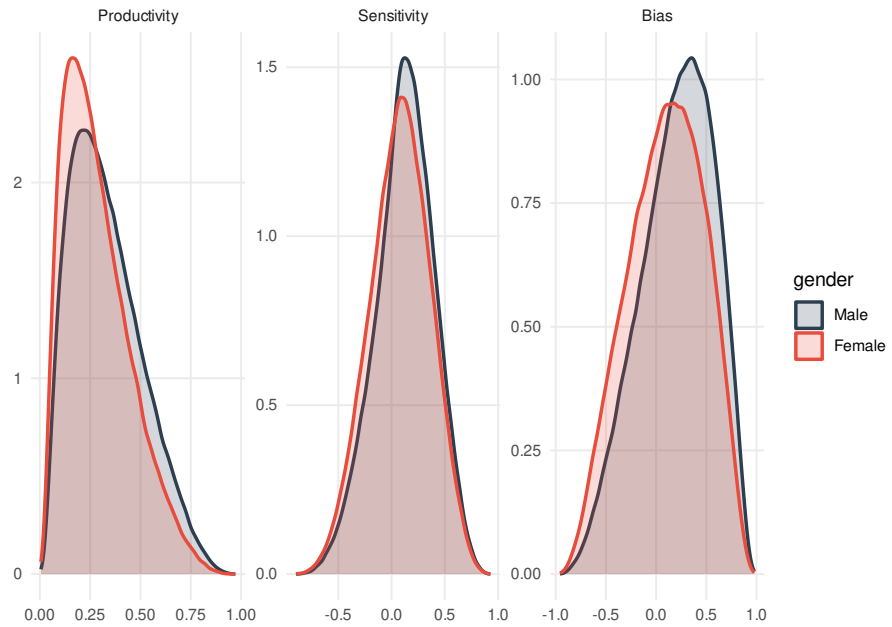


Figure A9: Density of Model Parameters by Gender (ii)

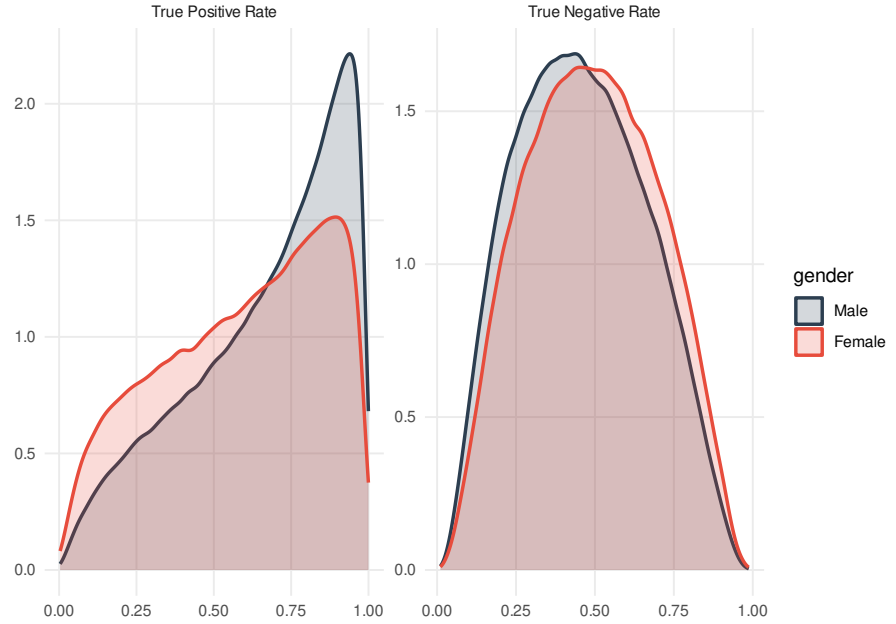


Figure A10: Average Posterior Means of Model Parameters by Gender



Figure A11: Posterior Means Density of Model Parameters by Gender (i)

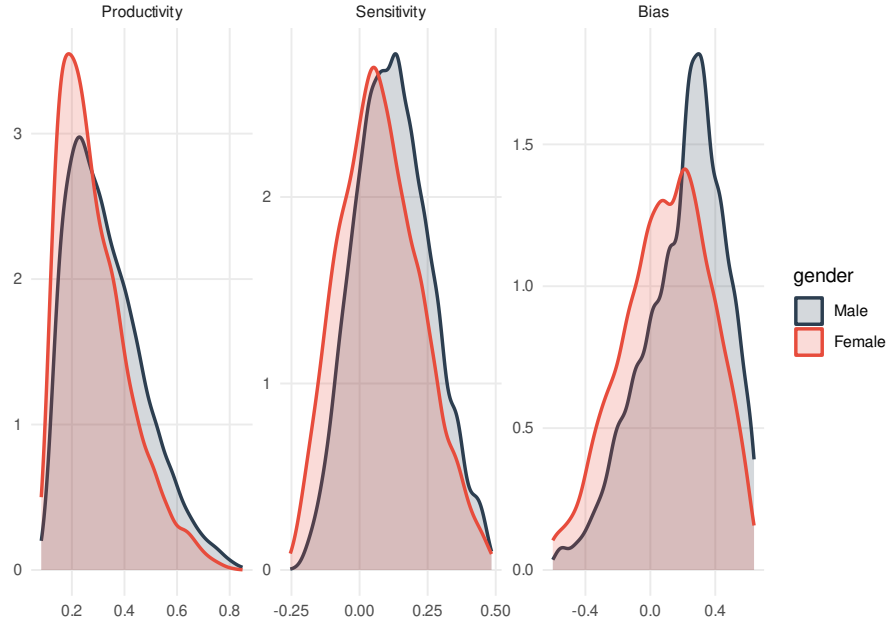


Figure A12: Posterior Means Density of Model Parameters by Gender (ii)

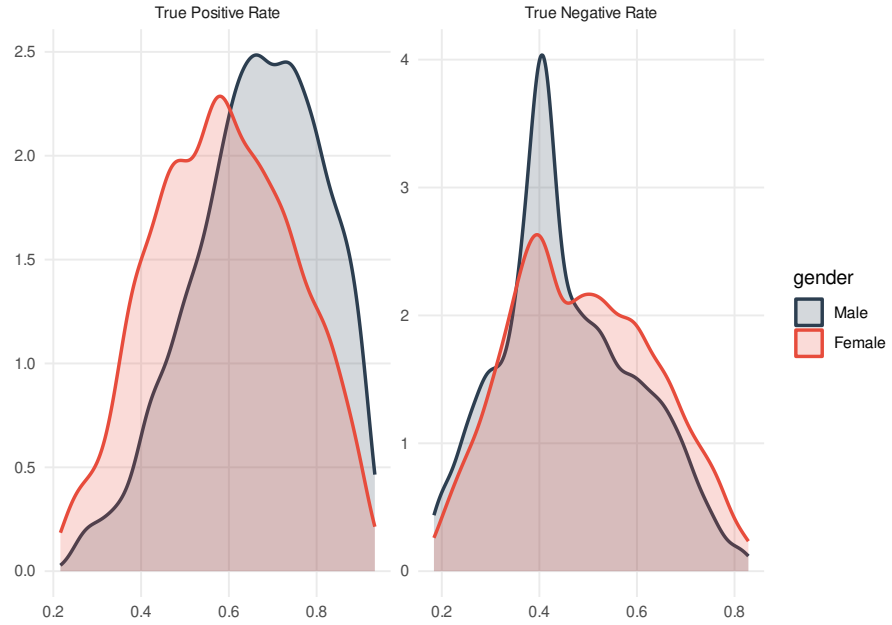


Table A6: Model Estimates Validation

	zAUC	zOC
zProductivity	0.021	−0.405***
	(0.048)	(0.026)
zSensitivity	0.235***	−0.151***
	(0.040)	(0.022)
zBias	−0.031	0.153***
	(0.051)	(0.009)
Num.Obs.	6,770	3,781
R2	0.154	0.309
Clustering	Student + Sit.	Student + Sit.

* p <0.1, ** p <0.05, *** p <0.01 Student-sitting FE

Figure A13: Empirical and Model-Implied Gender Gap

