Imperfect Metacognition and Task Selection

Francesco Bilotta*

March 15, 2025

Click here for the most recent version.

Abstract

I propose a model of task selection where efficiency derives from metacognition – the skill of monitoring one's own cognitive performance – and I underpin it with an information-processing foundation. My model provides a test for metacognition based on the elasticity of task performance to shifts in incentives. Moreover it predicts a complementarity between metacognition and ability: accurately monitoring high (low) performance is relatively more valuable for high (low) ability decision makers. I operationalize the theory analyzing field data from multiple choice tests. My analysis documents sizable metacognitive heterogeneity among students and shows that metacognitive differences may rationalize the gender gap in performance under penalties for errors, explaining why it widens in students' ability.

Keywords: Metacognition, Gender Gaps, Higher-Order Skills, Multiple-Choice Tests

*Department of Economics, Bocconi University; francesco.bilotta2@phd.unibocconi.it

I am especially grateful to my advisors Nicola Pavoni, Massimo Morelli and Luca Braghieri for constant guidance and encouragement, and to Nagore Iriberri and Pedro Rey-Biel, who generously shared their data. I thank, for valuable suggestions, Salvador Candelas, Carlo Cusumano, Stefano Della Vigna, Astrid Esparza Sánchez, Ignacio Esponda, Nicola Gennaioli, Pamela Giustinelli, Katrin Gödker, Yves Le Yaouanq, Pietro Ortoleva, Justus Preusser, Alex Smolin, as well as participants to the 1st Northwestern-Kellogg School in Economic Theory and the 2nd Bocconi-CCA-Cornell Workshop in Political Economy. The financial support of Premio Rotary Dott. Gabriele Corbelli is gratefully acknowledged. All errors remain my own.

OCEANUS

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 and cultures, a recurring distinction has emerged between the capacity to process information and the capacity to operationalize it through sound decision-making.[1] Traditionally, economics has concentrated on the former aspect, emphasizing the role of cognitive skills as a determinant of productivity within tasks. However, a growing body of evidence suggests that efficiency in selecting which tasks to perform and which to avoid is an increasingly valuable component of agents' human capital (Deming and Silliman, 2024), often critical for social welfare (e.g. medical decision-making; Currie et al., 2024).[2]

A natural question, then, is what determines efficiency in task selection contexts. In this paper, I propose a model that identifies metacognition – the skill of monitoring one's own cognitive performance (Fleming, 2024) – as a key determinant. My results provide a behavioral test for metacognitive accuracy, based on choice-data, and make predictions about the impact of metacognitive imperfections on welfare. Moreover, they show how such imperfections may emerge from Bayesian information processing. As an application, I use my model to interpret field data on student's performance in multiple choice tests, a context where efficiency in task selection is particularly consequential. Analyzing the dataset of Iriberri and Rey-Biel (2021), I document sizable heterogeneity in accuracy among students. In turn, I offer a metacognitive explanation for the gender gap under penalties for errors (negative marking) rationalizing why it widens in students' ability – a puzzling pattern for preference-based explanations.[3]

To get started, the first question is conceptual: how can we model metacognition? A natural point of departure is offered by its quantitative measurement in cognitive neuroscience, which is based on the association between objective performance and subjective confidence across experimental task trials (Fleming and Lau, 2014). Essentially, a subject is metacognitively accurate if, when relatively more confident, they are relatively more likely to be successful.[4] In my model, I adapt this idea to get an operational definition of metacognition.

I consider a decision maker (DM, they) who faces a set of tasks and has to choose which tasks to perform, accepting an outside option on the remaining ones. The objective probability of success, called performance level, is task specific. Crucially, I assume that the DM has no

[1] In the Nicomachean Ethics (≈350 BC) Aristotle contraposes the virtues of ability ("sophia") and practical wisdom ("phronesis"). Similar dichotomies appear in Confucian ("wen" vs. "zhi") and Indian ("jnana" vs. "prajna") thought. In modern psychology, Stanovich (1999) distinguishes between intelligence and rationality, echoing these ideas.

[2] Deming (2021) documents that between 1960 and 2018 (i) the share of U.S. jobs requiring decision-making skills rose from 9% to 34% and nearly half of the shift concentrates after 2007 (ii) the total wage bill for management-related occupations grew from 15% to 32%. In contrast, returns to cognitive skills remained flat or declined slightly (Edin et al., 2022). In gynecology, Currie and MacLeod (2017) find that one standard deviation improvement in decision-making would reduce unnecessary C-sections by 15.5% and increase necessary ones by 5.5%, leading to fewer complications and lower mortality.

[3] On top of Iriberri and Rey-Biel (2021), the widening gender gap is documented, for instance, by Akyol et al. (2016); Baldiga (2014); Coffman and Klinowski (2020); Conde-Ruiz et al. (2020); Riener and Wagner (2017). Moreover, the widening gap arises also as a consequence of implicit costs, such as pressure (Díez-Rituerto et al., 2025; Galasso and Profeta, 2024) or stereotypes in the phrasing of questions (Ruiz-Tagle, 2024).

[4] I use the word "accuracy" instead of the cognitive science terminology – "sensitivity" – to avoid confusion with the economic meaning of the term. Apart from this, I try to maximize coherence with the original terminology.

conception of it.[5] Instead, when deciding whether to engage in a task, the decision maker relies on a subjective forecast of their success, were they forced to act. I call such forecasts confidence judgments, coherently with the notion of propositional confidence in cognitive neuroscience (Pouget et al., 2016).

In this picture, I capture differences in ability positing that performance levels are stochastic and extracted from a decision maker-specific distribution. Then, I describe metacognition through a stochastic map collecting the distribution of the confidence judgment, conditional on any performance level. This object allows to capture the degree of association between confidence and performance in a non-parametric way. The standard introspection assumption, namely that the decision maker perfectly forecasts their odds of success, is nested as an extreme case. Taken together, ability and metacognition imply a joint distribution between confidence and performance, which is the primitive of my model, called skill-type.

This approach allows to directly incorporate metacognition in the traditional paradigm of subjective expected utility maximization. Under natural assumptions, optimal behavior is simple: the DM performs only those task where confidence exceeds a threshold that grows in risk aversion and in the relative attractiveness of the outside option (Proposition 1). Importantly, the threshold is independent from the number of task faced. As a consequence, in the long-run, the DM's behavior reveals two statistics about their skill-type. The fraction of tasks omitted (omit-rate) approximates the confidence threshold quantile, while the fraction of successful tasks among the ones performed (hit-rate) approximates the expectation of the performance levels conditional on above-threshold confidence. Moreover, defining welfare as a linear combination of the number of success, failures and omitted tasks, the hit-rate and the omit-rate pin down its asymptotic average (Proposition 2).[6]

Leveraging these behavioral statistics, my main result shows that metacognitive accuracy is measured by the elasticity of task performance to shifts in incentives (Proposition 3).

The intuition is straightforward. When task failure is not penalized compared to the outside option, performing all tasks is a dominant strategy. As a consequence, the decision maker achieves a hit-rate equal to the average performance level implied by their ability distribution. When penalties are turned on, instead, confidence needs to be sufficiently high in order to perform tasks. As a consequence, if confidence and performance are positively associated, task selection should improve the hit-rate compared to the no penalties case. Hence, while under no penalties the hit-rate elicits a measure of ability, the sign of its differential after a penalty shift is a within-individual indicator of metacognitive accuracy.[7] Additionally, fixed metacognition, the magnitude of the hit-rate differential grows for spreads in the ability distribution, capturing the intuition that metacognition matters more as performance becomes

[5]This assumption is meant to capture a form of implicit/unconscious ability within my framework. Belardinelli and Schipper (2024) formalize implicit ability in unawareness structures and provide an insightful overview of approaches to the concept in psychology, philosophy and computer science.

[6]This definition of welfare is particularly apt for contexts where every additional success yields incremental benefit. This especially true in tournament-like setting, where one could think that a fixed number of prizes is allocated to the best ranked participants. Admission tests are an example.

[7]My measure of ability is reminiscent of the way implicit ability is revealed in psychology, forced task performance (Augusto, 2010). Essentially, imposing no penalties for errors makes this procedure incentive compatible.

more variable (Proposition 4).

Whereas the hit-rate differential test has a qualitative flavor, indicating whether metacognition is overall "good" or "bad", a second point of interest is to establish what sort of metacognitive imperfections are most harmful to the decision maker.

On this point, the main observation is that a complementarity arises between ability and metacognition. The logic is particularly clear in the binary case, where there are just two performance levels and two confidence judgments. In this case, imperfect metacognition can be decomposed in Type-I (high confidence, low performance) and Type-II (low confidence, high performance) errors. I say that these denote, respectively, lack of "negative" and "positive" metacognition.[8] Consider a high ability DM, who would succeed quite often, if forced to act. Then, the major source of loss is lack of positive metacognition, as this leads to opt out of likely successful tasks, which are frequent. Conversely, for a low ability DM, loss arise mostly from lack of negative metacognition, which makes them select into tasks where they probably fail. Proposition 5 generalizes the idea that the relative value of accurately monitoring high vs low performance levels grows with ability. Behaviorally, the complementarity predicts that, comparing groups of DMs, who have similar metacognitive accuracy, but who are slanted, respectively, towards positive and negative metacognition, we should expect a widening gap in the achieved welfare as ability increases.

The operational nature of my model makes it directly applicable to interpret individual-level data on task selection problems: as an example, I consider multiple-choice tests. In particular, I re-analyze data from a large field experiment by Iriberri and Rey-Biel (2021) that features a unique within-individual variation in the scoring rule, allowing me to observe performance both with and without penalties.

First, I show that a striking fraction of students are inefficient in task selection which, according to my model, indicates bad metacognitive accuracy. Indeed, while, as one would expect, almost all of the students react to the shift in penalties omitting more questions, almost half of them displays a negative hit-rate differential (Figure 1). This finding is reinforced by a regression analysis. Whereas in a population of metacognitively accurate students the intensity of selection should show a sizable correlation with its quality, I find that on average, the effect of omitting more questions on the hit-rate differential is significantly small.

At the same time, the distribution of the differential is roughly bell-shaped around zero, suggesting sizable heterogeneity in metacognitive accuracy. Given this observation, I ask whether metacognitive differences may help to explain otherwise puzzling features of my dataset. To this end, I replicate the finding that females tend to omit more questions than males under negative marking, lowering their scores. Importantly, both these effects are much stronger at the top rather than at the bottom of the ability distribution, as measured by the hit-rate under no penalties.

On the one hand, traditional explanations, which impute the gap in omissions to gender

[8]The naming is motivated by a similar duality between the positive and negative introspection axioms for knowledge structures in epistemic logic and epistemic game theory (Battigalli and Bonanno, 1999; Van Benthem, 2010, for reference).

differences in risk aversion, need to implicitly assume that such differences are correlated with ability. I find no evidence of such correlation in my dataset; moreover I show that while risk aversion impacts the fraction of questions omitted as expected, it does not have a significant effect on scores (Figure 3).[9] On the other hand, thanks to the complementarity between ability and metacognition, the widening gaps are naturally rationalized in my model, hypothesizing that males' metacognition is more attuned to monitor high performance levels (i.e. positive) while female metacognition low performance ones (i.e. negative). To make this point quantitatively I provide maximum likelihood estimate of my model pooling observations by gender (Figure 4). Moreover, I find independent support for this hypothesis in the data of Bordalo et al. (2019). In their lab experiment, students answer a multiple choice questionnaire reporting the confidence that their answer is correct. Therein, I document that, while, on average confidence is too low on easy tasks and too high on difficult tasks – the hard-easy effect, (Moore and Healy, 2008) – the first dimension of the effect is stronger for females, while the second for males.

The discussion above demonstrates, I argue, that my approach to metacognition yields both interesting predictions and a direct connection to empirical evidence. Nonetheless, some conceptual point remained open: how can we give a foundation to the joint distribution between confidence and performance? What accounts for potential heterogeneity in its association?

As a final contribution, I provide an information processing rationale for skill-types, where tasks are decision problems in the sense of Blackwell (1953). The main idea is that, in each decision problem, the decision maker observes a signal from an experiment whose informativeness they ignore and they process (second-order) information about this feature of the environment before taking action. Then, ability derives from the distribution of experiments' informativeness across tasks, which determines performance, while metacognition is regulated by the accuracy of second-order information (Proposition 7). A sharper characterization is obtained for symmetric environments, where each action is ex-post optimal with the same frequency. In this case, I show that martingale coupling between confidence and performance (Strassen, 1965) is a necessary and sufficient condition for skill types to be "implementable" in the information processing model (Proposition 8).

Roadmap I conclude the Introduction discussing the related literature. Section 1 lays out the setup of my model. Section 2 develops the main tools of analysis. Section 3 illustrates the main intuitions of the paper in a binary setting. Section 4 presents the empirical analysis. The next two Sections present general results: Section 5 characterizes the relation between the hit-rate differential and metacognition quality; Section 6 studies the complementarity between ability and metacognition. Section 7 proposes the information-processing foundation. Section 8 concludes. I study the perfect metacognition benchmark in Appendix A. Proofs are

[9] On top of this, in the literature (Benjamin et al., 2013; Dohmen et al., 2010) higher cognitive ability is generally associated with lower risk aversion: this implies a compression of gender differences at the top.

contained in Appendix B. Details on the empirical analysis are provided in Appendix C.

Related Literature

This paper contributes to the growing field of cognitive economics (Enke, 2024; Woodford, 2020, for a review).

At a high level, my main contribution is to integrate the insights of research on metacognition (Fleming, 2024, for a review) in an economic model of decision-making, bridging a gap between the approach to confidence in the cognitive science and in behavioral economic theory. Traditionally, the literature on confidence within economics (Moore and Healy, 2008, for a review) mostly focused on the average displacement between confidence and performance. Instead, in my framework I explicitly model and microfound their joint distribution, importing an insight from the cognitive sciences. In the other direction, I show how the economic approach brings to the table (i) a test for metacognition which is based on "revealed confidence" rather than on confidence reports and (ii) a natural framework to assess consequences of metacognitive imperfections on welfare.

Part of the cognitive economics literature has shown that low confidence reports, i.e. cognitive uncertainty, systematically predict regress to a default action and that the effect becomes more sizable as the complexity of the task rises (Enke and Graeber, 2023). This attenuation pattern, has been replicated through a larger scale experiment in Enke et al. (2024), additionally showing that the elasticity of agents' decisions to fundamentals in problems with a normative solution (e.g. belief updating) is always smaller than what normative theory predicts. Whereas this work implicitly stands on the idea that a good association exists between confidence and performance, Enke et al. (2023) show that this relationship shows great heterogeneity across tasks.[10] This evidence can be seen as strong motivation for my model, which allows a fine-grained description of the relationship between confidence and performance. Moreover, a possible reading of the hit-rate differential test is that lack of metacognition generates a form of attenuation across tasks, rather than within: task selection quality responds to shift in incentives less than it would under the benchmark.

Analogously, whereas many models of attenuation are based on the idea that individuals produce cognitive signals on the optimal action to take within a problem (Ilut and Valchev, 2023, for instance), but are sure about the experiment which generates them, in my information processing foundation a second layer of noise exists: the quality of second-order information underlies metacognitive accuracy. Relatedly, Augenblick et al. (2024) studies and experimentally tests a Gaussian updating model where the strength of external information is imperfectly coded, in order to explain overinference from weak signals and underinference from strong signals (OWS-USS) through an attenuation logic.[11] In my information processing

[10] The authors then study experimentally the implications of this mechanism for group decisions, showing that their quality improves when individuals' confidence accurately tracks performance, since coalescence towards the most confident opinions filters bias. In psychology, a similar point is made by Koriat (2012).

[11] See Vaeth (2024) for an interesting contribution showing, how beyond Gaussian signals, attenuation is a prediction of additive noise models with symmetric, quasi-concave noise under a symmetric, log-concave prior.

model the focus is on cognitive signals, and second-order information does not necessarily have an additive structure. In symmetric environments, my model also shows OWS-USS and can accommodate asymmetries between the two side of the effects through a false positive/false negatives logic. My empirical analysis hints at gender-based heterogeneity in the hard-easy effect, motivating why such asymmetries may be of interest.

The focus on task selection efficiency relates my work to Caplin et al. (2023) who measure efficiency in the related problem allocating resources to productive inputs, couching their experimental analysis in a rational inattention framework. My model speaks to this research, suggesting that when it comes to self-productivity, such allocative skill may derive from metacognition, which can be assessed through the hit-rate differential.

The application to multiple choice tests relates my paper to work combining ideas from item response theory (Lord, 1980) with expected utility maximization. In particular, Espinosa and Gardeazabal (2010) and Akyol et al. (2016) propose models to explore the role of risk aversion in tests, and then apply them in numerical and structural analysis. My theory focuses attention on question selection, which did not receive much consideration in the literature, especially if compared to omissions. In particular, this paper is the first to analyze the hit-rate differential, and to propose a cognitive explanation for the puzzling fact that gender gaps widen in ability (to the best of my knowledge). In this sense, it contributes also to the empirical literature on negative marking (Baldiga, 2014; Coffman and Klinowski, 2020; Conde-Ruiz et al., 2020; Iriberri and Rey-Biel, 2021; Riener and Wagner, 2017, and references therein).

Finally, on a technical level, I work with notions of dependence between random variables and multivariate stochastic orders (de Castro, 2009, for a review). In particular, to establish my main result, I introduce a notion which weakens the classical notion of positive quadrant dependence (Lehmann, 1966).

1 A Model of Task Selection

In this Section, I introduce the setup of my model.

Tasks A decision maker faces a menu of $N \in \mathbb{N}$ independent lotteries. Each lottery represents a *task*. The DM's problem is to choose which tasks to perform and which to avoid, accepting an outside option. Conditional on performing a task, the DM may either succeed or fail. The probability of success in task $i \in [N] \equiv \{1, \dots, N\}$ is called *performance level* and denoted by $p_i \in [0, 1]$. This quantity describes the difficulty of the task for the DM.

Welfare When the decision maker performs a task, they receive $h \in \mathbb{R}$ (for "hit") units of welfare in case of success and m (for "missed") units of welfare in case of failure. Instead,

when they opt out, they receive o units of welfare. The triple $\Delta = (m, o, h)$ defines a *decision environment*. To avoid trivialities, I assume $m \leq o < h$, so that (generically) opting out does not dominate opting in. I say that the environment has *penalties* if $m < o$; otherwise I say it has no penalties. Welfare is accumulated additively across tasks. As a consequence, when the DM chooses to perform a portfolio of tasks $A \subseteq [N]$, their *total welfare* is a random variable defined by

$$W(A; (p_i)_{i=1}^N, \Delta) = oN - \Delta^o |A| + \Delta^h \text{ PB}(A; (p_i)_{i=1}^N). \tag{Welfare}$$

In the Equation, $\Delta^o = o - m$ and $\Delta^h = h - m$ are, respectively, the relative gains from opting out and succeeding in a task, compared to failing. The random variable $\text{PB}(A; (p_i)_{i=1}^N)$ counts the number of successes in the portfolio A. Note that this is a Poisson-Binomial, i.e. the sum of independent Bernoulli variables with parameters $(p_i)_{i \in A}$, hence the notation.[12]

Ability I model the decision maker's ability as the (stochastic) process that determines performance levels. In particular, I posit that Nature draws the vector $(p_i)_{i=1}^N$ i.i.d. from a full support *ability distribution* $\pi \in \Delta(P)$, where $P \subset [0, 1]$ is a finite set of admissible performance levels.[13] Hence, $\pi(p)$ describes the limit frequency of tasks where probability of success is p. As a consequence, the larger the ability distribution π (in stochastic terms) the more often high performance levels are accessed by the DM. Slightly abusing notation, P also denotes a random variable of law π.

Metacognition In my framework, ability is implicit, in the sense that the decision maker does not perceive the realized performance levels. Instead, in each task $i \in [N]$ they form a probabilistic forecast of success, $c_i \in [0, 1]$, called *confidence judgment*. I model the DM's *metacognition* as the (stochastic) process which regulates confidence, conditional on performance. In particular, I consider a *metacognition map* $\psi : P \to \Delta(C)$, where $C \subset [0, 1]$ is a finite set of admissible confidence judgments. I posit that, given the vector $(p_i)_{i=1}^N$, Nature draws c_i from $\psi(p_i)$, independently across tasks. Hence, $\psi(c|p)$ describes the limit frequency of tasks where the DM feels confidence c, among those where he succeeds with probability p. In this sense, the metacognition map ψ describes how accurately the DM monitors their own ability. As a special case one gets the (standard) case of an agent who can perfectly forecast their odds of success: I say that a DM has *perfect metacognition* if $\psi \equiv id : p \mapsto \chi_p$ (where $\chi_\bullet$ denotes the Dirac's distribution). This benchmark is explored in Appendix A; in the main text, I mostly focus on departures from it.

[12] The Welfare Equation admits a financial interpretation. Think oN in analogy to initial wealth, $\Delta^o|A|$ as the deterministic opportunity cost accrued to perform portfolio A, and $\Delta^h \text{PB}(A; (p_i)_{i=1}^N)$ as the random gains associated with performance.

[13] Throughout the paper, I assume finiteness for expository convenience: my arguments generalize with standard technical modifications.

Skill-Type Ability and metacognition, together, imply a joint distribution between performance levels and confidence judgments, $\tau \in \Delta(P \times C)$, called a *skill-type*. This object can be equivalently taken as primitive.[14] When simpler, I formulate results in terms of skill-types. Observe that the second marginal of τ describes the unconditional distribution of confidence judgments: I denote it by γ. As for performance levels, slightly abusing notation, C also denotes a random variable of law γ. Then, τ is the law of the random vector (P, C).

The DM's Problem The decision maker chooses a portfolio of tasks to maximize their subjective expected utility from total welfare levels. I allow for risk aversion, restricting to CARA preferences for tractability. Given a *risk aversion coefficient* $\alpha \geq 0$, the DM's utility function is

$$u_\alpha(W) = \begin{cases} 1 - e^{-\alpha W} & if \quad \alpha > 0, \\ W & if \quad \alpha = 0. \end{cases} \tag{1}$$

Since the DM uses the vector of confidence judgments to forecast their probability of success in each task, they perceive a subjective analogue to Equation Welfare, where the number of successes is distributed according to $\text{PB}(A; (c_i)_{i=1}^N)$. Hence, the DM's objective is

$$V((c_i)_{i=1}^N; \alpha, \Delta) \equiv \max_{A \subseteq [N]} \mathbb{E}\left[u_\alpha(W(A; (c_i)_{i=1}^N, \Delta))\right]. \tag{DM's Problem}$$

Since the action space is finite, a solution always exists. For uniqueness, I assume that ties are broken in favor of inaction: when indifferent between A and $A \cup \{j\}$ for some $j \in [N] \setminus A$, the DM chooses A.[15] I denote by $A^*((c_i)_{i=1}^N; \alpha, \Delta)$ the subjectively optimal portfolio. Note that, conditional on a realization of confidence judgments, the DM's Problem is independent from the skill-type; hence its solution, too. Substituting $A^*((c_i)_{i=1}^N; \alpha, \Delta)$ in the expression for Welfare, one obtains that the *outcome* of the decision problem is a realization of the random variable $W(A^*; (p_i)_{i=1}^N, \Delta)$. Hence, while task selection is guided by confidence judgments, performance levels determine the number of successes, within the selected tasks.

Timing To sum up, my model has the following timing: (i) Nature draws N performance-confidence pairs $(p_i, c_i)_{i=1}^N$ independently from τ; (ii) given $(c_i)_{i=1}^N$ the DM solves DM's Problem, where Welfare is perceived as $W(A; (c_i)_{i=1}^N, \Delta)$; (iii) given the DM's decision, A^*, Nature draws a Welfare realization from $W(A^*; (p_i)_{i=1}^N, \Delta)$.

DM's Sophistication For full clarity, I stress that given a confidence judgment, the decision maker does not make inferences about the latent performance level; instead they take it at face value. Accordingly (i) the distribution π is not a prior (ii) the stochastic map ψ is not a Blackwell experiment and (iii) confidence judgments c's are not signals. Nonetheless, in

[14] For any (π, ψ) we can define $\tau(p, c) = \pi(p)\psi(c|p)$. Conversely, given any $\tau \in \Delta(O)$, for some $O \subset [0,1]^2$, its first marginal defines a ability distribution $\pi \in \Delta(P)$, while an metacognition map $\psi : P \to \Delta(C)$ is defined by conditioning: $\psi(p) = \tau|p$.

[15] Ties are non generic, as arbitrarily small perturbations in the support of τ resolve them.

Section 7 I show that some instances of my model are compatible with a Bayesian information processing foundation. Hence, despite appearances, I argue that my DM is not naive.

Example: Multiple Choice Tests To keep an example in mind, in my application (Section 4) the DM is a student, tasks are multiple choice questions and their collection is a test. In this context, decision environments are identified with scoring rules and welfare with the test score. Performance levels can be interpreted as a measure of question-specific ability, whereas confidence judgments gauge its feeling.

2 Optimal Decision Rule and Long-Run Behavior

In this Section I develop the main tools of analysis. First, I obtain the optimal decision rule for the DM's Problem. Then, I leverage it to characterize three long-run behavioral statistics: the rate of task omission (omit-rate), the rate of success in performed tasks (hit-rate) and the average welfare. These statistics are the focus of my results in the next Sections.

Optimal Decision Rule Under my assumptions, subjectively optimal behavior has tractable and intuitive structure. Given the independence of tasks and the additivity of the total welfare, a first order stochastic dominance argument reveals that the DM performs all those tasks where they feel sufficiently confident, as measured by a threshold. In principle, this threshold depends on the number of tasks faced. Thanks to the CARA assumption, accumulating additional welfare does not alter the marginal evaluation of each task, hence uniformity is achieved.[16] Moreover, its value can be computed, exploiting the Poisson-Binomial moment generating function.

Proposition 1 *Fix a number of tasks N. Consider a decision maker with risk aversion coefficient $\alpha \geq 0$. The optimal decision rule has a threshold structure,*

$$A^*((c_i)_{i=1}^N; \alpha, \Delta) = \{i \in [N] \mid c_i > c^*(\alpha, \Delta)\} \quad \textit{where} \quad c^*(\alpha, \Delta) = \begin{cases} \frac{1-e^{-\alpha\Delta^o}}{1-e^{-\alpha\Delta^h}} & if \quad \alpha > 0, \\ \frac{\Delta^o}{\Delta^h} & if \quad \alpha = 0. \end{cases} \tag{Decision Rule}$$

Naturally, the confidence threshold grows in the risk aversion coefficient (α) and in the relative attractiveness of opting out a task (Δ^o) versus succeeding (Δ^h); on the other hand, it is independent from the skill-type. In this sense, the model neatly separates preferences from metacognition: the first aspect determine a criterion the DM would like to adhere to in order to perform tasks, while the second their efficiency in the selection process. Moreover, while selection is simultaneous in the timing above, the decision rule can be equivalently implemented sequentially.

[16] If the utility is not CARA, there exists a sequence of thresholds $(c_N^{*,u})_{N\in\mathbb{N}}$ that describe the DM's behavior in tests of different length. If the sequence converges to a limit $c^{*,u}$, the main results of the paper still hold, up to approximation, for the threshold $c^{*,u}$.

Behavioral Statistics The uniformity of the decision rule allows to focus on task selection in the large-N limit, abstracting from sample effects. To this end, I define two behavioral statistics of interest: the *omit-rate*, which is the limit fraction of tasks where the outside option is accepted and the *hit-rate*, which is the limit fraction of successful tasks among attempted ones: respectively,

$$\overline{O}(\tau, \alpha, \Delta) = 1 - \operatorname*{plim}_{N\to\infty} \frac{|A^*|}{N} \quad \text{and} \quad \overline{H}(\tau, \alpha, \Delta) = \operatorname*{plim}_{N\to\infty} \frac{\mathrm{PB}(A^*; (p_i)_{i=1}^N)}{|A^*|}.$$

Note that both quantities can be straightforwardly estimated observing behavior in task selection contexts (such as multiple-choice tests). The next result leverages Proposition 1 to show that the two behavioral statistics are functionals of the DM's skill-type, establishing a mapping between skill-types and behavior. In turn, it connects the statistics to welfare. Note that, since success happens infinitely often, Welfare diverges. To this end, I impose a $1/N$-normalization and focus on the *asymptotic average welfare*,

$$\overline{W}(\tau, \alpha, \Delta) = \operatorname*{plim}_{N\to\infty} \frac{W(A^*; (p_i)_{i=1}^N, \Delta)}{N}.$$

Proposition 2 *Let τ be the decision maker's skill-type. It holds*

$$\overline{O}(\tau, \alpha, \Delta) = F_\gamma(c^*) \quad \textit{and} \quad \overline{H}(\tau, \alpha, \Delta) = \mathbb{E}_\tau[p | c > c^*].$$

In turn, the asymptotic average welfare rewrites

$$\overline{W}(\tau, \alpha, \Delta) = m + \Delta^o \overline{O} + \Delta^h (1 - \overline{O}) \overline{H}. \tag{AAW}$$

The first part of the result is a direct consequence of the (weak) Law of Large numbers. The welfare decomposition, while intuitive, is not obvious ex-ante, since the distribution $W(A^*)$ is itself stochastic (conditional on each number of tasks N, it depends on the realization of $(p_i, c_i)_{i=1}^N$). The proof builds on a double limit argument. Since the variability in the welfare per-task vanishes, using Chebyshev's inequality I show that the sequence of random variables $W(A^*; (p_i)_{i=1}^N, \Delta)/N$ has the same probability limit as the sequence of their expectations. These are functions of the fraction of task omitted and hit, then one can conclude leveraging the first point. The main upshot of the decomposition in Equation AAW is that my environment is "ergodic": the asymptotic averaged welfare is just the expected welfare per task.

3 Binary Skill-Types

In this Section I study binary skill-types. This serves two purposes: the first is to develop parametrically the intuitions underlying the general results, which are postponed to Sections

5 and 6; the second is to provide a shortcut to the empirical analysis, anticipated to Section 4.

Imagine that tasks can be partitioned in two groups, easy and difficult, so that the decision maker performance comes in two levels, $\overline{p} > \underline{p}$. Similarly, suppose that the DM's confidence judgments are also binary $\overline{c} > \underline{c}$ (and possibly detached from performance). In this case, the ability distribution is identified by the frequency of high performance, $\overline{\pi} \equiv \pi(\overline{p}) \in (0,1)$ while the metacognition map is a row-stochastic matrix,

$$\psi = \left(\begin{array}{c|cc} P/C & \underline{c} & \overline{c} \\ \hline \underline{p} & \psi_- & 1-\psi_- \\ \overline{p} & 1-\psi_+ & \psi_+ \end{array}\right) \quad for \quad \psi_+, \psi_- \in (0,1). \qquad \text{(Binary Model)}$$

Then, the performance-confidence relationship is described by two parameters: the true positive rate ψ_+ and the true negative rate ψ_-, capturing, respectively, the ability of monitoring high and low performance. I call ψ_+ *positive metacognition* and ψ_- *negative metacognition*. In turn, the sum $\psi_+ + \psi_-$ is a natural measure for the overall accuracy of the performance-confidence relationship, which I call *total metacognition*.

In this setting, Proposition 2 yields simple closed-form expressions for the behavioral statistics, which can be used to illustrate the main messages of the paper. To start, let's substantiate the intuition that metacognition can be read in the elasticity of the hit-rate to shifts in incentives. As anticipated, when errors are not penalized with respect to omissions ($\Delta^o = 0$) the outside option is dominated and there is no scope for task selection. Then, the hit rate is just the average performance level,

$$\overline{H}^0 \equiv \mathbb{E}_\pi[p] = (1-\overline{\pi})\underline{p} + \overline{\pi}\overline{p}. \tag{2}$$

Hence, while performance under no penalties reveals a measure of ability, it does not carry any information about the DM's metacognition.

Suppose that penalties are progressively increased. Whatever the DM's risk attitudes, at some point the confidence threshold will rise enough to exceed the low confidence judgment, while keeping below the high one, $c^* \in (\underline{c}, \overline{c})$. Behaviorally, we will observe the DM starts opting out of some – but not all – tasks. When this happens, tasks are performed only conditional on high confidence. Hence the hit-rate shifts from $\overline{H}^0$ to

$$\overline{H}^{>0} = (1-\overline{\pi})\frac{1}{(1-\overline{\pi})+\overline{\pi}\rho}\underline{p} + \overline{\pi}\frac{\rho}{(1-\overline{\pi})+\overline{\pi}\rho}\overline{p}. \tag{3}$$

where $\rho = \psi_+/1-\psi_- \in (0,\infty)$ is the ratio between the true positives and the false positives of the metacognition map. This ratio controls the (affine) transformation of the hit rate from H^0 to $H^{>0}$. Comparing Equations 2 and 3 shows that, the higher ρ, the more the convex weight on the low performance level is shrunk in favor of dilating that on the high performance level.

This suggests that (total) metacognition is captured by the hit-rate differential,

$$\overline{D} \equiv H^{>0} - H^{0} = \underbrace{\frac{\overline{\pi}(1-\overline{\pi})\overbrace{(\overline{p}-\underline{p})}^{\text{cognitive variance}}}{(1-\overline{\pi})(1-\psi_{-})+\overline{\pi}\psi_{+}}}_{1-\text{omit rate}} \cdot \underbrace{[-1+(\psi_{+}+\psi_{-})]}_{\text{total metacognition}}. \tag{4}$$

This expression provides an identification between the sign of the hit-rate differential and metacognitive accuracy. Take as reference "null" metacognition, namely the case when confidence and performance match at random ($\psi_{-} = \psi_{+} = 1/2$). Then, task selection neither improves nor it decrements the hit-rate, implying a null differential. In turn, a necessary and sufficient condition for "good" metacognition – i.e. more accurate than a random matching – is that the hit-rate differential is positive. Viceversa for the case of "bad" metacognition. From an empirical perspective, this result provides a theory-based test for metacognitive accuracy, developed in the next Section (Figure 1).

As an additional observation, note that, fixed metacognition, the absolute value of the hit-rate differential grows when the ability distribution is stretched, dilating $\overline{p} - \underline{p}$. This captures the intuition that metacognition is particularly relevant for those that have sometimes very high and sometimes very low performance – high cognitive variance.

The equivalence between metacognitive accuracy and efficiency in task selection is generalized to arbitrary skill-types in Proposition 3, which states that a DM shows a nonnegative hit-rate differential in any decision environment if and only if their skill-type has Weak Positive Dependence. The point on the effect of cognitive variance is generalized in Proposition 4, defining the notion of Centrifugal Spread.

When skill-types are binary, by construction, lack of metacognitive accuracy is driven by two decision-making errors. The first, generated by the false positives of the metacognition map (lack of negative metacognition) is "neglecting threats", i.e. failing to opting out low performance tasks. The second, generated by the false negatives of the metacognition map (lack of positive metacognition) is "missing opportunities", i.e. failing to opting in high performance tasks. Whereas both kinds of error contribute to reduce the hit-rate differential, intuition suggests they have different consequences depending on the relative abundance of "opportunities" vs. "threats" faced by the decision maker. Indeed, in my model a natural complementarity arises between ability and metacognition, so that lack of positive (negative) metacognition is particularly damaging for high (low) ability skill-types.

To grasp why this is the case, note that when no tasks are omitted (as in the $\Delta^{o} = 0$ case) the asymptotic average welfare depends only on ability: it is a linear transformation of the hit rate $\overline{H}^{0}$,

$$\overline{W}^{0} = m + \Delta^{h}\mathbb{E}_{\pi}[p].$$

To understand returns to metacognition under task selection, observe that, in this case the welfare level can be decomposed as a sum of $\overline{W}^0$ with a term capturing the (linear) interaction of ability and metacognition,

$$\overline{W} = \overline{W}^0 + \psi_- \underbrace{(1-\overline{\pi})[\Delta^o - \underline{p}\Delta^h]}_{\partial\overline{W}/\partial\psi_-} - (1-\psi_+)\underbrace{\overline{\pi}[\overline{p}\Delta^h - \Delta^o]}_{\partial\overline{W}/\partial\psi_+}. \tag{5}$$

Thus, returns to negative introspection reflect the benefit of avoiding losses in the face of a threat, while returns to positive introspection reflect the benefit of saving the cost of a foregone opportunity. Both returns can be decomposed in an extensive margin (respectively $1-\overline{\pi}$ and π) that determines how often negative and positive metacognition is required to the DM, and an intensive margin (respectively $\Delta^o - \underline{p}\Delta^h$a and $\overline{p}\Delta^h - \Delta^o$) determining the stakes associated to threats and opportunities. Hence, intuition suggests that, whatever the intensive margin, for extreme values of $\overline{\pi}$ the extensive margin dominates and either negative or positive introspection is more valuable. Indeed, taking the difference between the returns, we obtain

$$\frac{\partial\overline{W}}{\partial\psi_+} - \frac{\partial\overline{W}}{\partial\psi_-} = \Delta^h \mathbb{E}_\pi[p] - \Delta^o. \tag{6}$$

Note that the right-hand side equals the difference in welfare between the case where the DM performs all tasks and that where they perform none. Hence, this expression shows that the value of positive relative to negative metacognition is monotonic in the expected performance level, and that positive metacognition becomes more valuable than negative as soon as $\mathbb{E}_\pi[p] \geq \Delta^o/\Delta^h$ – a simple condition for "sufficient ability". Observe that, in general, there are cases where either positive or negative metacognition have negative returns. A necessary and sufficient condition for the positivity of both is $\underline{p} < \frac{\Delta^o}{\Delta^h} < \overline{p}$, which is satisfied restricting to sufficiently spread ability distributions.

The argument above shows that the widening gap in returns is essentially driven by how omission behavior varies with ability. As a final point, let's record that the omit-rate is an average of the true negative rate and the false positive one, weighted by the frequency of high performance; hence, it can be rewritten as

$$\overline{O} = \psi_- + \overline{\pi}[1 - (\psi_+ + \psi_-)]. \tag{7}$$

It follows that, while metacognitive accuracy regulates the elasticity of omission behavior across ability levels, relatively more negatively metacognitive DMs unconditionally tend to omit more tasks.

The complementarity between ability and metacognition is generalized to arbitrary skill-types in Proposition 5, introducing the notion of Metacognition Shift. The same Proposition also shows an inequality linking the gaps in omissions and welfare. Empirically, the interest of the complementarity is that it provides an explanation to the widening gender gaps in scores

and omissions observed in multiple-choice exams, as I argue in the next Section (Figure 2).

4 Metacognition in Multiple-Choice Tests

In this Section, I apply my framework to interpret existing field data on student's behavior in multiple-choice tests.

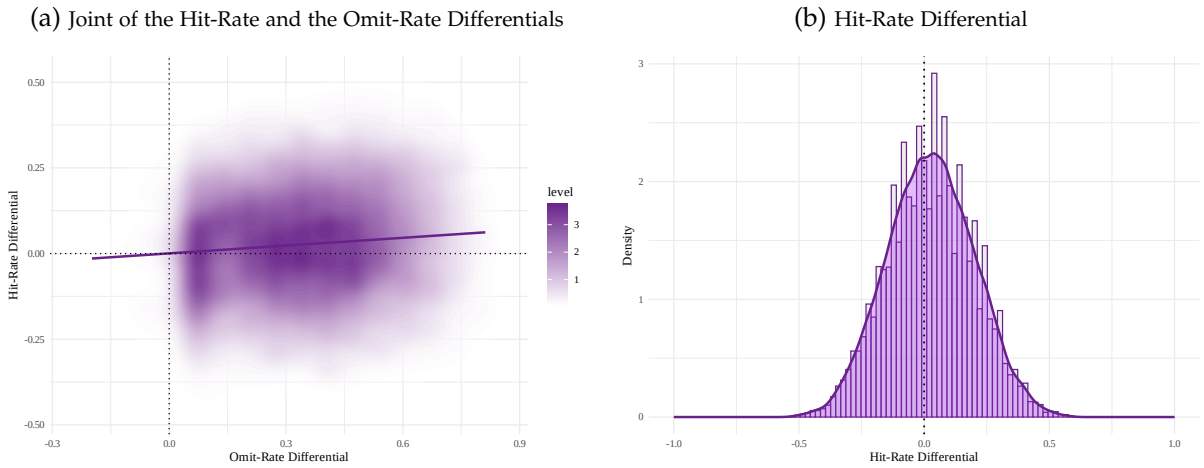
The Dataset I analyze data from a large field experiment, conducted by Iriberri and Rey-Biel (2021). The authors collaborate with the organizers of the 2016, 2017 and 2018 editions of the *Concurso de Primavera de Matemáticas*, a regional math contest held annually in the Madrid area. For around 10000 students, I observe question-level performance in a multiple-choice math test. The test features a unique within-individual variation in the scoring rule: the first part of the test (13 questions) has no penalty for errors relative to omissions, while the second part (12 questions) does (i.e. there is negative marking).[17] The two blocks are designed to be comparable under any other respect, including average question difficulty. Students are free to answer questions in the order they prefer: the structure of the test does not constrain them in any sense. Naturally, they receive no feedback whatsoever during the test. On top of performance, the dataset records demographics including gender, age group (4 groups spanning between 10 and 17 years old), participation time and average math grades at school (normalized at school level). Moreover, for participants in the 2017 and 2018 editions (approximately 5300 observations) I observe a survey measure of risk aversion, on a decreasing scale from 1 to 5.[18]

Theory-Data Mapping To map my theory into the data, I interpret each question in the test as a task. Hence, the scoring rule implemented in each part of the test defines a decision environment and, based on Proposition 1, I can estimate the omit-rate, the hit-rate and the (average) welfare, in both parts of the test, at the individual level. To attenuate measurement error concerns, I employ an empirical Bayes estimation strategy, effectively shrinking the noisier estimates towards the sample mean (Gelman et al., 1995, see Appendix C for details). Since students received a different test based on their age group and year of participation, I normalize variables at the group-year level for between-students comparisons. I use plug-in estimators for functions of the behavioral statistics.

[17] In the first part, correct answers are evaluated 5 points, omitted or wrong ones 0. In the second part, the score for correct and wrong answers is the same as in the first, but the one for omitted answers is increased to 1; this is an instance of *formula scoring* (Holzinger, 1924; Thurstone, 1919): a risk neutral individual is indifferent between attempting every question at random and leaving every question blank.

[18] The question used to elicit risk aversion is the following: *When omitting a question was worth 1 point, I answered the question (1) when I was absolutely sure (2) when I was almost sure (3) when I was uncertain between 2 answers (4) when I was uncertain between 3 answers (5) always*. Note that, interpreting a question as a lottery, (5) reveals risk neutrality, since each question has 5 candidate answers and the scoring rule is 5 points for hits, 1 for omission and 0 for errors.

Figure 1: Hit-Rate Differential Test



Notes: Panel (a) shows levels for the estimated density (100 bins) of the joint distribution of the hit-rate and omit-rate differentials. Moreover, it displays a simple regression line to show the correlation between the two variables. Panel (b) shows the density of the hit-rate differential in the sample. Since the variable has discrete, but very large support, I plot both a kernel density estimate of the probability mass function, and a 100 bins histogram.

Efficiency in Task Selection

The first part of my analysis concerns efficiency in task selection. I show that, while variation in penalties generates a within-individual increase in the fraction of omitted questions, selection is inefficient. I take this as evidence of imperfect metacognition in my sample.

Impact of Penalties on Hits and Omissions To start with, I estimate the hit-rate differential at the individual level, taking the difference between the empirical Bayes estimate of the hit-rate in the two parts of the test. Similarly, I estimate, at the individual level, the omit-rate differential, defined by the difference of the omit-rate in the two parts of the test.[19] Figure 1a reports their estimated joint-distribution, within my sample. Observe that, as expected, students react to the shift in penalties omitting sensibly more questions: the omit-rate differential is nonnegative for virtually all of the sample (98%) and is around 0.3 on average. At the same time, a sizable fraction of students shows a negative hit-rate differential, whatever the fraction of questions omitted. Zooming into this finding, Figure 1b reports the marginal distribution of the hit-rate differential within my sample. I find that, strikingly, 46% of the students display a negative hit-rate differential; moreover, its distribution is roughly bell-shaped around the median, 0.02, and supported in the interval [-0.5,0.5] (first tertile -0.05, second tertile 0.1). Hence, while students responds to shift in penalties, they do so inefficiently, reducing their hit-rate.

[19]My theory predicts no omissions under no penalties, hence the omit-rate differential should be equal to the omit-rate under penalties. In the sample, I observe that 22% of students omit at least one question even under no negative marking. Nonetheless, 90% omit just one or two questions therein. As anticipated, omission under no penalties may arise due to implicit costs like pressure or stereotypes. These would be constant between the two parts of the test. Hence they should wash out in the before-after comparison.

Table 1: Omissions and the Hit-Rate

	Hit-Rate Differential		
	(Full Sample)	(High Ability)	(Low Ability)
Omit-Rate Differential	0.178***	0.112***	0.167***
	(0.018)	(0.020)	(0.018)
Gender	0.005	0.016*	0.003
	(0.009)	(0.009)	(0.009)
High Ability	−0.053***		
	(0.011)		
Observations	9,797	4,898	4,899
R^2	0.242	0.220	0.183
Adjusted R^2	0.240	0.218	0.181
Residual Std. Error	0.150 (df = 9776)	0.152 (df = 4881)	0.148 (df = 4882)
F Statistic	155.895*** (df = 20; 9776)	86.118*** (df = 16; 4881)	68.551*** (df = 16; 4882)

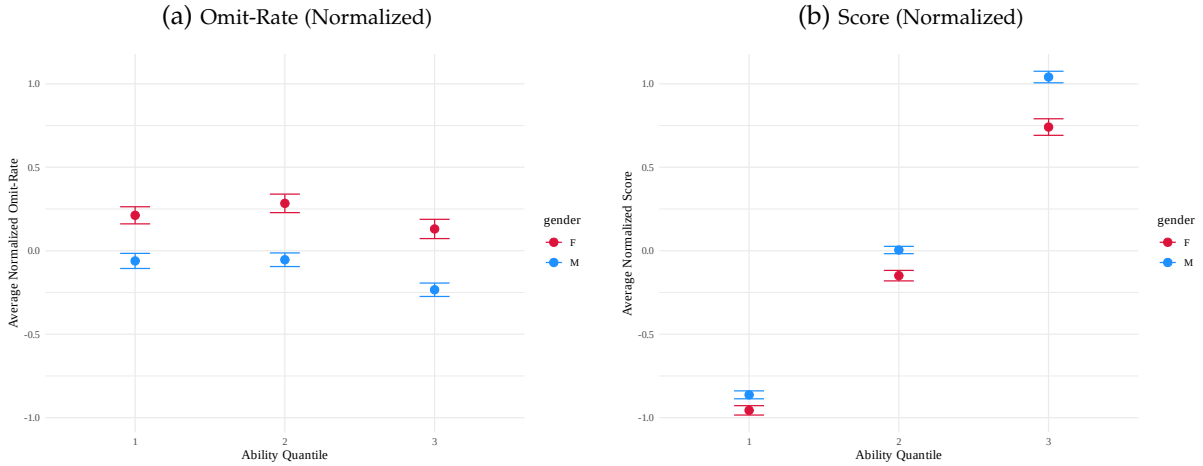
Note: interactions, fixed effects and constant are omitted. *p<0.1; **p<0.05; ***p<0.01

Elasticity of Selection Quality to Omission To complete the picture, I perform a regression analysis that is informative about the average metacognitive accuracy in my population. The previous finding, suggests that this is close to null. Indeed, I find a significantly small elasticity of the quality of selection to its extensive margin. To this end, I run a regression of the hit-rate differential on the omit-rate differential interacted with dummies for gender and high ability (defined as above median normalized hit-rate in the first part of the test) including test-level and participation time fixed effects. Table 1 reports the results. A difference in omissions corresponding to half of the questions in the negative marking test, leads to a 0.09 improvement in the hit-rate differential, corresponding to just one half standard deviation of the outcome. High ability has a mechanical negative effect, by the definition of hit-rate differential. Hence, I also run separate regressions for the high ability and the low ability group. I find very similar results, suggesting that metacognitive accuracy and ability are not strongly correlated.[20] Interestingly, while gender has no significant effect in the full sample, selection quality seem to be slightly better for males among the high ability students. This is coherent with the next subsection, showing that, the difference between males and females outcomes tends to be stronger in the high ability groups.

Robustness Checks In Appendix C, I show that the results are robust to (i) focusing on both genders separately (ii) restricting to students who participated in the math test more than once (hence who should not have problems understanding negative marking) (iii) dropping students who omit questions under no negative marking (hence choosing a dominated strategy). Moreover, while an analysis of the relation between metacognition and external outcomes is beyond the scope of this paper, I show that better metacognition is associated with higher math grades in school.

[20]This is coherent with psychological evidence, dating back to Lichtenstein and Fischhoff (1977), reinforced by the fact that metacognition has a distinct neural basis distinct from that of task performance (Fleming and Dolan, 2012).

Figure 2: Gender Gaps in Omissions and Scores vs. Ability



Notes: Panel (a) and (b) show respectively, the average normalized omitted rate and the average normalized score under negative marking across tertiles of the normalized hit-rate under no negative marking, separately for both genders. Variables are normalized at the (year, age group) level, which identifies tests.

Taken together, my findings falsify the predictions of a model based on perfect metacognition (see Appendix A). Instead, they are compatible with a scenario where metacognitive accuracy is heterogeneous across students, and null on average.

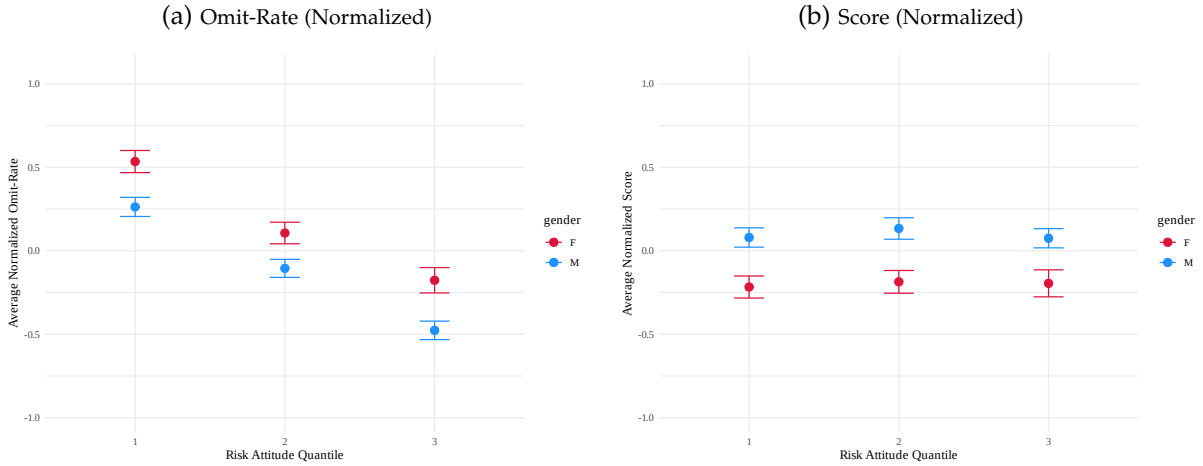
Gender Gaps in Score and Omissions

The literature on multiple-choice tests recurrently documented that penalties open a gender gap in scores and omissions that widens in students' ability. This correlation is puzzling to explain via preference-based explanations. The second part of my analysis employs my model to rationalize this pattern, based on the complementarity between ability and metacognition which naturally arises in my model.

Gender Gaps under Negative Marking In Figure 2, I plot the gap in average normalized omit-rate (2a) and in normalized scores (2b) under negative marking for each tertile of ability (i.e. normalized no negative marking hit-rate) separately for both genders. Coherently with the original paper (Iriberri and Rey-Biel, 2021) I show that under negative marking, males have higher normalized score and omit less questions than females, across all ability tertiles. Crucially, both gaps are sizably stronger for higher ability students. Specifically, moving from the bottom to the top ability tertile, the gap in scores triplicates (from 0.09 to 0.31 standard deviations of the mean) while the gap in omissions grows by an half (from -0.25 to -0.36 standard deviations of the mean).

Risk Aversion? Explaining these patterns in a model based only on differences in ability and risk aversion is difficult. Indeed, as explained in Appendix A, to rationalize the patterns maintaining the perfect metacognition assumption, one should assume that gender differences in risk preferences are positively correlated with ability. I do not observe such correlation in my

Figure 3: Gender Gaps in Omissions and Score vs. Risk Aversion



Notes: Panel (a) and (b) show respectively, the average normalized omitted rate and the average normalized score under negative marking across tertiles of risk aversion (on a decreasing scale), separately for both genders. Variables are normalized at the (year, age group) level, which identifies tests.

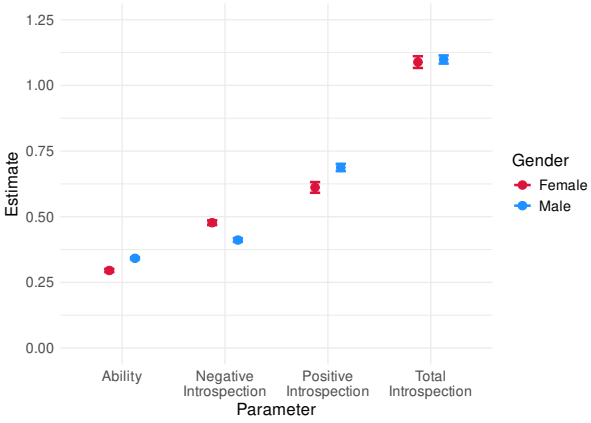
dataset: for both genders, the correlation between my measures of ability and risk aversion is significantly null (-0.04 for males and -0.06 for females). Moreover, Figure 3 shows that, while, risk aversion is positively correlated with omissions (as expected) it does not impact scores, nor for males or females. On top of this, within each risk aversion tertile, females still omit more than males.

Hypothesis: Gender Differences in Metacognition Leveraging the complementarity between ability and metacognition predicted by my model, I propose to rationalize Figures 2 and 3 positing that while comparable in their overall metacognitive accuracy, female metacognition is more slanted towards the negative side while male metacognition to the positive one. To fix ideas, consider a representative male and a representative female who differ only in metacognition. Given Equations 5 and 7, a back-of-the-envelope computation shows that we can write gaps in omissions and score between the two individuals as

$$\begin{cases} \delta\overline{O} & = \delta\psi_- - \overline{\pi}(\delta\psi_+ + \delta\psi_-), \\ \delta\overline{W} & = \delta\psi_-[\Delta^o - \underline{p}\Delta^h] + \overline{\pi}\left[\delta\psi_+(\overline{p}\Delta^h - \Delta^o) - \delta\psi_-(\Delta^o - \underline{p}\Delta^h)\right], \end{cases}$$

where $\overline{\pi}, \underline{p}, \overline{p}$ are common ability parameters and $\delta\psi_+$ ($\delta\psi_+$) is the difference in positive (negative) metacognition. Then, the gap in Figure 2a (and 3a) suggests that females are on average more negative metacognitive than males, $\delta\psi_- < 0$, while its widening pattern suggests $\delta\psi_+ >_\approx -\delta\psi_- > 0$ so that males are on average more positive metacognitive than females, and, on top of that, have slightly higher total metacognition. This assumption yields the widening size of the gap in scores (Figure 2b), when $\underline{p}, \overline{p}$ are spread enough.

Figure 4: Maximum Likelihood Estimates for the Binary Model.



Notes: Estimates for $\underline{p} = 0.2$ and $\overline{p} = 1$. Confidence intervals are built using the delta method.

Maximum Likelihood Estimation To make these points quantitatively, I provide a maximum likelihood estimation of my model (see Appendix C for details). I pool data by gender, as if they were generated by the two representative individuals, a male and a female. To identify my system, I assume that both have the same values of $\underline{p}$ and $\overline{p}$. For simplicity, I assume $\underline{p} = 1/5$ (corresponding to randomly guessing an answer) and $\overline{p} = 1$ (corresponding to perfectly knowing an answer).[21] The results, shown in Figure 4, are by and large coherent with the intuitive picture above. In particular, while the representative male is slightly more able and metacognitive than the representative female, the imbalance between positive and negative metacognition is as expected. Moreover, note that total metacognition is estimated to be close to one for both genders, implying a null hit-rate differential, coherently with Figure 1b.

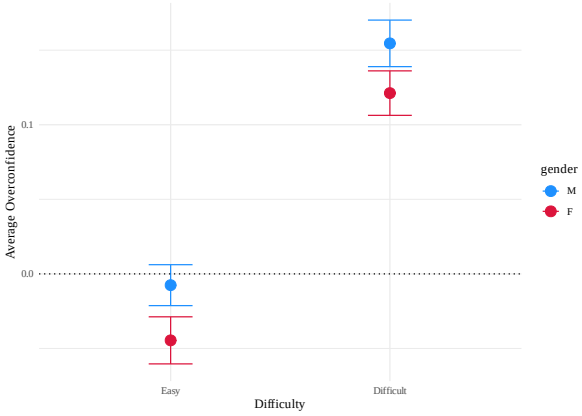
Evidence from Bordalo et al. (2019) To further support my hypothesis, I provide evidence about gender-based heterogeneity in the relation between confidence and performance, exploiting the data of Bordalo et al. (2019). In their experiment, participants answer to a 40 item questionnaire (without skipping) and report confidence question-by-question. Then we can directly analyze confidence reports by gender and and task difficulty. Say that a question is difficult/low-performance (easy/high-performance) if less (more) than half of the sample got it correct. Focusing on male-typed topics (including math) Figure 5a shows that male students are on average calibrated on easy task but overconfident on difficult ones, while female students are underconfident on easy task but less overconfident than males on difficult ones. The comparison of averages, while informative, may be masking interesting heterogeneity. Figure 5b shows that there is a mass of female students which are especially underconfident on easy tasks. This provide suggestive evidence that female (male) confidence tracks

[21] In Appendix C, I run robustness checks with other choices of $\underline{p}, \overline{p}$.

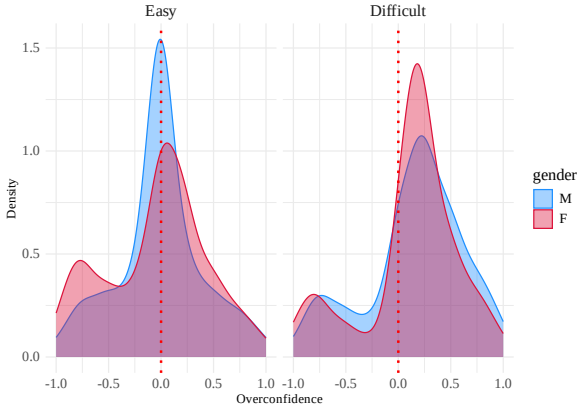
performance relatively better on difficult (easy) tasks.[22]

Figure 5: Gender Differences in Metacognition

(a) Average Overconfidence by Difficulty and Gender



(b) Overconfidence Distribution by Difficulty and Gender



Notes: Panel (a) and (b) show the average overconfidence and the density of the overconfidence distribution, separately for both genders in difficult vs easy tasks. Difficult (easy) tasks are those where at least (most) half of the sample fails.

5 The Hit-Rate Differential

In this Section, I generalize the study of the hit-rate differential to arbitrary skill-types.

First, let's point out that two arguments developed for binary skill types are actually completely general. In particular, (i) whenever there are no omissions, the hit-rate corresponds to the average performance level implied by the ability distribution; (ii) whenever there are no penalties for errors, opting out tasks is dominated, hence there are no omissions. The following is an immediate consequence of Proposition 2.

Corollary 1 *Consider a DM with skill-type τ. It holds*

$$\overline{O} = 0 \Longrightarrow \overline{H} = \mathbb{E}_\pi[p] \;\; \textit{and} \;\; \overline{W} = m + \Delta^h \mathbb{E}_\pi[p].$$

Moreover no penalties, $\Delta^o = 0$, imply no omissions, $\overline{O} = 0$.

Hence, the fact that performance under no penalties elicits ability is completely general. Then, we can define the hit-rate differential as the difference in the hit-rate between an environment with and without penalties, $\Delta^o > 0$,

$$\overline{D}(\tau, \alpha, \Delta) = \mathbb{E}_\tau[p | c > c^*] - \mathbb{E}_\pi[p]. \qquad \text{(Hit-Rate Differential)}$$

Note that, unlike the binary case (Equation 4) in general this quantity depends on how large penalties are and on risk aversion, because these parameters affect the threshold level c^*.

[22] This finding can be restated as documenting gender-based heterogeneity in the hard-easy effect (Moore and Healy, 2008, for reference).

To generalize the insight that (the sign of) the hit-rate differential can be taken as the behavioral equivalent of metacognitive accuracy, the first step is to define an appropriate nonparametric notion of accuracy. Indeed, outside low cardinality cases, it is unclear how to formulate a condition expressing the rate of "ordinal matching" between confidence and performance, like that used in Section 3; moreover, such a condition may become too requiring as the number of performance levels and confidence judgments grows.

Weak Positive Dependence In statistics, a widespread notion expressing that two variables "move together" is *positive quadrant dependence* (Lehmann, 1966). This says that, given any pair of quantiles, the probability of the two variables being simultaneously above (or simultaneously below) such quantiles is at least as large as it would be if they were independent. The notion captures the idea that confidence tracks performance "better-than-random"; for instance it is implied when the joint distribution of the variables dominates the independent coupling of their marginals in the supermodular order (Meyer and Strulovici, 2012).[23] As shown by Esary et al. (1967), positive quadrant dependence is equivalent to requiring non-negative covariance between any non-decreasing pair of transformations of the random variables. I introduce a slightly weaker version of this notion, stating that a skill-type τ has *weak positive dependence* if, given any non-decreasing affine $f : [0,1] \to \mathbb{R}$ and any non-decreasing $g : [0,1] \to \mathbb{R}$ it holds

$$\text{Cov}_\tau[f(P), g(C)] \geq 0. \qquad \text{(Weak Positive Dependence)}$$

My main result shows that Weak Positive Dependence is equivalent to the non-negativity of the hit-rate differential, for any decision environment and any risk aversion coefficient.

Proposition 3 *Consider a DM with skill-type τ, then it holds*

$$\tau \text{ has Weak Positive Dependence} \iff \forall \Delta, \alpha \quad \overline{D}(\tau, \alpha, \Delta) \geq 0. \qquad \text{(Hit-Rate Differential Test)}$$

Intuitively, compared to positive quadrant dependence, this weaker notion allows to set quantile thresholds only on the confidence variable, but not on performance. Nonetheless, this is sufficient for the non-negative hit-rate condition: just choose the step function of confidence defined by the confidence threshold implied by payoffs and risk aversion. The converse direction, which is less obvious, follows from Hoeffding's covariance identity.

Centrifugal Spreads The second intuition suggested by Equation 4 is that, fixed metacognition, the magnitude of the hit-rate differential grows when better performance levels are further away from worse ones. To generalize this notion, I need three ingredients. First, I

[23] A coupling of two distributions is any distribution having them as marginals (Ambrosio et al., 2021). Recall that, fixed two distributions $\mu, \nu \in \Delta([0,1])$ and given two couplings $\lambda, \lambda' \in \Gamma(\mu, \nu)$, the supermodular order is defined $\lambda' \succeq_{cor} \lambda$ if $\mathbb{E}_{\lambda'}[f] \geq \mathbb{E}_\lambda[f]$ for any supermodular $f : [0,1]^2 \to \mathbb{R}$. For bivariate distributions, the order is equivalent to many other notions of interdependence. To see the implication, observe that since fg is supermodular, $\tau \succeq_{cor} \pi \otimes \gamma$ implies $E_\tau[fg] \geq E_\pi[f]E_\gamma[g]$, hence Weak Positive Dependence.

define a stochastic order which captures the idea that a distribution becomes "more bimodal" than another, while keeping its mean constant. Say that π' is a *centrifugal spread*[24] of π , $\pi' \succeq_{CS} \pi$, if

$$\exists\, \rho > 1 \quad s.t. \quad \pi' = l_{\#}(\pi) \quad for \quad l(p) = E_{\pi}[p] + \rho(p - E_{\pi}[p]). \qquad \text{(Centrifugal Spread)}$$

The idea beyond the order is that mass is pushed away from the mean to the extreme in a centrifugal fashion, just as it happens for the surface of a fluid in a mixer. The parameter ρ captures intensity. The fact that centrifugal spreads preserve the mean is a consequence of the linearity of expectation. Second, I define a strengthening of Weak Positive Dependence. Say that metacognition is *increasing* if

$$p' \geq p \Longrightarrow \psi(p') \succeq_{FOSD} \psi(p).$$

Hence, to higher performance levels, correspond greater confidence distributions, in the stochastic sense.[25] In the terminology of Lehmann (1966), this condition says that confidence is positively-regression-dependent on performance.[26] Third, I need to express that two DMs have the same metacognition even when their ability distributions differ cardinally. To compare skill types with different marginals, I say that τ, τ' are *metacognition-equivalent* if

$$\exists\, f : P \to P' \quad \text{increasing bijection} \implies \psi' \circ f = \psi$$

This notion generalizes the fact that, in the binary example, metacognition is essentially an ordinal concept.

Proposition 4 *Consider two metacognition-equivalent DMs with risk aversion coefficient $\alpha \geq 0$ and positive metacognition. It holds*

$$\pi' \succeq_{CS} \pi \Longrightarrow \forall \Delta \quad \overline{D}(\tau', \alpha, \Delta) \geq \overline{D}(\tau, \alpha, \Delta).$$

Note that a mirror version of the results presented in this Section holds considering notions opposite to those of Weak Positive Dependence and increasing metacognition.

6 Ability-Metacognition Complementarity

In this Section, I generalize the study of the hit-rate differential to arbitrary skill-types.

[24] In the context of common value auctions, Cusumano and Pieroth (2024) independently introduce the notion of *mean-preserving k-scaling* to study when due diligence is optimal, depending on the size of the stakes. A centrifugal spread is a mean preserving scaling where $k > 1$.

[25] Recall that given $\lambda, \lambda' \in [0,1]$, $\lambda' \succeq_{FOSD} \lambda$ iff $\forall x \in [0,1] F_{\lambda'}(x) \leq F_{\lambda}(x)$.

[26] To see why increasing metacognition implies WPD, observe that $\overline{D} = Cov(P, \overline{F}_{\psi}(c^*|P))/\mathbb{E}[\overline{F}(c^*|P)]$, where $\overline{F}_{\psi}$ is the system of survival functions associated to ψ. Increasing metacognition implies that given any c^*, $\overline{F}(c^*|P)$ is increasing. Then, $\overline{D}$ is positive as its numerator is, being the covariance between a random variable and a monotone transformation.

In the Binary Model the clarity of the comparison expressed by Equation 6 derives from a well-defined notion of true positives and true negatives for the metacognition map. In the general case, there is no well-established way to compare distributions based on their interdependence strength of for high relative to low values.[27] Nonetheless, the main intuition is general.

Metacognition Shifts The key observation is that, in the binary case, when a decision maker is more positive metacognitive but less negative metacognitive than another – so that there is no dominance between the two, in terms of accuracy – the confidence distributions of the first DM first-order stochastic dominate that of the second DM at any performance level. I leverage this feature of binary type comparisons to generalize the idea that one DM's metacognition is more slanted towards the positive/negative side compared to another. Given two skill-types supported on the same set of performance levels I say that ψ' is a *positive shift* of ψ if

$$\forall c \quad \delta(c|p) \equiv \overline{F}_{\psi'(p)}(c) - \overline{F}_{\psi(p)}(c) \geq 0 \text{ and increasing in } p. \qquad \text{(Metacognition Shift)}$$

Equivalently, I could say that ψ is a negative shift of ψ'.

Proposition 5 *Fix a decision environment Δ. Consider two DMs with ability π and risk aversion coefficient $\alpha \geq 0$. Define the welfare gap $\delta\overline{W}(\psi, \psi'|\pi, \alpha) = \overline{W}(\tau', \alpha, \Delta) - \overline{W}(\tau, \alpha, \Delta)$. Then, it holds*

$$\delta\overline{W}(\psi, \psi'|\pi, \alpha) = \mathbb{E}_\pi \left[(p\Delta^h - \Delta^o)\delta(c^*|p) \right]. \qquad \text{(AAW Difference)}$$

Hence, if ψ' is a positive shift of ψ, we have the inequality

$$\delta\overline{W}(\psi, \psi'|\pi, \alpha) \geq -(\Delta^h \mathbb{E}_\pi[p] - \Delta^o)\delta\overline{O}(\psi, \psi'|\pi, \alpha) \qquad \text{(Welfare Gap Inequality)}$$

where $\delta\overline{O}(\psi, \psi'|\pi, \alpha) = \overline{O}(\tau', \alpha, \Delta) - \overline{O}(\tau, \alpha, \Delta)$, the omission gap, is negative and decreasing in FOSD-improvements of π. Moreover, the welfare gap is eventually increasing

$$\exists \tilde{\pi} \quad \pi' \succeq^{FOSD} \pi \succeq^{FOSD} \tilde{\pi} \Longrightarrow \delta\overline{W}(\psi, \psi'|\pi', \alpha) \geq \delta\overline{W}(\psi, \psi'|\pi, \alpha)$$

The AAW Difference expresses compactly how the impact on welfare of a shift in metacognition, ceteris paribus, depends on the expectation of the product of two terms. The first, $p\Delta^h - \Delta^o$ captures the expected payoff between performing vs. opting out a task as a function of the performance level, p. The second is the probability of actually performing a task of level p. In turn, Welfare Gap Inequality gap follows since these two terms, under Metacognition Shift, are both an increasing transformation of the random performance. Hence, they are affiliated (Milgrom and Weber, 1982) implying a positive covariance. Just as in Section 3, when DMs are sufficiently able, $\mathbb{E}_\pi[p] \geq \Delta^o/\Delta^h$, this bound grows in FOSD

[27] In the copulas literature (Embrechts et al., 2001) tail dependence is the closest notion. On top of being a limit concept, here its application is not fruitful.

improvements in ability. This captures the main force at play, which "eventually" ensures monotonicity of the gap. Again, flipping signs one obtains a statement in terms of negative shifts.

Maximal Types Another immediate implication of AAW Difference is that some overconfidence helps balancing out the excessive conservatism imposed by risk aversion. This aspect was obfuscated by the extreme nature of the binary model, where two confidence judgments do not give sufficient nuance to reveal it. To let it emerge, I characterize maximal skill-types, fixed ability and risk aversion.

Proposition 6 *Among DMs with ability π and risk aversion coefficient $\alpha \geq 0$, welfare $\overline{W}$ is maximized iff*

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

As a consequence, unless $\alpha = 0$, perfect metacognition, $\psi^* = id$, is not maximal.

To see that maximal types involve overconfidence, observe that risk aversion imposes $c^* > p^*$. At the same time note that, among the risk neutrals, perfect metacognition is maximal for any ability and any decision environment – indeed, under perfect metacognition the DM's Problem maximization is against the same distribution Nature draws from. This suggests that the value of perfect metacognition is in its "robustness", but that this feature is not dominant, in general.

7 An Information Processing Foundation

Having concluded the analysis of my model and its implications, in this Section I propose an information processing rationale for skill-types.

Tasks as Decision Problems I enrich the setup described in Section 1, modeling each task $i \in [N]$ as an instance of a decision problem. In particular, I consider a finite set of states of the world $\Omega = [M] \equiv \{1, \ldots, M\}$, a set of actions $\mathcal{A} = [M] \cup \{0\}$, an outcome function

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

and a payoff function defined as $u(a, \omega) \equiv u_\alpha(\zeta(a, \omega))$, where u_α is as in Equation 1. The payoff describes a (generalized) hypothesis testing problem, enriched with an outside option, for a potentially risk averse agent. Note the ex-post of optimal action in each task is to match

the state of the world, $a^*(\omega) = \omega$. Hence, the state of the world can be interpreted as the ex-post optimal action itself. Actions include an outside option, 0, which is never ex-post optimal. The state of the world is drawn independently across tasks from a full-support distribution $d_0 \in \Delta(\Omega)$, which describes the statistical frequency with which each action is ex-post optimal. The decision maker takes d_0 as a prior about ω.

Cognitive Signals In each task, before taking action, the decision maker observes a cognitive signal about the state of the world, s_ω, from a finite set, S_Ω. The signal is extracted from a Blackwell experiment whose informativeness is parameterized by a set of *informativeness types*, $\Theta \ni \theta$. Formally, I consider a map

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

The informativeness type is task-specific, capturing the idea that the quality of the DM's "thinking" is heterogeneous across tasks. In particular, the type realization is drawn independently from a full-support distribution $t_0 \in \Delta(\Theta)$. I assume that the identity of the ex-post optimal action and the quality of the DM's thinking in a task are not related.

Assumption 1 *The state of the world and the informativeness type are independent across tasks.*

The assumption allows tractability, and is natural in all those context where actions are "labels". For instance, in multiple-choice tests, the numbering of answers can be shuffled within a question, effectively breaking any statistical link between the action corresponding to the correct answer and the quality of a student's knowledge about the problem. Denote by $d_0 \otimes t_0 \in \Delta(\Omega \times \Theta)$ the joint distribution of states and types.

Information Processing The decision maker does not observe the task-specific realization of θ. Hence they are uncertain about the experiment which generated the observed signal. To model imperfections in the DM's (metacognitive) accuracy about the noise in their (cognitive) signals, I posit that the DM Bayes-updates beliefs about θ. In particular, I consider a meta-experiment

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

that produces a signal s_θ from a finite set S_Θ, conveying information about the experiment's type. Hence, σ and μ, together with $d_0 \otimes t_0$ endow the set $\Omega \times \Theta \times S_\Omega \times S_\Theta$ with the following probability measure,

$$\mathbb{P}(\omega, \theta, s_\omega, s_\theta) = d_0(\omega) t_0(\theta) \sigma(s_\omega | \omega, \theta) \mu(s_\theta | \theta). \qquad \text{(Joint Distribution)}$$

I denote by $d^{s_\theta}_{s_\omega} \in \Delta(\Omega)$ the posterior about ω after observing (s_ω, s_θ). To sum up, the model proceeds as follows. For each task $i \in [N]$,

1. Nature extracts (ω^i, θ^i) from $d_0 \otimes t_0$.
2. Nature extracts s^i_ω from $\sigma(\omega^i, \theta^i)$ and s^i_θ from $\mu(\theta^i)$.
3. The decision maker observes (s^i_ω, s^i_θ) and updates beliefs about actions to $d^{s^i_\theta}_{s^i_\omega}$.

I call the task-specific realization $(\theta^i, s^i_\omega, s^i_\theta)$ the *sort* of task i. Note that only (s^i_ω, s^i_θ) is part of the DM's information set in the task. Moreover, observe that, given the finiteness of $\Theta \times S_\Omega \times S_\Theta$ each task sort in the support of $\mathbb{P}$ is faced infinitely often by the DM.

Performance and Confidence This setting allows for a Bayesian definition of performance levels and confidence judgments. To this end, given beliefs $d \in \Delta(\Omega)$ define the *forced-optimal action* as

$$a^\circ(d) = \arg\max_{\mathcal{A}\setminus\{0\}} \mathbb{E}_d[u(a, \omega)].$$

This is the action the decision maker would take, were the outside option made unavailable. Given the payoff structure, it is immediate to observe that $a^\circ(d) = mode(d)$. Moreover, note that the posterior optimal action writes $\hat{a}(d) = a^\circ(d)\mathbf{1}\{d(a^\circ(d)) > c^*\}$, where c^* has the same expression as in Decision Rule. Then, we can define *success* as the event where the forced-optimal action and the ex-post optimal action coincide,

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

Observe that E is a cylinder set in Θ, since neither the ex-post optimal action or the forced-optimal one depend on the experiment's informativeness type. In turn, the confidence judgment in a task-sort $(\theta, s_\omega, s_\theta)$ can be defined as the subjective belief the the DM entertains about E,[28]

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

Instead, the performance level in a task-sort $(\theta, s_\omega, s_\theta)$ can be defined as the limit frequency of success in tasks of that sort

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

Contrary to confidence judgments, performance levels do depend on the informativeness type, which determines the frequency with which a sort belongs to the success event. Then, it is straightforward to associate to each instance of the information processing model a skill-type, pushing forward the joint distribution of information sets and informativeness types to one on confidence and performance. Denote by $d^\theta_{s_\omega} = \mathrm{BU}(d_0, \sigma_\theta, s_\omega)$ the posterior of a

[28] Obviously, confidence judgments are independent of the latent experiment's informativeness types, conditional on the DM's information set. Still, regarding confidence as a function of the full sort facilitates exposition.

decision maker accessing the perfectly informative meta-experiment $\mu^* : \theta \mapsto \chi_\theta$.

Proposition 7 *Consider an instance (d_0, t_0, σ, μ) of the information processing model. Fix any task-sort $(\theta, s_\omega, s_\theta)$. The performance level and the confidence judgment are, respectively*

$$p(\theta, s_\omega, s_\theta) = d^{\theta}_{s_\omega}(mode(d^{s_\theta}_{s_\omega})) \qquad c(\theta, s_\omega, s_\theta) = d^{s_\theta}_{s_\omega}(mode(d^{s_\theta}_{s_\omega}))$$

In turn, the DM skill-type is the pushforward

$$\tau(d_0, t_0, \sigma, \mu) = (p, c)_\#[\mathbb{P}(\theta, s_\omega, s_\theta)].$$

The proof follows from the form of a^* and a°, applying the Law of Large Numbers.

Symmetric Environments While Proposition 7 has a conceptual value, it may be abstract for modeling purposes. To make things more concrete, I restrict to symmetric environments, where d_0 is the discrete uniform prior. These are settings where there is no default action. To have a concrete example, think again about multiple choice tests, where the labels associated to correct answers can be determined through a randomization procedure. In these environments, it is convenient to work with a class of experiments which generalizes symmetric binary experiments. Fixed $\theta \in [0, 1]$ say that an experiment $\sigma(\cdot|\cdot)$ is *symmetric diagonally dominant* of informativeness θ if $S_\Omega = \Omega = [M]$ and

$$\forall i \in [M] \quad \theta = \sigma(i|i) \geq \max_{i \neq j} \sigma(i|j).$$

Let $m(t_0, \mu)$ be the stochastic map which associates to each informativeness type θ the *conditional distribution of posterior means* induced by the Bayesian updating of t_0 through μ. Then we have the following result.

Proposition 8 *Assume that d_0 is uniform.*

1. *If σ is a symmetric diagonally dominant family. Then, the DM's skill type $\tau(d_0, t_0, \sigma, \mu)$ is such that t_0 is the DM's ability distribution, while $m(t_0, \mu)$ is the DM's metacognition map. In particular, τ is a martingale coupling.*

2. *Conversely, any skill-type τ which is a martingale coupling is implemented by an information model with symmetric diagonally dominant σ.*

The result follows since, given the independence of d_0 and t_0, we can perform Bayesian updating hierarchically, first updating beliefs about the experiment's informativeness through the meta-experiment μ and then updating beliefs about optimal actions according to the implied expected experiment. The intuition is that, while both signals s_ω and s_θ carry information on the informativeness type θ, when the objective is to update beliefs on ω, the interim-expected

experiment is a "sufficient statistics". This fact, combined with properties of symmetric diagonally dominant experiments yields the result.[29]

The main upshot of Proposition 8 is that, in symmetric environments, first and second order information separately control ability and metacognitive accuracy. Note that, in this case, since the skill type is a martingale coupling, the confidence distribution is a mean preserving contraction of the performance one, hence we get a form of overinference from weak signals and underinference from strong ones. Nonetheless, the details of the meta-experiment determine which dimension of error prevails. For instance, false positives and false negatives in the meta-experiment give rise to asymmetries between positive and negative metacognition in binary skill-types.

8 Conclusion

In this paper, I proposed a theory of decision making centered on the idea that agents differ not only in cognitive skill, but also in the skill of monitoring their own cognitive performance – metacognition. First, I showed that efficient task selection is equivalent to good metacognition, defined as a better-than-random matching between confidence and performance; moreover I demonstrated that metacognition matters more, the more bimodal ability is. Second, I showed that a complementarity between ability and metacognition exists, so that monitoring high (low) performance tasks is relatively more valuable for those who are relatively more (less) able. Third, I supported both the premises and the predictions of my theory analyzing data on multiple choice tests, shedding light on puzzling gender gaps recurringly observed in that context. Finally, I proposed a foundation for my theory, based on second-order information processing.

I foresee numerous avenues for studying the economic implications of imperfect metacognition, in future research. First, my framework could be applied with little or no modifications to organizational economics models where monitoring task specific ability is of central importance. For instance, consider knowledge hierarchies (Garicano, 2000), where passing on tasks involves recognizing ignorance, or the trading tasks model (Deming, 2017) where efficiency derives from the mutual recognition of relative comparative advantage between workers. Moreover, the information processing foundation may be useful to explain expert's under-reliance on AI signals in human-AI interaction (Agarwal et al., 2023). Second, it may be interesting to understand how metacognition adapts to different environments in a (reinforcement) learning model. For instance, differences in the intensity of punishment for error vs. inaction may motivate differences in positive vs. negative metacognition. In this sense, research documenting different (reputational) incentives faced by genders (Sarsons, 2017) is a particularly interesting motivation. Moreover, it may be valuable to explore how motivated

[29] In particular, I exploit that for symmetric diagonally dominant families, the expected experiment is still symmetric diagonally dominant, with precision given by the expected informativeness type. Moreover, given a uniform prior, the posterior mode after any signal from a symmetric diagonally dominant experiment corresponds to the signal itself, and it has mass given by the informativeness parameter.

memory (Amelio and Zimmermann, 2023) may prevent recording failures and hence forming negative metacognition. Finally, a more ambitious step would be to endogenize the precision of cognitive signal, in the second-order information processing model. This would amount to a theory of "productive" rational inattention where an agent pays an effort cost to produce a signal, but lacks metacognition and hence is uncertain about its quality. This may answer the conceptual challenge of incorporating miscalibration in rational inattention models, as posed by Caplin (2023).

Bibliography

Agarwal, N., A. Moehring, P. Rajpurkar, and T. Salz (2023). Combining human expertise with artificial intelligence: Experimental evidence from radiology. Technical report, National Bureau of Economic Research.

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.

Ambrosio, L., E. Brué, and D. Semola (2021). *Lectures on optimal transport*. Springer.

Amelio, A. and F. Zimmermann (2023). Motivated memory in economicsa review. *Games 14*(1), 15.

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

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

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

Battigalli, P. and G. Bonanno (1999). Recent results on belief, knowledge and the epistemic foundations of game theory. *Research in Economics 53*(2), 149–225.

Belardinelli, G. and B. C. Schipper (2024). Implicit knowledge in unawareness structures. *Synthese 204*(5), 141.

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.

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

Caplin, A. (2023). *Science Of Mistakes, The: Lecture Notes On Economic Data Engineering*, Volume 16. World Scientific.

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

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

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

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

Currie, J., W. B. MacLeod, and K. Musen (2024). First do no harm? doctor decision making and patient outcomes. Technical report, National Bureau of Economic Research.

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

de Castro, L. I. (2009). Affiliation and dependence in economic models. Technical report, Discussion paper.

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. 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 (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.

Edin, P.-A., P. Fredriksson, M. Nybom, and B. Öckert (2022). The rising return to noncognitive skill. *American Economic Journal: Applied Economics 14*(2), 78–100.

Embrechts, P., F. Lindskog, and A. McNeil (2001). Modelling dependence with copulas. *Rapport technique, Département de mathématiques, Institut Fédéral de Technologie de Zurich, Zurich 14*, 1–50.

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.

Enke, B., T. Graeber, R. Oprea, and J. Yang (2024). Behavioral attenuation. Technical report, National Bureau of Economic Research.

Esary, J. D., F. Proschan, and D. W. Walkup (1967). Association of random variables, with applications. *The Annals of Mathematical Statistics 38*(5), 1466–1474.

Espinosa, M. P. and J. Gardeazabal (2010). Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical psychology 54*(5), 415–425.

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

Fleming, S. M. and R. J. Dolan (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences 367*(1594), 1338–1349.

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

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.

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

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

Ilut, C. and R. Valchev (2023). Economic agents as imperfect problem solvers. *The Quarterly Journal of Economics 138*(1), 313–362.

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.

Koriat, A. (2012). When are two heads better than one and why? *Science 336*(6079), 360–362.

Lehmann, E. L. (1966). Some concepts of dependence. *The Annals of Mathematical Statistics 37*(5), 1137–1153.

Lichtenstein, S. and B. Fischhoff (1977). Do those who know more also know more about how much they know? *Organizational behavior and human performance 20*(2), 159–183.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Meyer, M. and B. Strulovici (2012). Increasing interdependence of multivariate distributions. *Journal of Economic Theory 147*(4), 1460–1489.

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.

Pouget, A., J. Drugowitsch, and A. Kepecs (2016). Confidence and certainty: distinct probabilistic quantities for different goals. *Nature neuroscience 19*(3), 366–374.

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.

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

Sarsons, H. (2017). Interpreting signals in the labor market: evidence from medical referrals. Technical report.

Stanovich, K. E. (1999). *Who is rational?: Studies of individual differences in reasoning*. Psychology Press.

Strassen, V. (1965). The existence of probability measures with given marginals. *The Annals of Mathematical Statistics 36*(2), 423–439.

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

Vaeth, M. (2024). Imprecision attenuates updating.

Van Benthem, J. (2010). *Modal logic for open minds*. Number 199. Center for the Study of Language and Information Stanford.

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

A The Perfect metacognition Benchmark

In this Appendix Section, I briefly go through the prediction of my model in the case decision makers could perfectly forecast their performance. This is useful mostly to reject such predictions, corroborating the main analysis. In particular, I show that (i) binary comparisons among a population of perfectly metacognitive DMs should display a rigid structure (ii) under perfect metacognition risk aversion should strongly explain variation in scores. These points follow from the comparative statics of the perfect metacognition model, summarised in the following result.[30]

Proposition A9 *Fix any decision environment with penalties, $\Delta^o > 0$. Consider a population of DMs with perfect metacognition, $\psi^* : p \mapsto \delta_p$.*

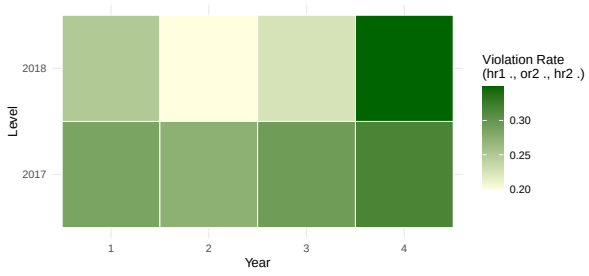
(a) For each DM the omit-rate is $\overline{O} = F_\pi(c^*)$ and the hit-rate is $\overline{H} = \mathbb{E}_\pi[p|p \geq c^*]$.

(b) Fix two DMs with the same risk aversion coefficient α. If $\pi' \succeq_{FOSD} \pi$, then $\overline{O}' \leq \overline{O}$ and $\overline{W}' \geq \overline{W}$. Moreover, if $\pi' \succeq_{HR} \pi$, also $\overline{H}' \geq \overline{H}$.

(c) Fix two DMs with the same ability π. If $\alpha' \leq \alpha$, then $\overline{O}' \geq \overline{O}$, $\overline{H}' \geq \overline{H}$ and $\overline{W}' \leq \overline{W}$.

Comparison of Perfectly Metacognitive DMs The main interest of the result is in offering a between-individual test for perfect metacognition which corroborates the principal one. Indeed, consider a pair of DMs. Suppose that the first (i) does (considerably) better than the second under no penalties; (ii) omits (considerably) more than the second under penalties. Then, taken together Points (a) and (b) suggest that the first student should have an higher hit-rate then the second. In my sample, I show that a considerable share of student pairs fails this pattern. Figure A1 shows a heatmap representing the share of violating pair at the test level (i.e. at the year-age group level): between 20% and 35% of pairs violate the comparison, even biasing the analysis against the result (asking that that comparisons hold even up to a robustness threshold).

Figure A1: Violations of Comparison Patterns



Notes: The heatmap is obtained self-joining the dataset and imposing an identification condition for violations, in particular I consider $\overline{H}^{0'} > \overline{H}^0 + 0.05$, $\overline{O}' > \overline{O} + 0.05$, but $\overline{H}' > \overline{H} + 0.05$.

[30] Recall *hazard rate dominance* $\pi' \succeq^{HR} \pi$ iff for all $x \in [0,1]$, $H_{\pi'}(x) \geq H_\pi(x)$, where, H_π is the hazard rate function, defined as $\lim_{\delta \to 0^+} \frac{\mathbb{P}_\pi(x \leq X < x+\delta | X \geq x)}{\delta}$, where X is a variable of law π.

Evidence on Risk Aversion Given that risk and ability are essentially independent in my sample, one would expect that, since more risk averse students omit more questions, they should experience a drop in score, on average. Figure 3 shows that the premise of the argument is true: for both genders, the level of omissions increases across the tertile of risk aversion (0.75 standard deviations of the mean passing from the least to the most risk averse tertile). As an upshot, this serves as validation of the non standard measure of risk aversion employed by the authors. Nonetheless, it also shows that the consequent is false: for both genders, performance in the second part of the test seems essentially independent from risk attitudes. I summarize these points in the following

B Proofs

Proof of Proposition 1 Fix a vector $(c_i)_{i=1}^N$ of confidence realizations. Without loss arrange it in decreasing order. I show that $\forall i, j \in [N]$ if $c_i > c_j$ and $j \in A^*((c_i)_{i=1}^N; \alpha, \Delta)$ then $i \in A^*((c_i)_{i=1}^N; \alpha, \Delta)$. Suppose this is not the case. Consider the perturbation $A' = (A^* \setminus \{\hat{j}\}) \cup \{\hat{i}\}$. Observe that $|A'| = |A^*|$, but $\sum_{i \in A'} Z(c_i) \succ_{FOSD} \sum_{i \in A^*} Z(c_i)$. Hence $W^{subj}(A') \succ_{FOSD} W^{subj}(A^*)$. In turn, u_α strictly increasing $V(A') > V(A^*)$, which is a contradiction. Then, it follows that the optimal portfolio has an interval structure, $A^*((c_i)_{i=1}^N; \alpha, \Delta) = \{i \in N \mid c_i > c^*\}$. Let $j(c^*) = |A^*((c_i)_{i=1}^N; \alpha, \Delta)|$, exploiting the decreasing rearrangement of $(c_i)_{i=1}^N$ and the tie breaking assumption we can determine c^* imposing

$$V(\{i \in N \mid c_i > c^*\}) \geq V(\{i \in N \mid c_i > c^*\} \cup \{j(c^*) + 1\}) \tag{8}$$

To characterize c^*, I exploit a closed form solution for the value of the DM's Problem.

Lemma A1 *Fix any $A \subseteq [N]$ any $\alpha > 0$ and any $(c_i)_{i=1}^N$. Then it holds*

$$V((c_i)_{i=1}^N; \alpha, \Delta) = 1 - e^{-\alpha No} e^{-\alpha |A|(m-o)} \prod_{i \in A} \left(1 - c_i + c_i e^{-\alpha(h-m)}\right) \tag{9}$$

Proof of the Lemma Define $H_A = \sum_{i \in A} Z(c_i)$, $a = (h - m)$ and $b = No + |A|(m - o)$. Then it holds

$$V(A) = \mathbb{E}\left[1 - e^{-\alpha W^{subj}(A)}\right] = 1 - \mathbb{E}\left[e^{-\alpha(aH_A + b)}\right] = 1 - e^{-\alpha b}\mathbb{E}\left[e^{-\alpha a H_A}\right] = 1 - e^{-\alpha b} \prod_{i \in A} (1 - c_i + c_i e^{-\alpha a})$$

where the last step follows by the moment generating function for the Poisson Binomial distribution (or for the sum of independent Bernoullis). Then Equation 9 follows substituting back a and b. □

To proceed with the proof, it is Then, substitute Equation 9 in Equation 8 to obtain

$$\begin{aligned}
1 - e^{-\alpha No} e^{-\alpha j(m-o)} \prod_{i=1}^{j} \left(1 - c_i + c_i e^{-\alpha(h-m)}\right) \geq 1 - e^{-\alpha No} e^{-\alpha(j+1)(m-o)} \prod_{i=1}^{j+1} \left(1 - c_i + c_i e^{-\alpha(h-m)}\right) \Leftrightarrow \\
e^{-\alpha j(m-o)} \prod_{i=1}^{j} \left(1 - c_i + c_i e^{-\alpha(h-m)}\right) \leq e^{-\alpha(j+1)(m-o)} \prod_{i=1}^{j+1} \left(1 - c_i + c_i e^{-\alpha(h-m)}\right) \Leftrightarrow \\
1 \leq e^{-\alpha(m-o)}(1 - c_{j+1} + c_{j+1} e^{-\alpha(h-m)}) \Leftrightarrow \\
c_{j+1} \leq \frac{1 - e^{-\alpha(o-m)}}{1 - e^{-\alpha(h-m)}}
\end{aligned}$$

It follows that $c^*(\alpha, \Delta) = \frac{1 - e^{-\alpha(o-m)}}{1 - e^{-\alpha(h-m)}}$. To see the case $\alpha = 0$, either repeat the same passages for any affine utility, or argue by continuity, observing that $\lim_{\alpha \to 0} \frac{1 - e^{-\alpha(o-m)}}{1 - e^{-\alpha(h-m)}} = \frac{o-m}{h-m}$. □

Proof of Proposition 2 Recall that $(p_i, c_i)_{i=1}^N$ are i.i.d draws from τ. Since $i \in A^*$ if and only if $c_i > c^*$ and by the (weak) Law of Large Numbers it holds

$$\overline{O}(\tau, \alpha, \Delta) = 1 - \operatorname*{plim}_{N\to\infty} \frac{|A^*|}{N} = \operatorname*{plim}_{N\to\infty} \frac{\sum_{i=1}^N \mathbf{1}_{c_i \leq c^*}}{N} = \mathbb{P}_\tau(\{c \leq c^*\}) = F_\gamma(c^*)$$
$$\overline{H}(\tau, \alpha, \Delta) = \operatorname*{plim}_{N\to\infty} \frac{\sum_{i\in A^*} Z(p_i)}{|A^*|} = \operatorname*{plim}_{N\to\infty} \frac{\sum_{i=1}^N p_i \mathbf{1}_{c_i > c^*}}{\sum_{i=1}^N \mathbf{1}_{c_i > c^*}} = \mathbb{E}_\tau[p | c < c^*]$$

Observe the random variables $\frac{W(A^*)}{N}$ have mean and variance

$$\mu_N = o + (m - o)\frac{|A^*|}{N} + (h - m)\frac{1}{N}\sum_{i\in\hat{A}_N} p_i$$
$$\sigma_N^2 = (h - m)^2 \frac{1}{N^2}\sum_{i\in\hat{A}_N} p_i(1 - p_i)$$

Note that, for any N it holds $0 \leq \sigma_N^2 \leq \frac{(h-m)^2}{4N}$, hence, $\sigma_N^2 \to 0$ as $N \to \infty$. As a consequence, by Chebyshev's inequality, fixed any $\varepsilon > 0$

$$\mathbb{P}\left(\left|\frac{W(A^*)}{N} - \mu_N\right| > \epsilon\right) \to 0 \quad as \quad N \to \infty$$

Hence the sequence $\frac{W(A^*)}{N}$ has the same probability limit as the sequence μ_N. From point (b) it follows,

$$\operatorname*{plim}_{N\to\infty} \frac{\hat{W}(A^*)}{N} = m + \Delta^o\overline{O} + \Delta^h(1 - \overline{O})\overline{H}.$$

□

Proof of Proposition 3 Preliminarily, observe that $\forall\Delta, \alpha\ \overline{D}(\tau, \alpha, \Delta) \geq 0$ is equivalent to

$$\forall c \in [0, 1] \quad \mathbb{E}_\tau[P|C > c] - \mathbb{E}_\pi[P] \geq 0 \tag{10}$$

To see that Weak Positive Dependence implies Condition 10, it is sufficient to consider cases when $\overline{O} > 0$ (otherwise, the claim follow trivially from the the zero omissions polar case). Fix any c and consider $f(P) = P$ and $g(C) = \mathbf{1}_{C>c}$. Then Weak Positive Dependence implies $\mathbb{E}[P\mathbf{1}_{C>y}] - \mathbb{E}[P](1 - F_\gamma(c)) \geq 0$ and Equation 10 follows dividing by $(1 - F_\gamma(c))$ – which is positive since $\overline{O} > 0$. To see that Condition 10. implies Weak Positive Dependence, fix any two functions f, g as in the hypothesis. If f is constant, Weak Positive Dependence follows trivially. We can then reduce to the case where f is strictly increasing. In this case, $f(P) = aP + b$ for some $a > 0$. Then $Cov(f(P), g(C)) = aCov(P, g(C))$ and it is sufficient to

prove the claim for $Cov(P, g(C))$. Write

$$\begin{aligned} Cov(P, g(C)) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Cov(\mathbf{1}_{P>s}, \mathbf{1}_{g(C)>t})dsdt = \\ &= \int_{-\infty}^{\infty}\int_{0}^{\infty} Cov(\mathbf{1}_{P>s}, \mathbf{1}_{C>g^{-1}(t)})dsdt = \\ &= \int_{-\infty}^{\infty}\int_{0}^{\infty} \mathbb{P}(P > s, C > g^{-1}(t)) - \mathbb{P}(P > f^{-1}(s))\mathbb{P}(C > g^{-1}(t))dsdt \end{aligned}$$

The first equality follows by Hoeffding's covariance identity; the second because $Cov(\mathbf{1}_{P>s}, \mathbf{1}_{g(C)>t})$ is almost surely null for negative values of s, given that P is almost surely positive; the third just by rewriting. To show that the last quantity is positive, rewrite Condition 10 exploiting the fact that P is almost surely nonnegative, so that the layer-cake formula for expectations holds

$$\begin{aligned} \forall t \quad & \int_{0}^{\infty} \mathbb{P}(P > s | C > g^{-1}(t)) - \mathbb{P}(P > s)ds \geq 0 \Leftrightarrow \\ \forall t \quad & \int_{0}^{\infty} \mathbb{P}(P > s, Y > g^{-1}(t)) - \mathbb{P}(P > s)\mathbb{P}(Y > g^{-1}(t))ds \geq 0 \end{aligned}$$

Where the equivalence follows by multiplying by $\mathbb{P}(C > g^{-1}(t))$. Hence, integrating

$$\int_{-\infty}^{\infty}\int_{0}^{\infty} \mathbb{P}(X > s, Y > g^{-1}(t)) - \mathbb{P}(X > s)\mathbb{P}(Y > g^{-1}(t))dsdt \geq 0$$

□

Proof of Proposition 4 Since $\pi' \succeq_{CS} \pi$, $\mathbb{E}_{\pi'}[p] = \mathbb{E}_{\pi}[p]$. Hence, $\overline{D}(\tau', \alpha, \Delta) \geq \overline{D}(\tau, \alpha, \Delta)$ iff $\overline{H}(\tau', \alpha, \Delta) \geq \overline{H}(\tau, \alpha, \Delta)$. Fix any Δ. Observe that by equivalence, for any c^* and any increasing bijection f is holds $\overline{F}_{\psi(p)}(c^*) = \overline{F}_{\psi'(f(p))}(c^*)$. Hence, choosing $f = l$, $\overline{O}(\tau, \alpha, \Delta) = 1 - \mathbb{E}_{\pi}[\overline{F}_{\psi(p)}(c^*)] = 1 - \mathbb{E}_{\pi}[\overline{F}_{\psi'(l(p))}(c^*)] = \overline{O}(\tau', \alpha, \Delta)$ where the first equality follows since $\overline{F}_{\psi(p)}(c^*)$ is the probability of performing a task at performance level p; the last by this fact and the definition of pushforward. If $\overline{O} = 1$, the claim follows trivially (by the no performance polar case). Consider $\overline{O} < 1$. It holds

$$\overline{H}(\tau', \alpha, \Delta) = \frac{\mathbb{E}_{\pi'}[p\overline{F}_{\psi'(p)}(c^*)]}{\mathbb{E}_{\pi'}[\overline{F}_{\psi'(p)}(c^*)]} = \frac{\mathbb{E}_{\pi}[l(p)\overline{F}_{\psi(l(p))}(c^*)]}{\mathbb{E}_{\pi}[\overline{F}_{\psi(l(p))}(c^*)]} = \frac{\mathbb{E}_{\pi}[l(p)\overline{F}_{\psi(p)}(c^*)]}{\mathbb{E}_{\pi}[\overline{F}_{\psi(p)}(c^*)]}$$

where the first equality follows by definition of pushforward; the second since ψ is ordinal and l is increasing so that $\overline{F}_{\psi(p)} = \overline{F}_{\psi(l(p))}$. Then,

$$\overline{H}(\tau', \alpha, \Delta) - \overline{H}(\tau, \alpha, \Delta) = \frac{\mathbb{E}_{\pi}[(l(p) - p)\overline{F}_{\psi(p)}(c^*)]}{\mathbb{E}_{\pi}[\overline{F}_{\psi(p)}(c^*)]} = (\rho - 1)\frac{\mathbb{E}_{\pi}[(p - \mathbb{E}_{\pi}[p])\overline{F}_{\psi(p)}(c^*)]}{\mathbb{E}_{\pi}[\overline{F}_{\psi(p)}(c^*)]}$$

Observe that the difference has the same sign of the fraction's numerator, as the denominator is positive (being equal to $1 - \overline{O}(\tau, \alpha, \Delta) = 1 - \overline{O}(\tau', \alpha, \Delta)$) and $\rho > 1$. To see the numerator is positive, observe that since ψ is increasing, $\overline{F}_{\psi(P))}(c^*)$ is an increasing transformation of the

random variable P, just as $(P - \mathbb{E}_\pi[p])$. Then,

$$\mathbb{E}_\pi[(p - \mathbb{E}_\pi[p])\overline{F}_{\psi(p)}(c^*)] - \mathbb{E}_\pi[(p - \mathbb{E}_\pi[p])]\mathbb{E}_\pi[\overline{F}_{\psi(p)}(c^*)] = Cov((P - \mathbb{E}_\pi[p]), \overline{F}_{\psi(P))}(c^*)) \geq 0$$

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

Proof of Proposition 5 AAW Difference follows from Part (c) of Proposition 1 observing that $\overline{O} = 1 - \mathbb{E}_\pi[\overline{F}_{\psi(p)}]$ and $(1 - \overline{O})\overline{H} = \mathbb{E}_\pi[p\overline{F}_{\psi(p)}]$. To obtain Welfare Gap Inequality from AAW Difference, note that $\delta(c^*|P)$ is an increasing transformation of P, since ψ' is a positive shift of ψ. Hence, since also $(P\Delta^h - \Delta^o)$ is increasing we have

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

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

Proof of Proposition 6 Fix ψ as in the statement and any ψ'. From AAW Difference it follows that

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

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

Proof of Proposition 7 Fix any action $a \neq 0$. Observe that $\mathbb{E}_d[u(a, \omega)] = m + (h - m)d(\{a\})$. Since $h > m$, the expectation is maximized by $a = mode(d)$. Hence, success in a task i of type $(\theta, s_\omega, s_\theta)$ is the event $E_i = \{mode(d^{s_\theta}_{s_\omega})\}$. The ex-interim belief that the event realizes hence, by definition, $c(\theta, s_\omega, s_\theta)$. Instead, let $I(\theta, s_\omega, s_\theta) \subseteq [N]$ the set of indices of tasks of the $(\theta, s_\omega, s_\theta)$ sort. Then,

$$p(\theta, s_\omega, s_\theta) = \operatorname*{plim}_{N\to\infty} \frac{\sum_{i \in I(\theta, s_\omega, s_\theta)} \mathbf{1}_{E_i}}{|I(\theta, s_\omega, s_\theta)|}$$

Observe that task of sort $(\theta, s_\omega, s_\theta)$ happen infinitely often. By the weak Law of Large Numbers, this limit is $\mathbb{P}(E_i|\theta, s_\omega, s_\theta)$, which is precisely $d^\theta_{s_\omega}(E_i) = d^\theta_{s_\omega}(mode(d^{s_\theta}_{s_\omega}))$ since $d^\theta_{s_\omega}$ is the Bayesian posterior of Joint Distribution. □

Proof of Proposition 8 To show the result, I first introduce the following simple fact, which relies on the independence of ω and θ in the Joint Distribution.

Lemma A2 *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

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

this shows the point. □

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

Lemma A3 *Let $\sigma : \Omega \times \Theta \to \Delta(S_\Omega)$ be a family of symmetric diagonally dominant 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 diagonally dominant, with parameter $\mathbb{E}_t[\theta]$.*
2. *Let $d_0 \in \Delta(\Omega)$ be the uniform prior, and consider any 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

$$\mathbb{E}_t[\sigma_\theta](i|j) = \sum_{\theta\in\Theta} t(\theta)\sigma_\theta(i|j).$$

Hence it follows that

$$\forall i \in [M] \quad \mathbb{E}_t[\sigma_\theta](i|i) = \sum_{\theta\in\Theta} t(\theta)\sigma_\theta(i|i) = \sum_{\theta\in\Theta} t(\theta)\theta = \mathbb{E}_t[\theta]$$

Moreover,

$$\begin{aligned}\forall i \in [M] \quad \mathbb{E}_t[\sigma_\theta](i|i) = \sum_{\theta\in\Theta} t(\theta)\sigma_\theta(i|i) &\geq \sum_{\theta\in\Theta} t(\theta)\max_{i\neq j}\sigma_\theta(i|j) \geq \\ &\geq \max_{i\neq j}\sum_{\theta\in\Theta} t(\theta)\sigma_\theta(i|j) = \max_{i\neq j}\mathbb{E}_t[\sigma_\theta](i|j)\end{aligned}$$

where the last inequality is Jensen's inequality. This shows that the expected experiment is symmetric diagonally dominant.
To see the second point, observe that, given that the prior is uniform $d_{s_\omega}(\omega) = \sigma(\omega|\sigma_\omega)$. This is maximized by $\omega = \sigma_\omega$, by symmetric diagonal dominance. By the same token, we see that the mass on the mode is θ. □

The result follows combining the two Lemmata with Proposition 7. First note that we have $p(\theta, s_\omega, s_\theta) = d^\theta_{s_\omega}(s_\omega) = \theta$ and $c(\theta, s_\omega, s_\theta) = d^{s_\theta}_{s_\omega}(s_\omega) = \mathbb{E}_{t_{s_\theta}}[\theta]$.

Then, to show the first direction, observe that the distribution of performance levels is just the unconditional distribution of θ, which is t_0. Instead, conditional on a performance level $p \equiv \theta$, the distribution of the confidence judgment is that of the conditional posterior mean $m(t_0, \mu)(\theta)$. To see that the induced skill type is a martingale coupling note that, given a confidence judgment realization c, $\mathbb{E}[p|c] = \mathbb{E}[\theta|c] = c$ by construction.

In the converse direction, suppose that $\tau \equiv (\pi, \psi)$ is a martingale coupling. Then, take π as a distribution over symmetric diagonally dominant experiments $\sigma : \Omega \times P \to \Delta(S_\Omega)$. Since $p(\theta, s_\omega, s_\theta) = \theta$, we get that the distribution of performance levels induced by the information processing model is precisely π. As a meta-experiment, take $\mu \equiv \psi$. To show that this indeed works, recall that $c(\theta, s_\omega, s_\theta) = \mathbb{E}_{t_{s_\theta}}[\theta]$. Then, observe that updating prior π by experiment ψ after a signal/confidence judgment c we have a posterior mean

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

where the last equality follows since τ, by hypothesis, is a martingale coupling.

Proof of Proposition A9 Point (a) follows by definition. The observations about the omit rate follow by definition of FOSD and since c^* increases in α. The observations about the omit rate follow by definition of HRD and since c^* increases in α. To see the points about welfare observe that AAW rewrites

$$\overline{W} = \mathbb{E}_\pi[\underbrace{o\mathbf{1}_{p<c^*} + (m + (h-m)p)\mathbf{1}_{c \geq c^*}}_{f_\alpha(p)}]$$

It is easy to see that for the risk neutral $c^* = \frac{o-m}{h-m}$, $f_0(p)$ is weakly increasing hence $f_\alpha(p)$ is also weakly increasing for any risk averse $c^* > \frac{o-m}{h-m}$. Hence, the scores weakly grows for FOSD improvements in π. Moreover, it holds pointwise $f_{\alpha'} \leq f_\alpha$ if $\alpha \leq \alpha'$, hence the quantity decreases in risk aversion.

C Empirical Appendix

Details on Empirical Bayes Estimation I use an empirical Bayes procedure, exploiting the Beta-Bernoulli conjugate, to estimate the hit-rates and the omit-rates at the individual level. For each observation, I compute the raw proportion of successes (in the case of hit-rates) or omissions (for omit-rates) $p_{\text{raw}} = S/N$. These raw proportions are assumed to be drawn from a Beta(α, β) distribution with mean $\alpha/(\alpha+\beta)$ and variance $\alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$. Using the method of moments on the collection of raw proportions across all observations, the hyperparameters are estimated by computing the sample mean $\bar{p}$ and variance s^2 and then solving $k = \bar{p}(1-\bar{p})/s^2 - 1$ with $\alpha = \bar{p}\,k$ and $\beta = (1-\bar{p})\,k$. Next, this is taken as the prior of a Beta-Bernoulli model. Hence, for each observation the posterior estimate is calculated as the posterior mean $\hat{p}_{\text{hier}} = \frac{S+\alpha}{N+\alpha+\beta}$. Note that this shrinks the raw estimate toward the overall prior mean by assigning more weight to the prior when N is small and less when N is large.

Robustness of the Hit-Rate Differential I perform various robustness checks on the hit-rate differential, replicating Figure 1a for different subsamples. In particular Figure A2 replicates the finding separately for both genders; Figure A3 focus on participants who took part in the math contest more than once; Figure A4 on participants who do not omit any question under no penalties.

Robustness of the Gender Gaps I perform various robustness checks on the gender gaps, replicating Figure 2 and 3 for different subsamples. In particular Figures A5 and A6 focus on participants who took part in the math contest more than once; Figure A7 and A8 on participants who do not omit any question under no penalties.

Math Performance in School While an analysis of the relation between metacognition and external outcomes is beyond the scope of this paper, I move in this direction exploiting math school grades, which are recorded for some units in my dataset. I normalize the grades at school level, to account for school effects. Then I regress this outcome on an indicator of high metacognition which takes value one if the hit-rate differential (normalized at test level) is above median interacted with gender and an indicator of risk aversion. Table A1 reports the results. It is interesting to note that the high metacognition group has better normalized math grade on average, though this effect is not strongly significant.

Identification of Maximum Likelihood In my maximum likelihood estimation procedure, I first pool the data by gender. Then I summarizing counts over the two sets of questions corresponding to the two parts of the test. For questions 1–13, let $n_{1,13}$ be the total number of observations, $\text{NA}_{1,13}$ the total number of missing responses, and $\text{ones}_{1,13}$ the total number of ones; I then compute the error rate under no penalties as

$$\hat{\varepsilon} = \frac{\text{NA}_{1,13}}{n_{1,13}},$$

and the hit rate under no penalties as

$$\hat{H}^0 = \frac{\text{ones}_{1,13}}{n_{1,13} - \text{NA}_{1,13}}.$$

Using the given performance levels $\underline{p}$ and $\overline{p}$, I invert Equation 2 to obtain the ability parameter

$$\hat{\pi} = \frac{\hat{H}^0 - \underline{p}}{\overline{p} - \underline{p}}.$$

Similarly, for questions 14–25, I define $n_{14,25}$ as the total observations, $\text{NA}_{14,25}$ as the total number of missing responses, and $\text{ones}_{14,25}$ as the number of ones; hence, the omission rate is estimated as

$$\hat{O} = \frac{\text{NA}_{14,25}}{n_{14,25}},$$

and the raw success rate (among non-missing responses) is

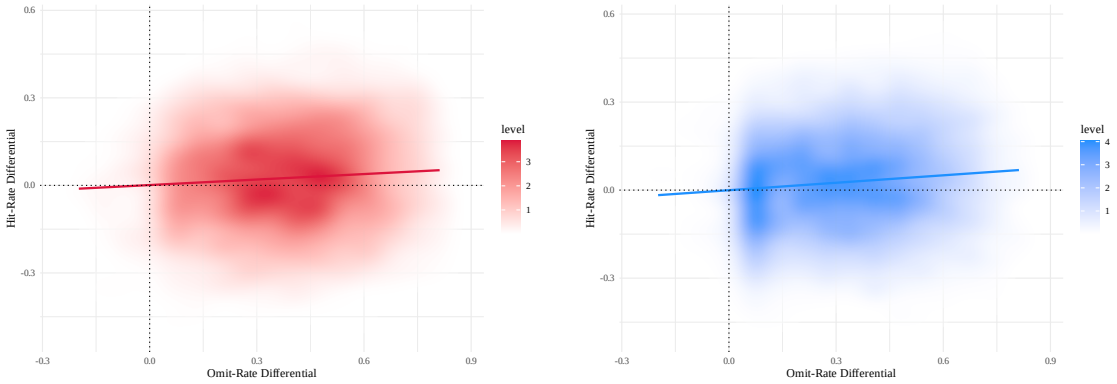
$$\hat{H}^{>0} = \frac{\text{ones}_{14,25}}{n_{14,25} - \text{NA}_{14,25}}.$$

In turn, inverting the system given by Equations 3 and 7, we obtain

$$\hat{\psi}_+ = \frac{(1-\hat{O})(\hat{H}^{>0} - \underline{p})}{\hat{\pi}(\overline{p} - \underline{p})} \quad \text{and} \quad \hat{\psi}_- = \frac{\hat{O} - \hat{\pi}(1 - \hat{psi}_+)}{1 - \hat{\pi}},$$

Hence, the system is identified. Confidence intervals are obtained by the Delta method.

Figure A2: Joint of Hit-Rate and Omit-Rate Differentials by Gender



(a) Females (b) Males

Figure A3: Joint of Hit-Rate and Omit-Rate Differentials – Experienced Participants

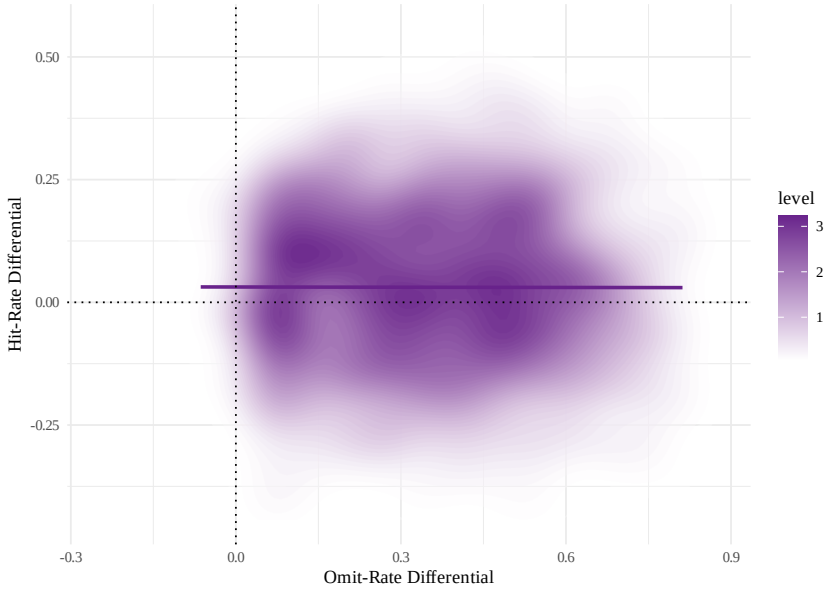


Figure A4: Joint of Hit-Rate and Omit-Rate Differentials – No Errors

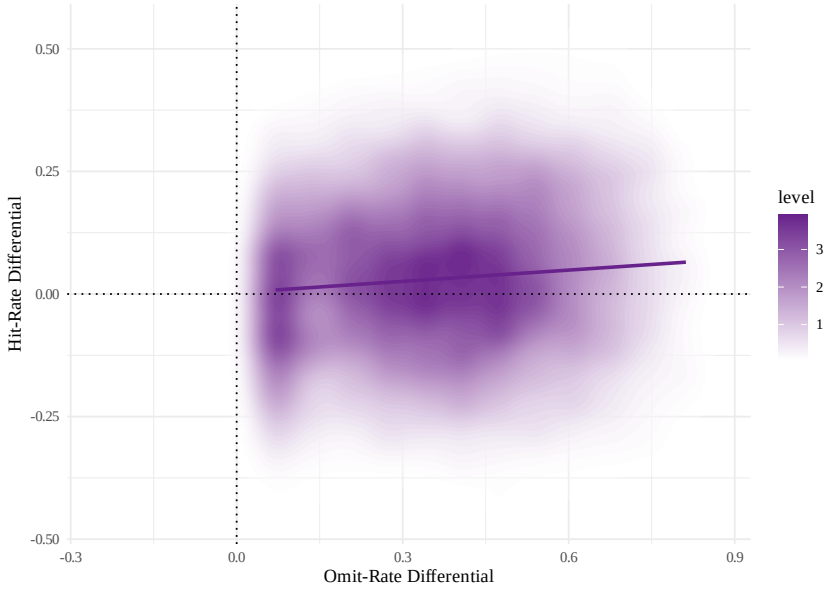
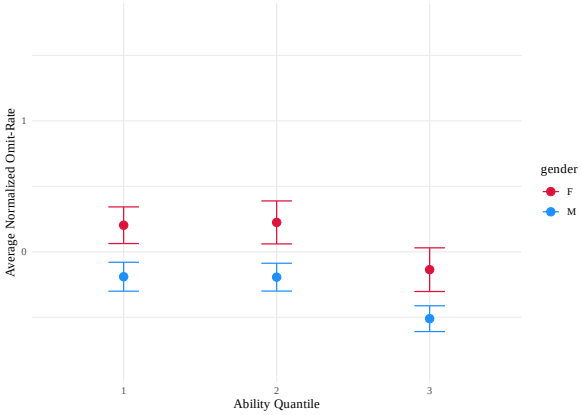


Figure A5: Gender Gaps vs Ability – Experienced Participants

(a) Omit-Rate (Normalized)



(b) Score (Normalized)

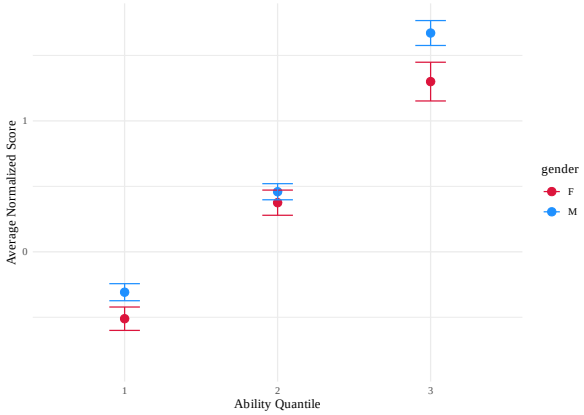
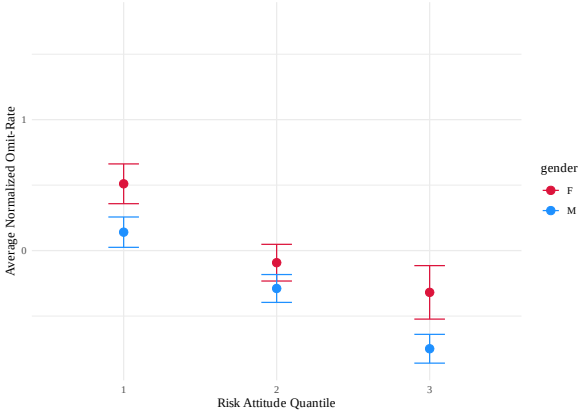


Figure A6: Gender Gaps vs Risk – Experienced Participants

(a) Omit-Rate (Normalized)



(b) Score (Normalized)

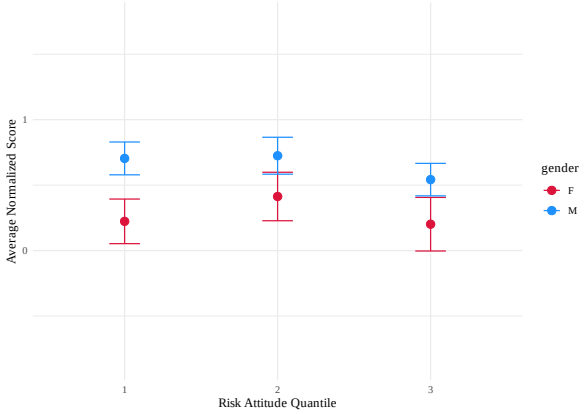
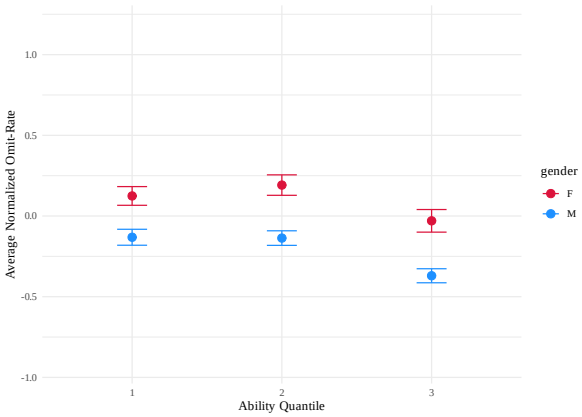


Figure A7: Gender Gaps vs Ability – No Errors

(a) Omit-Rate (Normalized)



(b) Score (Normalized)

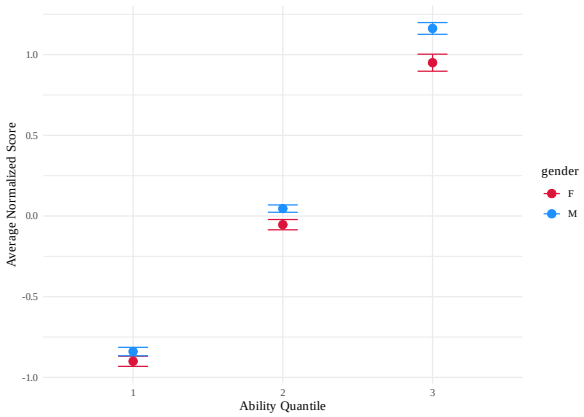
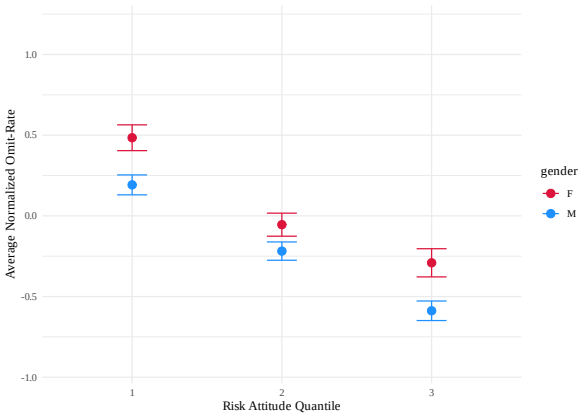


Figure A8: Gender Gaps vs Risk – No Errors

(a) Omit-Rate (Normalized)



(b) Score (Normalized)

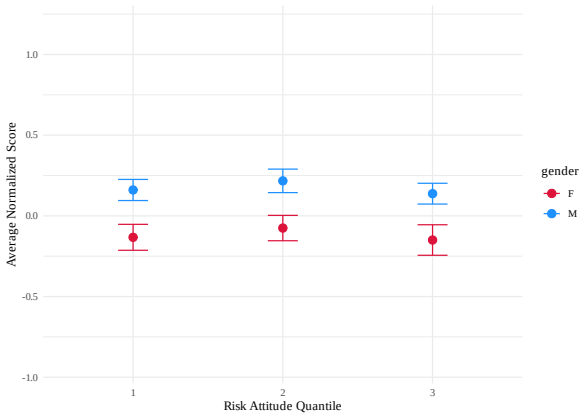


Table A1: Metacognition and Math Grades

	Normalized Math Grade
High Metacognition	0.110*
	(0.059)
Gender	−0.172***
	(0.050)
Risk Aversion	0.029
	(0.067)
Observations	4,686
R^2	0.012
Adjusted R^2	0.010
Residual Std. Error	0.944 (df = 4678)
F Statistic	7.884*** (df = 7; 4678)
Note: interactions and constant are omitted	*p<0.1; **p<0.05; ***p<0.01