

The Malleability of Motivated Beliefs

Yiming Liu* and Stephanie W. Wang†

Abstract

We study the malleability of motivated beliefs in a high-stakes environment where students estimate entrance exam scores for school choice decisions before receiving results. Students exhibit significant overconfidence when recalling past scores, with systematic patterns revealing motivated memory: bias increases with poor performance, greater ego-relevance, and ease of manipulation. Surprisingly, the same students state accurate beliefs when estimating entrance exam scores, despite supply-side mechanisms remaining active — students rely on biased recall that transmits through Bayesian updating at predicted rates. To resolve this puzzle of accurate beliefs despite biased inputs, we provide evidence for conscious downward adjustments that offset initial overconfidence. These results demonstrate that motivated beliefs are malleable: psychological biases persist but deliberative correction emerges when accuracy matters.

1 Introduction

Biased beliefs about ability and performance, such as overconfidence or optimism bias, have been documented in many domains (Moore and Healy, 2008) and affect economic and political behavior (Camerer and Lovallo, 1999; Malmendier and Tate, 2005; Ortoleva and Snowberg, 2015; Mueller et al., 2021; Bosch-Rosa et al., 2024; Pires, 2025). Representing the view that these are innate biases, Daniel Kahneman stated that overconfidence is “built so deeply into the structure of the mind that you

*Humboldt University of Berlin, WZB Berlin Social Science Center. Email: yiming.liu@wzb.eu.

†University of Pittsburgh. Email: swwang@pitt.edu. Wang completed this work as a Fellow at the Center for Advanced Study in the Behavioral Sciences.

couldn't change it without changing many other things.”¹

In contrast, models of motivated beliefs such as [Bénabou and Tirole \(2002\)](#), [Compte and Postlewaite \(2004\)](#), [Brunnermeier and Parker \(2005\)](#), [Kőszegi \(2006\)](#), [Gottlieb \(2010\)](#) and [Gossner and Steiner \(2018\)](#) suggest that overconfidence reflects an optimal tradeoff between psychological or instrumental benefits such as ego utility and the cost of decision mistakes.² A key prediction of these models is that biased beliefs should be malleable, adjusting when the stakes change, rather than fixed cognitive limitations. Yet empirical evidence for such malleability remains scarce, and more importantly, we know little about the mechanisms through which belief adjustment might occur. While the literature has identified supply-side channels that generate overconfident beliefs, such as biased memory ([Bénabou and Tirole, 2002](#)), we lack understanding of how individuals might correct these biases when accuracy becomes crucial. Do people shut down the supply side processes entirely, or do they maintain motivated processes while applying compensatory adjustments?

Studying the malleability of motivated beliefs requires a rare combination of empirical conditions. First, the domain must be ego-relevant to trigger motivated reasoning in the first place, yet involve sufficient consequences that accuracy matters for real outcomes. Second, we need randomized assignment to different stake levels or within-person variation in stakes to isolate how the environment affect belief formation while controlling for individual heterogeneity. Third, objective performance benchmarks are essential for measuring overconfidence. Fourth, to test whether beliefs follow Bayesian updating, we need to observe both individuals' priors and the signals they receive. Lastly, to test whether individuals shut down supply-side processes that produce motivated beliefs in response to high stakes, we must observe the supply-side mechanisms generating biased beliefs, e.g. how individuals access and process information from memory.

We study an environment that satisfies these conditions: high school admissions in a city in Hebei province, China. Elite schools admit students solely based on entrance exam scores through an immediate acceptance algorithm, with approximately 30% of students gaining admission. Critically,

¹See <https://www.theguardian.com/books/2015/jul/18/daniel-kahneman-books-interview>.

²The benefits of overconfidence include providing motivation for effort ([Bénabou and Tirole, 2002](#)), enhancing performance ([Compte and Postlewaite, 2004](#)), generating anticipatory or ego utility ([Brunnermeier and Parker, 2005](#); [Kőszegi, 2006](#); [Gottlieb, 2010](#); [Bracha and Brown, 2012](#); [Caplin and Leahy, 2019](#)), and minimizing expected losses under noisy perception ([Gossner and Steiner, 2018](#)).

students must submit school choices before learning their exact scores, after receiving only answer keys and scoring rubrics. This allows us to observe the same individuals in two contexts: recalling their mock exam performance from a month earlier, and estimating their entrance exam scores for school choice decisions. With standardized scoring across six subjects (Mathematics, Chinese, English, Physics & Chemistry, History & Politics, and Geography & Biology), we have objective benchmarks for measuring bias. These subjects vary naturally in ego-relevance (e.g. due to gender congruence) and cognitive manipulability. We observe the students' recalled mock scores (their prior), information about their entrance exam performance, and their final estimates (their posterior), allowing us to track how biased memory transmits through Bayesian updating. This setting thus provides unusual visibility into both the psychological mechanisms generating motivated beliefs and how these mechanisms respond to stakes.

We surveyed around 300 students immediately before they submitted their school choices, capturing their beliefs at this critical decision point. Students provided their estimated entrance exam scores across all six subjects and were asked to recall their mock exam scores across all six subjects as well from one month prior. These subjective beliefs, both recalls and estimates, were matched to administrative records of actual performance on both exams, providing objective benchmarks for measuring bias in recall and estimation. This within-person design allows us to track how the same individuals form beliefs across different tasks and how memory distortions are transmitted into consequential estimations. We also collected demographic information from administrative records (gender, age) and survey responses (risk attitudes, family background) to examine heterogeneity in belief formation and adjustment patterns.

To guide our empirical investigation, we develop a conceptual framework of belief formation under motivated reasoning. Individuals face a trade-off when recalling past performance: maintaining positive self-image provides ego utility, but distorting memories imposes cognitive costs. This generates motivated recall where individuals systematically inflate remembered performance, with the optimal bias balancing ego benefits against psychological costs of distortion. The magnitude of recall bias should vary predictably: increasing with ego-relevance (how much the domain matters for self-image), decreasing with cognitive costs (how difficult it is to distort memories in that domain), and increasing when actual performance is worse (where ego utility gains are largest). When forming beliefs about current performance, individuals combine their recalled past performance as a prior with

noisy signals through Bayesian updating. Crucially, this updating process takes the biased recall at face value, namely individuals are unaware their memories are distorted. Under standard Bayesian updating with normal distributions, the posterior belief becomes a weighted average of the biased prior and the signal, with weights determined by their relative precision. Since the prior is inflated by motivated recall, this bias transmits directly into posterior beliefs at a rate equal to the Bayesian weight on the prior.

Our empirical results provide strong support for motivated memory in recall. Students exhibit significant overconfidence when recalling their mock exam performance. Despite once knowing their exact scores, students' recalled scores are on average 0.1 standard deviation higher than actual scores. The magnitude of this bias exceeds what experts familiar with the setting expected, who predicted near-accurate recall. This overconfidence is pervasive: 59% of students recall higher scores than their actual performance.

The patterns in recall bias match the predictions of motivated memory. Students with worse actual performance show greater overconfidence in recall, and within individuals, the tendency to inflate recalled scores is stronger for subjects where they performed relatively poorly, suggesting that memory distortion serves to protect self-image where it is most threatened. Recall bias is significantly larger in non-STEM subjects compared to STEM subjects, consistent with mathematical and scientific knowledge being stored in more structured ways that make distortion more cognitively costly. Gender differences reveal the role of ego-relevance: male students exhibit substantially greater overconfidence than females, with this gap particularly pronounced in STEM subjects that are male-stereotyped. These systematic variations with performance, cognitive costs, and ego-relevance confirm that recall errors reflect motivated reasoning rather than random memory failures.

Given these biased priors and Bayesian updating, our conceptual framework predicts students should also exhibit overconfidence when estimating entrance exam scores. Surprisingly, we find that students show no average overconfidence in estimation, with estimated scores just 0.04 standard deviation below actual scores (not significantly different from zero). The distribution of estimation errors is more balanced than recall errors: while only 2% of students make exactly accurate predictions, there is no systematic tendency toward overestimation, with 48% over-estimating and 49% under-estimating their scores. This lack of systematic bias is also found in administrative data from one school where students reported their estimated scores to the school. This accuracy was unexpected by

experts, who predicted significant overconfidence in estimation, expecting students to overestimate by 0.13 standard deviation.

The absence of overconfidence in estimation despite significant overconfidence in recall presents a puzzle: how do students arrive at accurate beliefs when the psychological forces generating bias remain active? This suggests motivated beliefs may be malleable, adjusting to incentives and costs. But through what mechanism? One possibility is that students are sophisticated enough to recognize their biased recall would lead to bias in estimation and correct the recall before forming estimates. However, we can rule this out immediately — students remain significantly overconfident when recalling mock exam scores (0.1 s.d. overconfidence), showing no evidence of correcting their biased memories even when accuracy matters for school choice.

Alternatively, students might shut down the supply-side mechanism by relying on objective information rather than biased recall when forming estimates. After all, actual mock exam scores were readily available to students through official records. To test this, we examine how students combine information when forming estimates, regressing estimated entrance exam scores on both recalled and actual mock exam scores. Strikingly, students rely almost exclusively on their biased recall: recalled scores are highly significant predictors across all subjects, while actual scores have minimal predictive power despite being more accurate. This demonstrates that the supply-side mechanism, using biased memory as input to belief formation, remains operational.

We further test whether students follow Bayesian updating when combining their biased prior with new signals. Under Bayesian updating with normally distributed beliefs, the posterior should be a weighted average of prior and signal, with weights summing to one. Across all six subjects, we find precisely this pattern: students combine recalled mock scores (prior) and entrance exam performance (signal) with weights summing to approximately one, and we cannot reject Bayesian updating in any subject. Moreover, the transmission of recall bias into estimates occurs at exactly the rate predicted by the Bayesian weight on the prior.

These results deepen the puzzle: students use biased recall, update following Bayes' rule, yet achieve accurate beliefs. One way to reconcile these findings is through a conscious correction stage in the belief formation process. We develop a conceptual framework where belief formation involves two sequential processes. First, individuals automatically form posterior beliefs by combining biased recall with new information through Bayesian updating, unaware that their memories are distorted.

Second, a deliberative process inspects these posterior beliefs and makes conscious adjustments. The key insight is that individuals may possess partial meta-awareness — they sense they tend toward overconfidence without understanding that biased memory is the source. Unable to identify or correct the root cause, they instead make direct downward adjustments to their posterior beliefs. This adjustment process should respond to decision stakes: when consequences are minimal, the cognitive effort of correction may not be worthwhile, but when mistakes are costly, individuals consciously deflate their beliefs to offset suspected overconfidence.

To test this framework, we estimate the Bayesian updating regression while constraining weights to sum to one so any systematic deviation from the Bayesian posterior appears in the residual. Taking this residual as the correction term, we find that students make systematic downward adjustments to their posterior beliefs, with an average correction of -0.12 standard deviation, which is similar in magnitude to their initial overconfidence in recall.

The correction term exhibits patterns consistent with partial meta-awareness and responsiveness to incentives rather than random noise. Students across the first nine performance deciles make negative corrections, while only the top decile, who face minimal admission uncertainty, show non-negative adjustments. This variation with stakes indicates that corrections reflect conscious responses to decision stakes. Crucially, corrections also increase with subjective uncertainty: students reporting wider confidence intervals make substantially larger downward adjustments, suggesting they recognize that greater uncertainty signals more reliance on potentially biased priors. These systematic patterns validate our interpretation that the residuals capture meaningful belief adjustments rather than random error. Taken together, our results suggest students possess sophisticated partial meta-awareness: they not only recognize their general tendency toward overconfidence but also understand that this bias is larger when their beliefs rely more heavily on the prior.

Our study contributes to the theoretical and empirical literature on overconfidence as a motivated belief (Bénabou and Tirole, 2002; Compte and Postlewaite, 2004; Brunnermeier and Parker, 2005; Köszegi, 2006; Gottlieb, 2010; Bracha and Brown, 2012; Bénabou and Tirole, 2016; Caplin and Leahy, 2019; Orhun et al., 2024; Hagenbach and Saucet, 2025). While this literature has established that overconfidence can be understood as an optimal response to tradeoffs between ego utility and decision costs, a central prediction, that motivated beliefs should be malleable rather than fixed, has remained largely untested. We provide field evidence for this malleability: the same individuals who

exhibit significant overconfidence when recalling past performance state accurate beliefs on average when estimating performance for consequential decisions. Beyond documenting malleability, we investigate the underlying mechanisms: while we find no evidence that individuals correct their biased recall or shut down the transmission of recall bias into estimates, we find suggestive evidence for conscious downward adjustments that respond to stakes and uncertainty.

Our study also adds to the growing literature on biased memory as a supply-side mechanism that produces overconfident beliefs. Our within-individual analysis indicates that overconfidence in recall is correlated with overconfidence in estimation, consistent with biased memory serving as a tool for motivated belief formation (Bénabou and Tirole, 2002; Huang et al., 2020; Zimmermann, 2020; Hagenbach and Koessler, 2022; Huffman et al., 2022; Sial et al., 2023; Roy-Chowdhury, 2024; Gödker et al., 2025; Hagenbach et al., 2025). We contribute to the supply-side literature by showing that this mechanism remains robust even when decision stakes are high. Despite strong incentives for accuracy, individuals do not “turn off” this underlying force, suggesting limited awareness of its role in generating biased beliefs. In addition, we provide rich evidence for the motivated nature of memory distortion: recall bias varies systematically with performance, cognitive costs, and ego-relevance in patterns consistent with motivated distortions rather than random memory errors.

Lastly, our study speaks to the question of whether biases persist or diminish when the stakes are very high. Previous field studies focus on documenting biases in high-stakes contexts without systematically examining the same individuals across varying decision environments (Metrick, 1995; Berk et al., 1996; Levitt, 2004; Belot et al., 2010; Pope and Schweitzer, 2011; Graddy et al., 2014; Chen et al., 2016; Jetter and Walker, 2017; Heck et al., 2024; Klein Teeselink et al., 2024). Meanwhile, experimental work showed that higher incentives can mitigate but seldom eliminate cognitive biases (Camerer, 1987; Camerer and Hogarth, 1999; Ariely et al., 2009; Enke et al., 2023; Exley and Kessler, 2024; Gneezy et al., 2024). By observing how the same individuals form beliefs in different contexts, recalling past performance and estimating consequential exam scores, our paper complements this literature. We document that while average overconfidence disappears in high-stakes estimation, the psychological mechanisms generating biased beliefs remain fully operative, with accuracy emerging through compensatory adjustments rather than through motivated processes shutting down. This finding suggests that the relationship between stakes and bias is more nuanced than simple elimination, involving complex interactions between persistent psychological forces and conscious correction

mechanisms.

The rest of the paper proceeds as follows. Section 2 describes the high school admission system in China and our data collection. Section 3 presents our theoretical framework that guides the empirical analysis. Section 4 examines overconfidence in the recall of past performance. Section 5 analyzes whether students exhibit overconfidence in high-stakes estimation. Section 6 investigates the supply side of biased beliefs by testing the role of biased memory and Bayesian updating in belief formation. Section 7 identifies and examines the conscious correction mechanism. Section 8 concludes.

2 Background and Data

High school admission in Baoding, a city in Hebei province of China, is conducted through the immediate acceptance algorithm (IA). An identical priority ordering fully defines the high schools' preferences over students: the high school entrance exam score.

The High School Entrance Exam. All students take the two-day exam which consists of six subject exams: Mathematics, Chinese, English, Physics & Chemistry, History & Politics, and Geography & Biology. Students can score between 0 and 120 points in each subject exam except for the geography and biology exam, where the highest possible score is 60.

All students take a mock exam one month before the entrance exam that aims to mimic every aspect of the actual exam. The questions are drawn from the same question bank, and the pool of graders is also the same. Students learn their score and school ranking for the mock exam about two weeks before the entrance exam.

The Admission Mechanism. There are two tiers of high schools. Tier-1 schools, or the so-called "Provincial Key High Schools," are high-quality schools with limited seats. Only about thirty percent of middle school students can enter a Tier-1 high school. The second tier has lower-quality academic schools and vocational schools. Students and their parents universally prefer Tier-1 schools because it is practically the only route to college. Even though all students can apply to Tier-1 high schools, only the choices of the qualified students are considered in the admission process. To be qualified, a student's high school entrance exam score must be higher than the median score across all students in the city. Whether a student is qualified or not is not known at the time of application because the exams are not graded at this point.

Tier-1 high schools allocate admission quotas to each middle school, creating a system where

students compete against their peers within the same middle school for Tier-1 seats. These quotas are announced before the high school entrance exam. Schools rank students solely based on their entrance exam total score, which is unknown at the time of application. Students can apply to only one Tier-1 high school, regardless of how many such schools exist in their district.

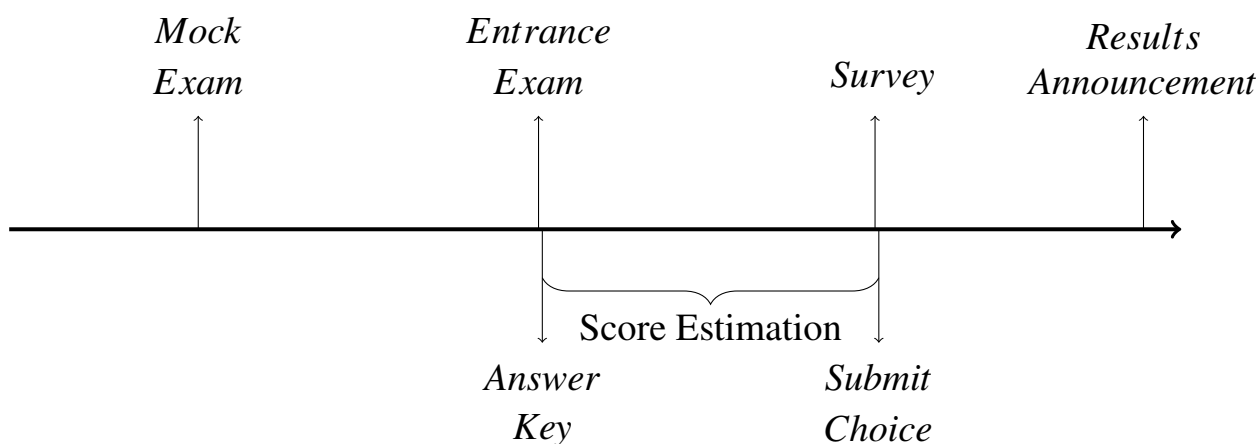
Students learn their entrance exam scores one week after submitting their school choices. The matching process then proceeds in two steps. First, only students who scored above the city-wide median are eligible for Tier-1 high schools. Among these eligible students, each Tier-1 school accepts the highest-ranked applicants from each middle school up to its predetermined quota. Students who are not accepted remain unmatched.

Second, an “aftermarket” round occurs if some Tier-1 high schools have unassigned seats after the first round. Students must indicate *ex ante*, before learning their match outcomes, whether they wish to participate in this aftermarket. Unmatched students who volunteered for this round are then assigned to remaining seats at Tier-1 high schools that still have vacancies. Crucially, students cannot choose which school they will attend in this round. Instead, the city education committee assigns students with higher scores to schools that are ranked higher according to the committee’s own ranking of schools. This aftermarket thus provides a safety net for unmatched students, though at the cost of school choice.

A distinctive feature of this matching mechanism is that students must submit their school choice before learning their exam scores. As shown in Figure 1, the admission process unfolds as follows: All students take the two-day high school entrance exam at the end of their third year of middle school. After the exam, students receive the scoring rubric and answer keys to help them estimate their performance. Seven days later, they must submit their school choice without knowing their actual exam scores.

The Immediate Acceptance (IA) mechanism creates strong incentives for accurate score estimation because students must strategically “game” the system. Unlike strategy-proof mechanisms where students only need to report their true preferences, under IA, the optimal strategy depends crucially on a student’s expected score. Consider a student who prefers school a to school b , where school a typically requires a score of 580 and school b requires 550. Under a strategy-proof mechanism, the student would simply list school a first. However, under IA, this is optimal only if she expects to score above 580; if her score is likely below 580, listing school b first might be her best strategy.

Figure 1: Timeline



Thus, accurate score estimation becomes crucial for optimal school choice.

However, two groups of students face lower stakes for accurate estimation. First, top-performing students, similar to beneficiaries of the “Texas top 10% rule,” have essentially guaranteed admission to any school. In our setting, since the most prestigious high school admits about 15% of students from each middle school, students ranking in the top have little incentive to estimate precisely. Second, students who expect to score below the Tier-2 high school threshold face low stakes because they will need to retake the exam the following year regardless of their estimation.

Data. We use various data sources to analyze how students estimate their high school entrance exam scores. We surveyed students in two districts of the city on the day students went back to school to submit their high school choice. The survey was given right before students submitted their high school choices. At this point, they had finished estimating their scores without knowing their final match. In the survey, we asked students to report their best-estimated scores for each subject exam and the possible score ranges (their highest and lowest possible scores). In addition, we collect demographic information, such as parental education, at the end of the survey.

It is important to note that we did not ask students to estimate their scores; instead, we simply asked them to report the scores they had already estimated. We also made it clear that as external researchers, we would de-identify the data and not share their estimates with their parents, classmates, or teachers. This measure is intended to maximize their willingness to report their estimates

truthfully.³

Apart from the estimations, we asked students to recall their exam scores on all six subjects in the mock exam. They were instructed to try their best to recall their scores without checking them. Because we conducted the survey on the day they went back to school to submit their school choice, students also did not have access to their mock exam scores when they answered the survey.

We obtained administrative data from that middle schools that included the students' actual high school entrance exam scores on all six subjects, their actual mock exam scores on all subjects, and basic demographic information, including gender and age.

We collected 305 valid survey responses out of 321 questionnaires distributed, after excluding 16 students who did not report estimated scores for all six subjects. From these 305 responses, we further excluded three students with missing entrance exam scores, resulting in a final sample of 302 students. Since each student estimated scores for six subjects, our final sample contains 1,812 score estimations (302 students \times 6 subjects).

However, we have significant missing data for the recall task, with 334 out of 1,812 observations (18.4%) missing recalled mock exam scores. The reasons for this are unclear: it could reflect difficulty in remembering specific scores, survey fatigue, or selective non-response. To assess whether this missing data threatens our results, Table A.1 compares the full sample with students who completed the recall task. Panel A shows that at the observation level (student-subject), students with recall data have slightly lower entrance exam scores ($P < 0.05$) and estimated scores ($P < 0.01$) compared to the full sample. However, these differences are economically small, representing less than 0.05 standard deviation of the respective scores. Mock exam scores show no significant difference. More importantly, Panel B presents student-level comparisons, where we observe no statistically significant differences in demographics between the full sample and the recall sample. We further discuss the nature of this missing recall data in Section 4 when presenting results on overconfidence in recall. We also replicate our findings on estimation accuracy in Section 5 by restricting our data to those with

³The admission algorithm incentivizes students to estimate their scores accurately, but they may also have reasons to misestimate or misreport them intentionally. Students often share their estimated scores with parents, teachers, and classmates, which can influence their reporting strategy. Some students may over-report their scores to please their parents temporarily, while others may underreport them to surprise them later. As a result, it is unclear how sharing behavior affects estimation and reporting accuracy.

recall data only and present the results in Appendix C.

Estimating scores vs. estimating placement. Ultimately, what matters for admission in our setting is each student's *placement*, raising the natural question of why we focus on overestimation of *absolute* scores rather than overplacement (i.e., one's rank). In practice, however, students in our study rarely need to assess how their scores compare to those of other students. Instead, there is a two-step process in which (i) students form their own best guess of their individual exam scores, and (ii) an external agent, such as their school or teacher, ranks students' estimated scores to determine likely placement or admission thresholds. Indeed, for one of the two schools in our sample, the school explicitly collected the students' estimated scores and provided a predicted ranking based on all submitted estimates. Moreover, even for those students whose school does not collect estimated scores, teachers typically have a good sense of the cutoff score. Since the median score historically varies by less than 20 points (out of a total of 720 points) in the three years prior to our study period, teachers can reliably map this year's performance onto past results. Consequently, students' core task is to accurately predict their own scores (step (i)), while step (ii), determining how those scores translate to placement, is largely handled by schools or teachers. Given this institutional arrangement, accurate self-estimation is effectively the high-stakes component from the student's perspective, motivating our focus on overestimation.

Information available for score estimation and recall. It is important to clarify the information available to students in each task. For the mock exam, students initially received their exact scores after taking the exam, along with answer keys and detailed feedback. Thus, when we ask them to recall their mock exam performance one month later, they are recalling scores they once knew with certainty. In contrast, for the entrance exam, students never learn their exact scores before submitting school choices — they only receive answer keys and scoring rubrics. This asymmetry in information availability is crucial for interpretation: students have access to more complete and precise information about their mock exam performance (which they may forget or distort) than about their entrance exam performance (which they must estimate). Higher accuracy in estimation than recall cannot be attributed to better information for the entrance exam; if anything, the information advantage lies with the mock exam.

Incentives for accurate estimation under immediate acceptance. The immediate acceptance mechanism creates strong incentives for accurate score estimation, and this holds regardless of the

students’ individual application strategies. We do not make assumptions about which strategies students adopt, nor do we attempt to distinguish between different strategic approaches in our analysis. Some students might aim for prestigious “reach” schools while others prefer guaranteed admission at “safety” schools, and risk preferences, family considerations, or school-specific factors may all influence these choices. Our key point is simply that any rational strategy, whether aggressive or conservative, requires accurate beliefs about one’s score as a crucial input. Consider a student deciding between schools with cutoffs at 550 and 580 points. Whether her goal is to maximize prestige or to guarantee admission, she needs accurate beliefs about where her score falls relative to these thresholds. Inaccurate beliefs lead to costly mistakes: overconfidence results in rejection, while underconfidence leads to unnecessarily settling for lower-ranked schools.

3 Conceptual Framework

We present a conceptual framework to describe the students’ decision environment. Consider an individual who needs to estimate their current task performance P_j for subject j based on two sources of information: a prior belief derived from recalled past performance, and a noisy signal about current performance from self-grading with answer keys.

We model belief formation as a two-stage process. First, individuals recall past performance. When accessing memories, ego-protective motives lead to systematically inflated recall, with the magnitude of distortion reflecting an optimal tradeoff between the psychological benefits of favorable memories and the cognitive costs of distorting reality. Second, this biased recall serves as the prior in Bayesian updating, which individuals combine with noisy signals about current performance to form posterior beliefs. Crucially, this updating process takes the biased recall at face value — individuals are unaware their memories might be distorted.

Building on models of motivated memory (Bénabou and Tirole, 2002) and affective decision making (Bracha and Brown, 2012; Caplin and Leahy, 2019), we formalize the recall process as follows. When recalling performance in subject j , the student faces the following maximization problem:

$$U_1 = \sum_j [w_j v(\tilde{P}_{past,j}) - \kappa(\tilde{P}_{past,j} - P_{past,j})^2], \quad (1)$$

where $P_{past,j}$ represents the true past performance, $\tilde{P}_{past,j}$ is the recalled performance, w_j captures

the ego-relevance weight for subject j , and $v(\cdot)$ is an increasing and concave function representing affective motivation. The term $\kappa(\tilde{P}_{past,j} - P_{past,j})^2$ represents the mental cost of holding distorted memories — the psychological effort required to justify beliefs that deviate from reality.

The first-order condition yields:

$$w_j v'(\tilde{P}_{past,j}) = 2\kappa(\tilde{P}_{past,j} - P_{past,j}) \quad (2)$$

Since $v'(\cdot) > 0$, this implies systematic overconfidence in recall:

$$\tilde{P}_{past,j} = P_{past,j} + b_j^*, \quad (3)$$

where $b_j^* = \frac{w_j v'(\tilde{P}_{past,j})}{2\kappa} > 0$ represents the optimal recall bias. This bias increases with ego-relevance (w_j) and decreases with mental costs (κ), consistent with motivated memory formation.

Prediction 1 (Overconfidence in recall): Individuals exhibit systematic overconfidence when recalling past performance. The magnitude of overconfidence is larger when individuals care more about the subject, when there are lower psychological costs of memory distortion, and when actual past performance is worse.

In the second stage, students form posterior beliefs by combining their recalled performance with new information through Bayesian updating. Based on their recalled mock exam score, they form a prior belief about current performance $P_j \sim N(\tilde{P}_{past,j}, \sigma_P^2)$, where σ_P^2 represents the uncertainty about how current performance relates to past performance. The signal s_j comes from self-grading with answer keys and scoring rubrics: students assess their performance but this assessment is imperfect, yielding a noisy signal $s_j = P_j + \varepsilon_j$ where $\varepsilon_j \sim N(0, \sigma_\varepsilon^2)$. The signal is unbiased on average (centered at true performance P_j) but contains error due to uncertainty about subjective questions like the essay, uncertainty about partial credit, misremembering specific answers, and difficulty in objective self-assessment. Given normally distributed prior and signal, Bayesian updating yields a posterior that is also normally distributed with mean:

$$\hat{P}_j = \alpha \cdot \tilde{P}_{past,j} + (1 - \alpha) \cdot s_j \quad (4)$$

where $\alpha = \frac{\sigma_\varepsilon^2}{\sigma_P^2 + \sigma_\varepsilon^2}$ represents the weight on the prior, increasing in signal noise and decreasing in prior uncertainty. The posterior variance is $\sigma_{post}^2 = \frac{\sigma_P^2 \sigma_\varepsilon^2}{\sigma_P^2 + \sigma_\varepsilon^2}$. While we assume normality for tractability, our

key prediction, that the posterior is a weighted average of prior and signal with weights summing to one, holds more generally for prior and signal error distributions satisfying the “Updating Toward the Signal” property (Chambers and Healy, 2012).

Crucially, this updating process takes the biased recall at face value. Thus unlike Bénabou and Tirole (2002), who assume sophistication about memory manipulation, but consistent with Bracha and Brown (2012) and Caplin and Leahy (2019), we assume that the student has no awareness that their prior formed through memory might be distorted.

To understand bias transmission, we calculate the expected deviation from actual performance:

$$E[\hat{P}_j - P_j] = \alpha \cdot E[\tilde{P}_{past,j}] + (1 - \alpha) \cdot E[s_j] - E[P_j] = \alpha(P_{past,j} - E[P_j]) + \alpha b_j^* \quad (5)$$

Since $b_j^* = \tilde{P}_{past,j} - P_{past,j} > 0$, Equation 5 shows that bias transmission occurs at rate $\alpha \in (0, 1)$, which is the same as the weight placed on the prior. This implies that there is stronger bias in the posterior when there is a larger weight on the prior, which happens when the signal is noisier or when the prior is more certain. Since the expected performance $E[P_j]$ in the entrance exam should equal to the true prior $P_{past,j}$, then $E[\hat{P}_j - P_j] = \alpha b_j^* > 0$.

We summarize the updating results below.

Prediction 2 (Bias Transmission through Bayesian Updating): Overconfidence from biased recall transmits to performance estimates through Bayesian updating. The transmission rate equals the weight placed on the prior.

4 Results: Overconfidence in recall

In the results section, we first examine whether students are overconfident when recalling past performance on the mock exam. As described in the background section, we asked students to try their best to recall their score in each of the six subjects in the mock exam. We assess their overconfidence in recall by comparing these recalled scores to their actual scores.

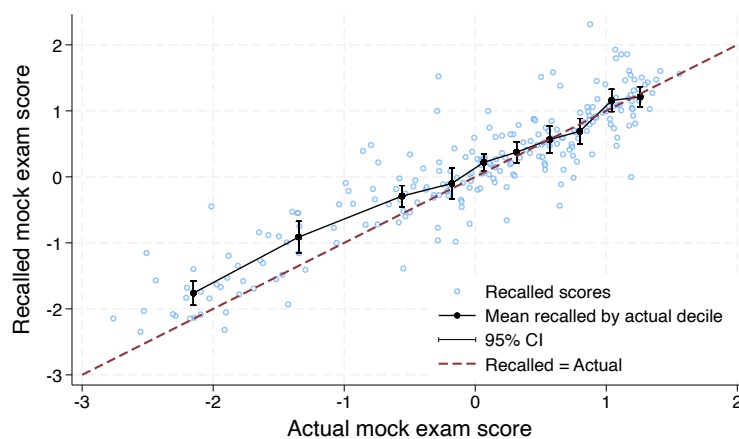
To benchmark our findings, we surveyed experts, primarily Chinese economists who are familiar with the high school admission system and had personal experience with exam score estimation, about their predictions for both recall and estimation accuracy. We posted the survey on the Social Science Prediction Platform and the Chinese experimental economists WeChat group (481 members)

and received 51 valid responses. After describing the study environment, we provided the mean and standard deviation of the mock exam scores and asked experts to predict the average error in recall.

Experts predicted weak if any overconfidence in the recall. On average, they predicted that the recalled score is 0.03 standard deviation higher than the actual mock exam score, which is not significantly different from 0 ($p = 0.119$, Wilcoxon signed-rank test).

We find much higher overconfidence in recall than what is predicted by the experts. Figure 2 displays the relationship between students' recalled mock exam total scores and their actual mock exam total scores. The 45-degree line represents accurate recall, with points above indicating overconfidence and points below indicating underconfidence. The graph reveals a systematic pattern of bias in students' recall of mock exam scores. The recalled scores are higher than the actual scores by 0.1 standard deviation on average, a difference that is statistically significant at the 1% level (Wilcoxon matched-pairs signed-ranks test). This pattern of recall appears systematic: while only 6% of students recalled their scores accurately, 59% recalled scores higher than their actual scores, and 35% recalled scores lower than their actual scores (see Appendix Figure A.1 for details). The substantially higher proportion of upward versus downward recall errors suggests that when students make mistakes in recalling their past performance, these mistakes are not randomly distributed around the true score but systematically biased upward.

Figure 2: Recalled vs. actual mock exam scores



Notes: This graph reports the actual and recalled total mock exam scores. The 45-degree line represents accurate recall of mock scores; points above the 45-degree line represent overconfidence in the recall and points below it represent underconfidence. The average recalled scores by deciles of the actual scores are also shown with 95% confidence intervals. Both actual and recalled mock scores are standardized with the mean and s.d. of actual mock exam scores. Thus 0 on the x-axis represents average performance in the mock exam and positive on the x-axis represents performing better than the average.

Overconfidence in recall at the subject level. We also observe a consistent pattern of overconfidence at the subject level. Students are overconfident at the subject level if we pool data for all six subjects, or if we look separately at individual subjects (see Appendix Figure A.2 for details). We find significant recall overconfidence in all subjects except for Math and Physics & Chemistry, thus recall overconfidence is not driven by overconfidence in a specific subject.

We also observe an interesting pattern not predicted by previous models of motivated memory in economics: students show significantly greater overconfidence in recalling non-STEM scores compared to STEM scores, even though all scores were originally presented in the same numerical format. This domain-specific variation in recall bias is consistent with our theoretical framework. STEM subjects, with their objective and precise nature, likely impose higher cognitive costs (κ) for memory distortion — it is psychologically more difficult to convince oneself that a clearly wrong mathematical answer was actually correct. In contrast, non-STEM subjects involve more subjective evaluation, making it easier to inflate recalled performance without experiencing the same cognitive dissonance. According to our conceptual framework, higher cognitive costs should lead to lower recall bias, which is exactly what we observe. To our knowledge, this is the first study in economics to document systematic domain-specific variation in motivated recall bias.

Overconfidence in recall and performance. Our conceptual framework predicts that recall bias should be decreasing in actual past performance. When performance is worse, the marginal ego utility from inflating recalled performance is higher, leading to greater overconfidence. This provides a strong test for motivated memory bias.

Figure 2 confirms this prediction. Students in the bottom five deciles of mock exam performance consistently recall scores higher than their actual ones, with overconfidence largest among the lowest performers and decreasing monotonically with performance. In contrast, students in the top five deciles show no systematic tendency toward overconfident recall, exhibiting both positive and negative recall errors around zero. This negative relationship between performance and recall bias holds across all subjects (see Appendix Table A.2 for details).

To rule out alternative explanations such as the Dunning-Kruger effect (Kruger and Dunning, 1999), we exploit within-individual variation across subjects. By including individual fixed effects, we control for student-specific factors like general meta-cognitive ability or memory sophistication. If motivated beliefs drive recall bias, students should show stronger upward bias in subjects where

they perform relatively poorly, as poor performance in any domain threatens their overall self-image. This within-person test isolates the motivational component from general cognitive limitations.

Table 1: Performance and bias in recall: Within-individual

Dependent variable	Overconfidence in recall		
	(1) Full	(2) High performers	(3) Low performers
Actual mock score	-0.288*** (0.033)	-0.439*** (0.054)	-0.277*** (0.048)
Individual fixed effect	Yes	Yes	Yes
Subject fixed effect	Yes	Yes	Yes
Observations	1478	745	733
R-squared	0.399	0.445	0.380

Notes: This table reports the relationship between bias in recall and performance in the mock exam. The dependent variable “Overconfidence in recall” is defined as the difference between recalled and actual mock exam scores. The independent variable “Actual mock score” is the actual mock exam score in a subject. The observation unit is student-subject. All specifications include individual and subject fixed effects to control for student-level and subject-specific factors that might affect recall accuracy. Column (1) presents results for the full sample, while columns (2) and (3) present results for students who scored above and below the median in the mock exam, respectively. Standard errors, reported in parentheses, are clustered at the student level to accommodate multiple subjects per student. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 1 presents evidence supporting the motivated beliefs explanation. Column (1) shows that within individuals, when a student performs one point better in one subject compared to another, they show 0.288 point less overconfidence in recall for that subject. This negative relationship between relative performance and recall bias suggests that students try to maintain a consistent self-image across subjects: when they perform relatively poorly in one subject compared to their other subjects, they are more likely to inflate their recalled performance in that subject.⁴ This pattern is particularly pronounced among high-performing students: they show a 0.439 point decrease in recall bias for every point increase in relative performance, compared to 0.277 for low-performing students. The stronger

⁴The within-individual pattern of stronger recall bias in weaker subjects can be understood through the lens of how memory processes complex information. Graeber et al. (2024) demonstrate that people better remember simplified stories than specific numerical details. In our context, students may form a simplified narrative about their overall performance level — for instance, “I scored around 80 in all subjects in the mock exam” — rather than maintaining precise memories of heterogeneous scores across subjects (e.g., 87 in Mathematics, 73 in English). This memory compression implies that when recalling subject-specific scores, students anchor on their simplified overall performance narrative and insufficiently adjust for subject-specific variation. Consequently, recall bias is strongest in subjects where actual performance deviates most from this internalized narrative: students who performed relatively poorly in a subject unconsciously adjust their recalled score upward toward their perceived average, while adjust their recall downward when the subject performance is above the average.

effect among high performers suggests that they are more invested in maintaining a consistent self-image across subjects, perhaps because performing poorly in any subject poses a greater threat to their overall identity as a strong student. This heightened sensitivity to relative performance among high performers is inconsistent with the Dunning-Kruger effect, which would predict that high performers should be more accurate in assessing their relative strengths and weaknesses across subjects.

Error in recall and ego-relevance. Our conceptual framework also predicts that recall bias should increase with ego-relevance: individuals should exhibit greater overconfidence in domains they care more about. We test this prediction by exploiting gender differences in subject-specific ego-relevance. Previous research suggests that STEM subjects are more ego-relevant for male students, while both genders care equally about non-STEM subjects (Bordalo et al., 2019; Coffman et al., 2024). If ego-relevance affects recall bias, we should observe a larger gender gap in overconfidence for STEM subjects compared to non-STEM subjects.

The data support this prediction. Male students exhibit significantly more overconfidence in recall compared to female students, but this gender gap varies systematically across domains. In STEM subjects, male students' recalled scores are 0.14 standard deviation higher than female students' given the same actual mock exam score. In non-STEM subjects, this gender gap shrinks to 0.09 standard deviation.

Controlling for individual fixed effects, Appendix Table A.3 shows that female students are significantly less overconfident in STEM subjects compared to non-STEM subjects, while male students show no significant difference in recall bias across the two domains. The interaction analysis confirms that male students are significantly more overconfident than female students in STEM versus non-STEM subjects, with the interaction effect significant at the 1% level. This pattern supports our ego-relevance hypothesis that individuals exhibit greater recall bias in domains they care more about.

Selection into recall: evidence for motivated forgetting. While our main analysis focuses on patterns in recalled scores, we acknowledge that the 18.4% missing recall data may not be random. This non-response can be interpreted as analogous to the “do not recall” option in classic memory tasks. We therefore analyze whether this non-recall exhibits systematic patterns consistent with motivated memory processes.

We examine predictors of not providing recall data at two levels. At the individual level, we estimate regressions where the dependent variable is an indicator for whether a student provides no recall

data for any subject, testing whether these students differ systematically from those who recall at least one subject. We find that students who performed better on the entrance exam relative to the mock exam are significantly more likely to not provide any recall. Specifically, in regressions controlling for demographics and risk preferences, a higher entrance exam score increases the probability of non-recall, while a higher mock exam score decreases it. This pattern, whereby students who improved from mock to entrance are less likely to recall, is robust whether we use estimated entrance scores (what students knew when deciding whether to recall) or actual entrance scores (addressing potential endogeneity concerns because estimations and recalls could be determined at the same time). See Appendix Table A.4 for detailed regression results.

The within-individual analysis examines students who recall some but not all subjects. We estimate regressions with individual fixed effects, where the dependent variable is an indicator for whether a student provides no recall data for a particular subject. This specification tests which subjects students selectively forget while controlling for all student-level characteristics. We find that students are more likely to provide no recall for subjects where they estimated higher entrance exam scores but had lower mock scores, suggesting selective forgetting when current performance exceeds past performance (see Appendix Table A.4 for details).

These patterns suggest that non-recall represents a form of motivated forgetting rather than random non-response. When current reality surpasses disappointing past performance, complete forgetting may serve the same ego-protective function as biased recall, representing an alternative strategy for managing self-image, in addition to recall with overconfidence.

To further address concerns about selection bias from missing recall data, we examine a conservative lower bound for overconfidence in recall. We replace all missing recall observations with the assumption of accurate recall (overconfidence = 0), effectively assuming that non-responders would have perfectly recalled their mock exam scores if forced to respond. This is a highly conservative assumption given that students who strategically forget disappointing performance likely exhibit similar motivated reasoning patterns as those who recall with positive bias. Even under this strict assumption, we find significant overconfidence in recall: the mean overconfidence is 0.091 standard deviation ($P < 0.01$). This lower bound analysis demonstrates that our main finding of motivated overconfidence in recall is robust: it would only be overturned if non-responders exhibited strong systematic underconfidence, recalling scores substantially below their actual performance. Given the motivated

forgetting patterns we document among non-responders (who preferentially forget disappointing performance), such systematic underconfidence seems implausible.

Taken together, we find significant overconfidence in students' recall of mock exam scores, a pattern that contrasts with experts' prediction of nearly unbiased recall. Three key patterns support the motivated nature of this bias. First, students with worse actual performance exhibit greater overconfidence, with this relationship holding both between and within individuals across subjects. Second, recall bias is significantly larger in non-STEM subjects compared to STEM subjects, consistent with lower cognitive costs of memory distortion in more subjective domains. Third, male students are more overconfident than female students, particularly in STEM subjects with higher ego-relevance.

5 Accuracy in Estimation

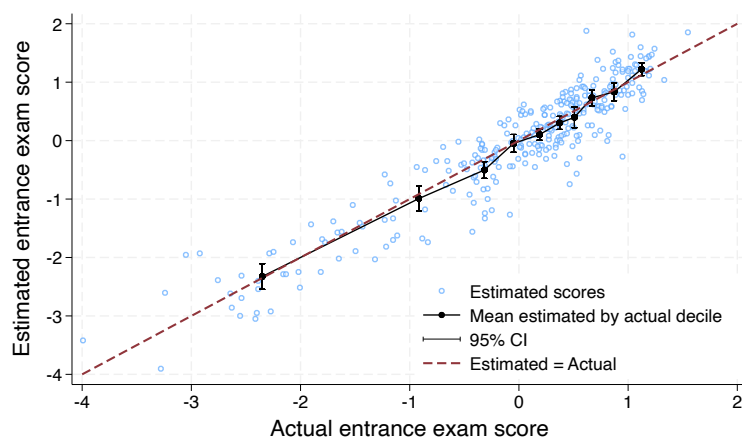
We have shown that students are overconfident in the low-stakes environment of recalling mock exam scores. We now test whether they remain overconfident when the stakes are high in estimating the entrance exam scores. As described earlier, we surveyed 51 experts about their predictions for both recall and estimation accuracy. For high-stakes estimation, experts predicted significant overconfidence, with an average estimation error of 0.13 standard deviation, significantly different from zero at the 1% level.

Figure 3 plots the relationship between estimated and actual entrance exam scores. In this high-stakes environment, the students' estimated scores are, on average, 0.04 standard deviation lower than their actual scores. The difference between the two is not significantly different from 0 (p -value = 0.282, Wilcoxon matched-pairs signed-ranks test). Thus we do not find overconfidence on average for the high-stakes estimation.

The distribution of errors in estimation is also more balanced than errors in recall (see Appendix Figure A.3). Only 2% of students accurately predict their score in the entrance exam. However, there is no clear tendency towards overestimation: 146 (48%) students over-estimated their score and 149 (49%) students under-estimated their score.

Estimation bias at the individual level could be due to overconfidence in some subjects and underconfidence in others. To explore this possibility, we also examined estimation accuracy at the subject level and find no evidence of overconfidence. Pooling all subject exam estimations together, the estimated scores are, on average, 0.02 standard deviation lower than the actual scores. Students

Figure 3: Estimated vs. actual entrance exam scores



Notes: This graph reports the actual and estimated total entrance exam scores. The 45-degree line represents accurate estimation of entrance exam scores; points above the 45-degree line represent overconfidence in the estimation and points below it represent underconfidence. The average estimated scores by decile of the actual scores are also shown with 95% confidence interval. Both actual and estimated scores are standardized with the mean and s.d. of actual entrance exam scores. Thus 0 on the x-axis represents average performance in the entrance exam and positive on the x-axis represents performing better than the average.

are slightly underconfident in non-STEM subjects, and are neither over- or underconfident in STEM subjects. Score estimations are largely accurate for all six subjects, with no systematic patterns for being over- or under-confident (see Appendix Figure A.4 for details).

Error in estimation and exam performance. While worse performance is associated with more overconfidence in recall both between and within students, we find a different pattern with estimation errors. Figure 3 shows that low-performing students in the bottom five deciles are, on average, either accurate or slightly underconfident in estimation, which contrasts sharply with their overconfidence in recall. Subject-level analysis indicates that the relationship between performance and estimation bias varies substantially across subjects: we find no significant relationship in Math, Physics & Chemistry, and History & Politics, but significant negative relationships in Chinese, English, and Geography & Biology (See Appendix Table A.5 for details). This variation across subjects stands in contrast to the recall task, where we observe consistently negative relationships across all subjects. However, it is important to note that unlike in recall where actual scores were once known for sure, estimation errors are not necessarily mistakes. As shown in our conceptual framework, students place more weight on their prior beliefs when signals about exam performance are noisier, implying that estimated scores should not perfectly reflect actual performance.

We then examine the relationship between performance and estimation errors within individuals

by regressing errors in estimation on actual entrance exam scores while controlling for individual fixed effects. Table 2 shows that within individuals, when a student performs one point higher in one subject compared to another, they show 0.234 point less overconfidence in their estimation for that subject. Similar to the pattern in recall, this relationship is particularly strong among high-performing students (coefficient = -0.453) compared to low-performing ones (coefficient = -0.242). While students show no systematic bias in their average estimations across all subjects, they tend to be overconfident in their relatively weak subjects and underconfident in their stronger ones. This pattern could reflect either persistent motivational forces in belief formation or rational Bayesian updating where students place positive weight on their prior beliefs. To distinguish between these explanations, we analyze the supply side of belief formation in Section 6 by examining how students combine their prior beliefs with new information.

Table 2: Performance and overconfidence in estimation

Dependent Variable	Overconfidence in estimation		
	(1) Full	(2) High performers	(3) Low performers
Actual entrance exam score	-0.234*** (0.028)	-0.453*** (0.045)	-0.242*** (0.034)
Observations	1812	900	912
R-squared	0.519	0.628	0.489

Notes: This table reports the relationship between overconfidence in estimation and performance in the entrance exam. The dependent variable “Overconfidence in estimation” is defined as the difference between estimated and actual entrance exam scores. The independent variable “Actual entrance exam score” is the actual entrance exam score in a subject. The observation unit is student-subject. All specifications include individual and subject fixed effects to control for student-level and subject-specific factors that might affect estimation errors. Column (1) presents results for the full sample, while columns (2) and (3) present results for students who scored above and below the median in the entrance exam, respectively. Standard errors, reported in parentheses, are clustered at the student level to accommodate multiple subjects per student. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Error in estimation and ego-relevance. Having established that ego-relevance influences recall bias, we now test whether these effects persist under high stakes. Our conceptual framework predicts they should: since biased recall serves as the prior in belief formation, gender differences in recall bias should transmit through Bayesian updating into estimation bias. Specifically, we should observe a larger gender gap in estimation bias for STEM versus non-STEM subjects.

We find evidence consistent with this prediction. Male students exhibit more overconfidence in estimation compared to female students, and this gender gap varies systematically across domains. In STEM subjects, male students’ estimated scores are 0.14 standard deviation higher than female

students' given the same actual entrance exam score (significantly different from 0 at the 1% level). In non-STEM subjects, this gender gap is only 0.04 standard deviation and not significantly different from 0 ($p = 0.587$). The gender gap in estimation bias is significantly larger in STEM subjects, with an interaction effect of 0.233 standard deviation (significant at the 1% level, see Appendix Table A.6 for details). This confirms that ego-relevance effects persist and transmit into high-stakes beliefs. We validate the transmission mechanism directly in Section 6 when examining the supply side of belief formation.

Stakes and mistakes. Our study presents a high-stakes environment where estimation errors can have significant consequences. In the academic year we studied, the average difference in admission cutoffs among the five Tier-1 high schools was approximately 0.3 standard deviation of the exam score. This narrow margin implies that even small estimation errors can substantially affect students' school placement outcomes.

While we do not observe students' actual applications, we can assess potential mistakes by assuming students apply based on their estimated scores. Specifically, we first identify the highest-ranked school (in terms of cutoff scores) a student could attend based on their actual score (the ex-post optimal choice). We then compare this to the highest-ranked school they could attend based on their estimated score (their likely choice if applying according to their estimates). We define the mistake as the difference between these two school ranks, where negative values indicate overconfidence (likely applying to overly competitive schools) and positive values indicate underconfidence. To assess the role of stakes in estimation accuracy, we compare this distribution of mistakes to a counterfactual scenario where students exhibit the same degree of overconfidence in estimation as observed in their recall.

This analysis yields three key findings about potential mistakes in school choice. First, consistent with the lack of systematic bias in estimation, students demonstrate remarkable accuracy: 82% would apply to schools that match their ex-post optimal choice based on their actual scores. Second, the distribution of potential mistakes is relatively symmetric, with similar frequencies of applying to schools that are too selective (10%) or not selective enough (8%). This balance suggests no systematic tendency toward over- or under-shooting in school choice. Third, in a counterfactual scenario where students exhibit the same degree of overconfidence in estimation as they do in recall, the model predicts substantially more mistakes, particularly in the direction of applying to overly selective schools.

The share of students who would apply to schools beyond their qualification would increase from 10% to 25%. The stark contrast between these scenarios, with actual estimation errors being both smaller and more balanced, suggests that students respond to high stakes by forming more accurate beliefs than their recall errors would predict. See Appendix Figure A.5 for details of the distribution of mistakes.

Despite the stakes being generally high, there are two groups of students for whom the stakes are lower. The first group includes top-performing students who have effectively secured a seat at their preferred school. We define this group as the top decile of students at each school. As shown in Figure 3, the top decile students overestimated their scores by 0.1 standard deviation and the overestimation is significantly different from 0 ($p = 0.059$).⁵

The second group consists of low-performing students who did not meet the cutoff for Tier-2 high schools. There were 24 students who fall into this category in our sample. On average, they overestimated their scores by 0.04 standard deviation. However, this difference is not significantly different from 0 ($p = 0.861$).

Social image concerns. While our main analysis focuses on self-image concerns driving motivated beliefs, social image concerns could also influence how students report their estimated scores. Students might overestimate to impress others (parents, teachers, or peers) or underestimate to avoid the embarrassment of falling short of inflated predictions. The direction of social image effects is theoretically ambiguous.

To investigate the role of social image concerns, we exploit a unique feature of our data collection. One of the two middle schools required students to officially report their estimated scores to teachers for coordination purposes. This creates a high social-image context where scores are shared with authority figures who know the students personally. In contrast, our survey explicitly minimized social image concerns: we identified ourselves as external researchers, assured students their responses would be anonymized, and promised not to share individual scores with parents, teachers, or classmates.

⁵The highest ranked high school admits about 15% of students in each middle school. However, we picked a cutoff of 10% because students who ranked between 10% and 15% might not be sure about their admissibility to their preferred school. The result is qualitatively consistent if we instead look at the top 15% students.

Comparing estimations across these two reporting contexts provides suggestive evidence about the importance of social image concerns. If such concerns substantially affect reported beliefs, we should observe systematic differences between scores reported to teachers versus those reported to anonymous researchers. We find that the distributions of estimation errors are remarkably similar across the two contexts. Students who officially reported to their school showed average overconfidence of 0.01 standard deviation, virtually identical to the -0.04 standard deviation in our survey data. A comparison of means shows no significant difference ($p = 0.680$, Wilcoxon ranksum test). Appendix Figure A.6 further shows that the distributions of estimation errors are remarkably similar across the two contexts. This similarity suggests that social image concerns do not substantially affect score estimation in our setting.

Robustness: Estimation patterns in the Recall sample. One potential concern about our results is that the contrast between overconfidence in recall and accurate estimation could be driven by sample differences due to missing recall data. To address this, we reproduce our estimation analysis using only the recall sample, namely students who provided at least one recall response. With this restricted sample, estimated scores are 0.06 standard deviation lower than actual scores, representing slight underconfidence that is significantly different from zero at the 5% level ($p = 0.044$). Importantly, we still find no evidence of overconfidence in estimation even when restricting to the recall sample. We replicate all analyses from this section using the recall sample, with results reported in Appendix C. The key patterns remain qualitatively unchanged. The contrast between significant overconfidence in recall and accurate or slightly underconfident estimation thus holds within the same sample of students, ruling out sample selection as an explanation for our main findings.

6 The Supply Side of Biased Beliefs

Having established that students exhibit motivated overconfidence in recall but achieve accurate beliefs in estimation, we now examine what mechanisms make motivated beliefs malleable. The absence of overconfidence in estimation despite biased recall suggests some form of adjustment to environmental pressures, but the specific mechanism remains unclear. One possibility is that students recognize the connection between recall and estimation accuracy, making recall itself high-stakes. If students understand they will rely on recalled performance when forming consequential estimates, they might strive for accurate recall even at the cost of ego utility. However, we have already shown

students remain significantly overconfident when recalling mock exam scores, with no evidence of correcting these biased memories.

A second possibility is that students shut down the supply-side channel from biased recall to biased estimation. Even if they cannot correct their motivated memories directly (perhaps due to the ego utility these inflated memories provide), they might recognize the importance of accuracy and switch to using objective information when forming estimates. After all, students had access to their actual mock exam scores through official records. If this mechanism operates, we should observe students relying on actual rather than recalled performance when forming estimates, effectively shutting down the biased memory channel. To test whether students shut down this supply-side channel, we examine how they actually combine information when forming estimates — specifically, whether they update using biased recall as the prior or more objective information.

Bayesian updating and belief formation. To test whether students follow Bayesian updating, we estimate the following regression specification separately for each subject j :

$$\hat{P}_{ij} = \beta_{1j}\tilde{P}_{past,ij} + \beta_{2j}s_{ij} + \mathbf{X}'_i\gamma_j + \varepsilon_{ij} \quad (6)$$

, where \hat{P}_{ij} is student i 's estimated score for subject j , $\tilde{P}_{past,ij}$ is their recalled mock exam score (prior), s_{ij} is their entrance exam score (signal), \mathbf{X}_i includes individual characteristics such as gender and class fixed effects, and ε_{ij} is the error term. Our conceptual framework predicts that the sum of the weights on the prior and the signal should equal to 1: $\beta_1 + \beta_2 = 1$ (Equation (4)).

Panel A of Table 3 tests this prediction by examining how students form their score estimates across all six subjects. The evidence strongly supports Bayesian updating. In all subjects, the estimated weights on recalled mock scores (β_1) and entrance exam scores (β_2) sum remarkably close to 1, ranging from 0.953 in English to 1.039 in Physics & Chemistry. Formal tests of the restriction $\beta_1 + \beta_2 = 1$ yield p-values ranging from 0.079 to 0.939, indicating that we cannot reject the null hypothesis of Bayesian updating in any subject at 5% level.

Panel B of Table 3 provides evidence that students rely on biased memory rather than objective information when forming beliefs. Even when controlling for actual mock exam scores, which students had easy access to when making their estimates, recalled mock scores remain highly significant predictors across all subjects. In stark contrast, the coefficient on actual mock exam scores are largely insignificant and economically small for most subjects, with only History & Politics showing a sig-

nificant coefficient at 5% level. This pattern demonstrates that students systematically choose to base their estimates on potentially biased recalled performance rather than on the objective record of their past performance. The persistence of recalled mock scores' influence, despite the availability of accurate information, supports our framework's prediction that biased memory serves as a supply-side mechanism for overconfident beliefs. Interestingly, this supply-side mechanism still operates in the high-stakes score estimation.

Table 3: The supply side of overconfident beliefs by subject

Dependent variable	Estimated score					
	Math (1)	Chinese (2)	English (3)	Phy&Che (4)	His&Pol (5)	Geo&Bio (6)
Panel A: Bayesian updating with recalled mock scores						
Recalled mock score (β_1)	0.427*** (0.055)	0.603*** (0.054)	0.430*** (0.050)	0.251*** (0.048)	0.410*** (0.054)	0.499*** (0.066)
Entrance exam score (β_2)	0.597*** (0.051)	0.399*** (0.055)	0.523*** (0.051)	0.788*** (0.048)	0.622*** (0.064)	0.470*** (0.063)
Sum ($\beta_1 + \beta_2$)	1.024	0.996	0.953	1.039	1.032	0.970
P-value ($\beta_1 + \beta_2 = 1$)	0.555	0.939	0.079	0.292	0.590	0.574
R-squared	0.856	0.746	0.906	0.885	0.733	0.688
Panel B: Role of biased memory						
Recalled mock score	0.419*** (0.063)	0.560*** (0.061)	0.386*** (0.058)	0.186*** (0.054)	0.323*** (0.056)	0.487*** (0.076)
Actual mock score	0.015 (0.067)	0.113 (0.076)	0.104 (0.071)	0.179* (0.074)	0.314*** (0.076)	0.030 (0.091)
Entrance exam score	0.592*** (0.056)	0.351*** (0.064)	0.476*** (0.060)	0.692*** (0.062)	0.466*** (0.072)	0.457*** (0.075)
Observations	249	244	248	245	246	246
R-squared	0.856	0.749	0.907	0.888	0.752	0.689

Notes: This table reports how students combine their recalled mock exam scores and the perceived entrance exam performance when forming their estimated scores by subject. Panel A shows the basic Bayesian updating regression. Panel B adds actual mock exam scores to test whether students rely on biased recall even when controlling for objective past performance. The dependent variable is the estimated entrance exam score. The observation unit is student-subject. Columns (1) through (6) present results separately for Mathematics, Chinese, English, Physics & Chemistry, History & Politics, and Geography & Biology, respectively. Standard errors are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Transmission of bias in recall to bias in estimation. Our conceptual framework predicts that overconfidence from biased recall should transmit to performance estimates through the Bayesian updating process. From Equation 5, the expected overconfidence in the posterior belief is:

$$E[\hat{P}_j] - E[P_j] = \alpha(P_{past,j} - E[P_j]) + \alpha b_j^*.$$

This equation decomposes estimation bias into two components: the first term captures rational updating based on performance changes between past and current exams, while the second term represents the transmission of motivated memory bias. The parameter α reflects the weight placed on the prior in Bayesian updating, which should equal the bias transmission rate.

To estimate these components empirically, we specify the following regression for each subject:

$$\text{Estimation Bias}_{ij} = \alpha_1(\text{Mock Score} - \text{Entrance Score})_{ij} + \alpha_2(\text{Bias in Recall})_{ij} + \mathbf{X}'_i\boldsymbol{\gamma}_j + \varepsilon_{ij}, \quad (7)$$

where Estimation Bias is the difference between estimated and actual entrance exam scores, Mock Score - Entrance Score captures performance changes between exams, and Bias in Recall measures the overconfidence in recalled mock exam scores. The coefficient α_1 measures how students rationally update their beliefs based on performance surprises, while α_2 captures the transmission rate of motivated memory bias.

To test whether the transmission of bias operates purely through the Bayesian updating channel, we conduct formal tests comparing the coefficients α_1 (on mock-entrance score difference) and α_2 (on bias in recall) to the Bayesian weight α estimated from our baseline specification. Table 4 presents these test results. We cannot reject the hypothesis that both coefficients equal the Bayesian α in most subjects, with p-values for the joint test ranging from 0.142 to 0.956, except for History & Politics where we reject the null at the 1% level. This indicates that in most domains, the transmission of bias from recall to estimation follows the predictions of Bayesian updating.

Despite the lack of statistical significance in most subjects, we observe systematic patterns in how the transmission coefficients deviate from the Bayesian weights. The coefficient on bias in recall (α_2) tends to be lower than the Bayesian α across subjects. One way to interpret this tendency is that students partially discount their biased memories when forming high-stakes estimates. Conversely, the coefficient on the performance change (α_1) generally exceeds the Bayesian α , suggesting that students place more weight on the prior when the entrance exam score deviates from the mock exam score. This pattern is consistent with a conservatism bias in belief updating, where individuals under-weight new information relative to priors, especially when the signal is precise (Benjamin, 2019; Ba et al., 2024; Augenblick et al., 2025). The lack of statistical significance in these deviations could reflect insufficient power rather than true equivalence as detecting subtle differences in updating

weights may require larger samples.

Table 4: The transmission of bias in recall to bias in estimation by subject

Dependent variable	Estimated – Actual entrance score					
	Math (1)	Chinese (2)	English (3)	Phy&Che (4)	His&Pol (5)	Geo&Bio (6)
Mock – Entrance score: α_1	0.414*** (0.056)	0.657*** (0.061)	0.518*** (0.060)	0.324*** (0.061)	0.577*** (0.068)	0.538*** (0.074)
Bias in recall: α_2	0.410*** (0.062)	0.553*** (0.059)	0.386*** (0.058)	0.167** (0.053)	0.303*** (0.055)	0.497*** (0.073)
Bayesian α	0.427	0.603	0.430	0.251	0.410	0.499
P-val: $\alpha_1 = \alpha$	0.789	0.392	0.448	0.113	0.054	0.974
P-val: $\alpha_2 = \alpha$	0.817	0.379	0.146	0.230	0.015	0.607
P-val: Joint test	0.956	0.360	0.142	0.090	0.001	0.861
Observations	249	244	248	245	246	246
R-squared	0.304	0.460	0.382	0.200	0.328	0.358

Notes: This table reports how bias in recall transmits to bias in estimation by examining the decomposition of estimation errors. The dependent variable is “Estimated – Actual entrance score,” defined as the difference between estimated and actual entrance exam scores. “Mock – Entrance score” (α_1) is the difference between the two standardized scores, and “Bias in recall (α_2)” is the difference between recalled and actual mock exam scores. “Bayesian α ” reports the weight on recalled mock scores from the Bayesian updating regression in Table 3 (called β_1 in Table 3). The p-values test whether each transmission coefficient equals the Bayesian weight, with “Joint test” testing the null hypothesis that both $\alpha_1 = \alpha$ and $\alpha_2 = \alpha$. All specifications include class fixed effects and gender controls. The observation unit is student-subject. Columns (1) through (6) present results separately for Mathematics, Chinese, English, Physics & Chemistry, History & Politics, and Geography & Biology. Standard errors, reported in parentheses, are clustered at the student level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

7 Final belief adjustment

Even though we find no overconfidence on average in high-stakes estimation, our analysis indicates that the supply-side channel of motivated biased memory still operates. Students remain overconfident in recall and continue to rely on this biased recall when forming estimates. The puzzle, then, is how this mechanism can be active without producing overconfidence in the final estimation under high stakes.

One potential resolution to this puzzle involves conscious adjustments to posterior beliefs. We develop a framework where individuals possess partial meta-awareness: they sense a general tendency toward overconfidence without understanding its source. Unable to identify that biased memory drives their overconfidence, they cannot correct the root cause but instead can make direct adjustments to their final beliefs.

In this framework, individuals with partial meta-awareness choose final beliefs to maximize ex-

pected utility, balancing three considerations: ego utility from optimistic beliefs, costs of decision mistakes, and cognitive effort required to adjust beliefs. When stakes are high, the cost of mistakes dominates, leading to systematic downward adjustments. Formally, if individuals receive posterior \hat{P}_j from Bayesian updating, the expected utility from choosing final belief P_j^* is:

$$E[U] = \sum_j [w_j v(P_j^*) - c \cdot E[L(P_j^* - P_j)|\hat{P}_j] - \kappa_2(P_j^* - \hat{P}_j)^2] \quad (8)$$

where c represents the stakes parameter (weight on decision costs), $L(\cdot)$ is the loss function from decision mistakes, and $\kappa_2(P_j^* - \hat{P}_j)^2$ represents the cognitive cost of consciously adjusting beliefs away from the biased Bayesian posterior. The expectation $E[L(P_j^* - P_j)|\hat{P}_j]$ reflects the individual's uncertainty about true performance given their posterior belief.

For quadratic loss from decision mistakes, $L(x) = x^2$, the first-order condition becomes:

$$\frac{w_j v'(P_j^*)}{2} = c(P_j^* - E[P_j|\hat{P}_j]) + \kappa_2(P_j^* - \hat{P}_j) \quad (9)$$

Let $\gamma = P_j^* - \hat{P}_j$ denote the correction term. Substituting and rearranging:

$$\frac{w_j v'(P_j^*)}{2} = c(\hat{P}_j - E[P_j|\hat{P}_j]) + (c + \kappa_2)\gamma \quad (10)$$

Solving for γ :

$$\gamma = \frac{1}{c + \kappa_2} \left[\frac{w_j v'(P_j^*)}{2} - c(\hat{P}_j - E[P_j|\hat{P}_j]) \right] \quad (11)$$

Given students' meta-awareness of overconfidence, we have $\hat{P}_j - E[P_j|\hat{P}_j] > 0$. As stakes become very high ($c \rightarrow \infty$):

$$\lim_{c \rightarrow \infty} \gamma = \lim_{c \rightarrow \infty} \frac{1}{c + \kappa_2} \left[\frac{w_j v'(P_j^*)}{2} - c(\hat{P}_j - E[P_j|\hat{P}_j]) \right] = -(\hat{P}_j - E[P_j|\hat{P}_j]) < 0 \quad (12)$$

We thus reach the following prediction:

Prediction 3 (Negative Correction under High Stakes): When decision stakes are high, individuals with partial meta-awareness of their overconfidence tendency make systematic downward adjustments to their Bayesian posterior beliefs.

Empirical estimation of the final correction term. According to our framework, students make a final conscious correction to their Bayesian posterior \hat{P}_j to arrive at their final estimate P_j^* . This cor-

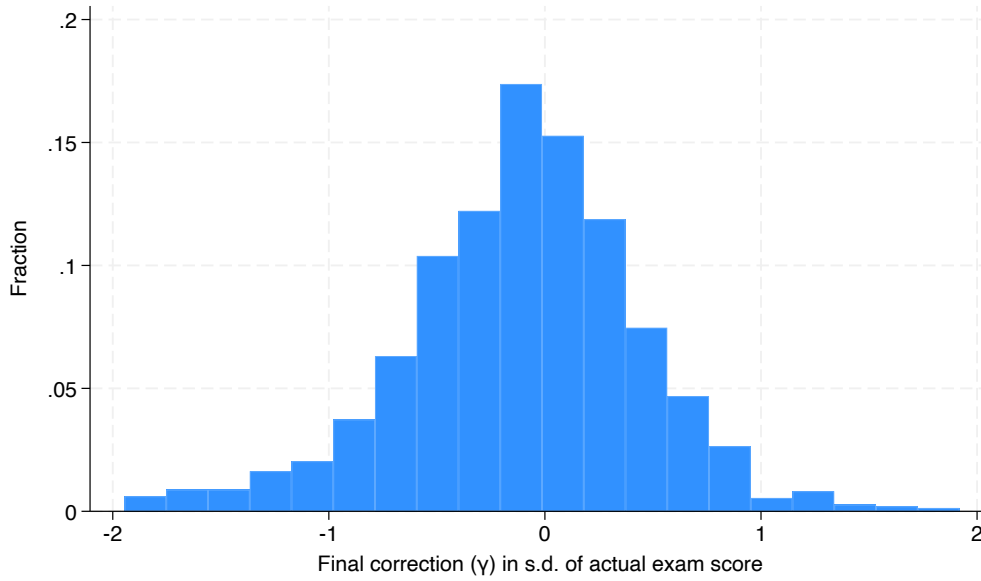
rection, driven by partial meta-awareness of overconfidence tendencies and the high stakes involved, is captured by $\gamma = P_j^* - \hat{P}_j$. To empirically identify this correction term, we estimate Equation 6 while imposing the Bayesian constraint $\beta_{1j} + \beta_{2j} = 1$ for each subject j . Under this constraint, the coefficients represent pure Bayesian weights, and any systematic deviation from the Bayesian posterior appears in the residual. Thus, we estimate γ_{ij} as the residual term from the constrained regression.

An empirical challenge is that the residual contains both the correction term γ_{ij} and other components v_{ij} such as measurement error, idiosyncratic shocks, and cognitive noise. We cannot directly separate γ_{ij} from v_{ij} , because we only observe their sum as the regression residual ε_{ij} . However, if the residuals primarily reflect final corrections rather than random noise, they should exhibit systematic patterns consistent with our theoretical predictions. Crucially, while OLS residuals from an unconstrained regression mechanically have mean zero, residuals from our constrained regression can have a non-zero mean if students systematically deviate from Bayesian updating. Finding average correction $\bar{\gamma} < 0$ would indicate downward adjustments, as predicted by our model when stakes are high. Additionally, the residuals should vary with stakes (less negative for top performers who face lower admission uncertainty). We therefore examine whether the empirical residuals display these predicted patterns to validate our interpretation that they capture meaningful belief adjustment behavior rather than merely statistical noise.

Figure 4 shows the distribution of the final correction term, γ . Students are more likely to adjust their estimated scores down, with 57% of adjustments being downwards. As a result, on average, students adjust their estimation down by 0.12 standard deviation of the actual exam score, which is significantly different from 0 ($p < 0.01$). The non-zero mean of the adjustment term is consistent with our theoretical prediction, suggesting that it is more than statistical noise. Interestingly, the size of the final correction is similar to the size of overconfidence in recall, which is consistent with supply-side factors still playing a role in generating the posterior, but the average belief being corrected through the final adjustment so that we do not observe average overconfidence.

To further show the final correction term γ represents a meaningful belief adjustment by students, we test the relationship of γ with stakes faced by students. Figure 5 displays how γ varies across entrance exam score deciles. We observe that students across the first nine deciles consistently make negative corrections to their estimated scores, with magnitudes ranging from -0.21 to -0.08 standard deviation. In stark contrast, students in the top performance decile make positive corrections of ap-

Figure 4: Final correction term: distribution



Notes: This graph reports the distribution of the final correction term γ , standardized by the actual exam score. Here a positive value means adjusting the estimation up, and a negative value means adjusting the estimation down.

proximately 0.03 standard deviation. This pattern is consistent with our theoretical framework: the highest-performing students face substantially lower stakes due to their near-certain admission to preferred schools, and are consequently the only group exhibiting overconfidence in their final estimates. Importantly, among the remaining 90% of students who face uncertainty about admission outcomes, we observe no systematic relationship between performance and correction magnitude. This suggests that the correction term captures conscious responses to stakes rather than a general correction tendency related to ability level.

Confidence intervals and the correction mechanism. Our analysis so far assumes students have only general awareness of overconfidence. We now extend this to examine what happens when students possess more sophisticated meta-awareness. Specifically, what happens when they understand that bias originates from the prior and increases with the prior’s weight in Bayesian updating.

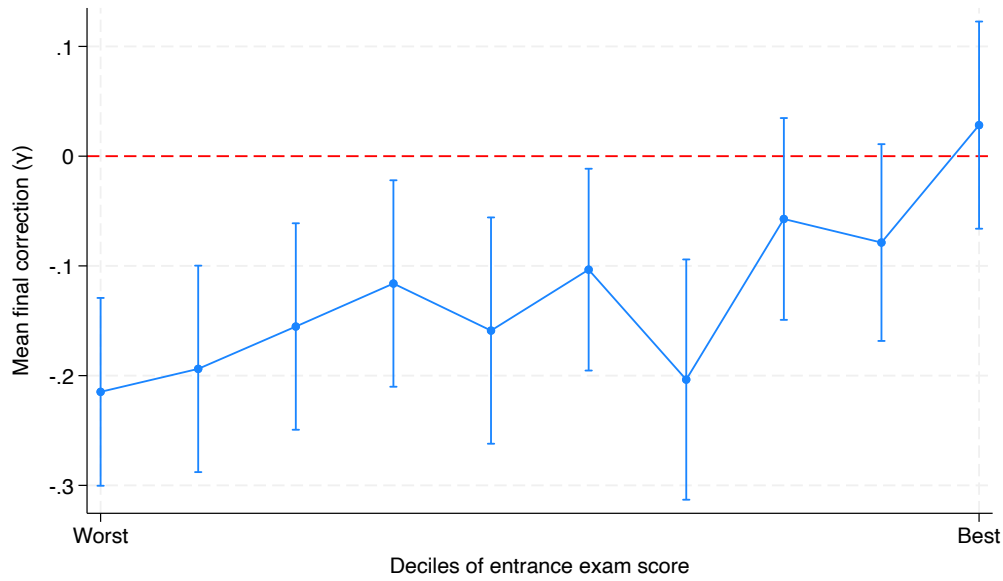
We first establish the relationship between bias and uncertainty in the posterior. From Bayesian updating, we know:

$$\alpha = \frac{\sigma_{\epsilon}^2}{\sigma_P^2 + \sigma_{\epsilon}^2} \quad \text{and} \quad \sigma_{post}^2 = \frac{\sigma_P^2 \sigma_{\epsilon}^2}{\sigma_P^2 + \sigma_{\epsilon}^2}$$

These can be combined to show: $\alpha = \frac{\sigma_{post}^2}{\sigma_P^2}$.

Since the transmitted bias is αb_j^* , and the posterior variance increases with signal noise, both

Figure 5: Performance and the final correction term



Notes: This graph reports the relationship of the final correction term γ to deciles of total entrance exam scores. A positive value of γ means adjusting the estimation up, and a negative value means adjusting the estimation down. Error bars indicate 95% confidence intervals.

the transmitted bias and posterior variance increase with signal noise. The transmitted bias can be expressed as: $\frac{\sigma_{post}^2}{\sigma_p^2} b_j^*$. Thus, a higher posterior variance, or a larger confidence interval, indicates a larger overconfidence bias in the posterior.

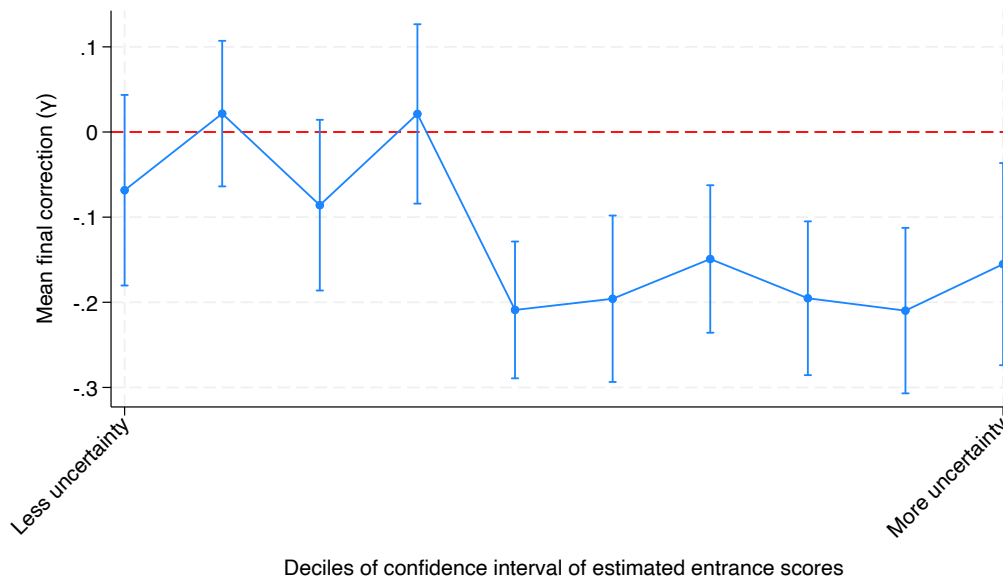
Suppose the individual understands that bias transmission occurs through the Bayesian weight of the prior, α . While they cannot directly observe α , they can infer it from the posterior variance. Then, for a given \hat{P}_j , $E[P_j | \hat{P}_j, \sigma_{post,j}^2]$ is decreasing in $\sigma_{post,j}^2$. Thus, Equation 11 indicates that the magnitude of the final correction is increasing in variance of the posterior. This generates a direct relationship between confidence interval width and correction magnitude: the wider the confidence interval, the larger the negative correction. This relationship arises because wider confidence intervals signal higher posterior uncertainty and thus greater reliance on the biased prior. Conscious adjustment responds by making larger corrections when observing wider intervals.

Confidence intervals and the correction mechanism: empirical results. To examine how perceived uncertainty relates to belief adjustments, we make use of students' subjective confidence intervals for their score estimates. Recognizing that students may not be familiar with formal statistical concepts like 95% confidence intervals, we asked them to report their "highest plausible score" and "lowest plausible score" for each subject. This framing captures their subjective uncertainty in intuitive terms. The responses suggest students understood the question as intended. While they could

have interpreted “highest plausible” as the maximum possible score (120 or 60 points), only 1% of students reported this value. Similarly, no student reported 0 as their lowest plausible score, indicating they provided meaningful bounds rather than theoretical extremes. We define the confidence interval width as: $W_{ij} = H_{ij} - L_{ij}$, where H_{ij} and L_{ij} are student i 's reported highest and lowest plausible scores for subject j .

To examine how subjective uncertainty relates to final corrections, we regress γ_j on deciles of confidence interval width, controlling for subject fixed effects, and report the results in Figure 6.

Figure 6: Perceived uncertainty and the final correction term



Notes: This graph reports the relationship of the final correction term γ to deciles of the width of the reported confidence interval of the estimated score. A positive value of γ means adjusting the estimation up, and a negative value means adjusting the estimation down. The width of the confidence interval is defined as: $W_{ij} = H_{ij} - L_{ij}$, where H_{ij} and L_{ij} are student i 's reported highest and lowest plausible scores for subject j . Error bars indicate 95% confidence intervals.

We find students who report greater uncertainty make substantially larger downward corrections. Students in the lowest uncertainty deciles (1-4) make corrections close to zero, with some even making small positive adjustments. However, starting from the fifth decile, corrections become increasingly negative, reaching approximately -0.2 standard deviation of the actual exam score for students in the highest uncertainty deciles.

This pattern supports our theoretical prediction that corrections respond to the perceived reliability of initial estimates. When uncertainty is high (wide intervals), students recognize that they likely placed substantial weight on a potentially biased prior, leading to larger downward corrections.

An important limitation of our interval analysis is that 7.2% of subject-level observations lack

confidence interval data. These missing responses are not random: students who did not report intervals show greater overconfidence in estimation (0.089 vs. -0.047 standard deviation, $p = 0.013$) and somewhat higher overconfidence in recall (0.187 vs. 0.106 standard deviation, $p = 0.249$). Several interpretations are possible. Students who did not provide intervals may have found it difficult to quantify their uncertainty, potentially reflecting lower meta-cognitive awareness. Alternatively, overconfident students might be less inclined to acknowledge uncertainty by providing ranges. The missing data could also reflect survey fatigue or misunderstanding of the question. While our main results about recall bias, Bayesian updating, and average corrections remain robust in both the full sample and the subsample with non-missing interval data, readers should interpret the uncertainty-correction relationship with this limitation in mind.

8 Conclusion

This study investigates the malleability of motivated beliefs by examining how students form beliefs in a high-stakes environment, where estimation errors directly affect educational opportunities. While prior literature documents widespread overconfidence across domains, we ask whether such biases persist when mistakes are costly and whether the psychological mechanisms generating overconfidence adapt to environmental pressures. Our setting allows us to observe belief formation at multiple stages for the same individual: when students recall past performance and when they estimate scores that become critical in the application process.

Our empirical results reveal a striking pattern: students exhibit significant overconfidence when recalling mock exam scores, yet state accurate beliefs when estimating high-stakes entrance exam scores. This contrast emerges despite the psychological mechanisms generating bias remaining fully active: students rely on their inflated recall when forming estimates, with bias transmitting through Bayesian updating at predicted rates. The puzzle of accurate beliefs despite biased inputs points to conscious adjustment processes. We find suggestive evidence for systematic downward corrections that offset initial overconfidence, with larger adjustments when uncertainty is higher, suggesting partial meta-awareness of overconfidence tendencies. These patterns demonstrate that motivated beliefs are malleable: while the supply-side forces generating bias persist even under high stakes, deliberative corrections can emerge to produce accurate or even under-confident beliefs when accuracy matters. The same individuals who cannot help but remember their past performance through rose-colored

glasses can nonetheless form realistic or pessimistic beliefs when consequential decisions demand it.

Beyond the main results, we uncover rich patterns in how motivated memory operate across domains and demographics. Recall bias increases systematically with poor performance: students with worse actual scores show greater overconfidence both between and within individuals across subjects, suggesting memory distortion serves to protect self-image where it is most threatened. Interestingly, students who improved from mock to entrance exam are more likely to not recall at all, pointing to motivated forgetting as an alternative ego-protective strategy when current performance surpasses disappointing past results. The gender gap in overconfidence in recall varies systematically with ego relevance: males exhibit substantially more overconfidence than females in STEM subjects that are considered to be male-typed, but similar levels in non-STEM domains. Subject characteristics also matter: recall bias is significantly larger in subjective non-STEM domains compared to the more objective STEM domains, consistent with higher cognitive costs of distorting memories about objective performance. These heterogeneous patterns reveal the sophisticated nature of motivated memory, with different individuals employing different strategies (biased recall versus selective forgetting) depending on their specific circumstances and the psychological costs involved.

While our setting is specialized, high-stakes testing in China, such specificity is arguably necessary to answer the set of questions we ask about belief malleability. Testing whether motivated beliefs respond to environmental pressures requires specific conditions: between or within-person variation across contexts, objective performance benchmarks, and high enough consequences for accuracy. Our setting naturally provides these elements, with students both recalling past performance and estimating future scores where mistakes directly affect important life outcomes. The mechanisms we identify should generalize broadly: whenever individuals face tradeoffs between self-image and decision quality, conscious adjustments may emerge to counteract psychological biases. While we find accurate beliefs on average in our high-stakes context, we do not claim that high-stakes always lead to accurate beliefs. The general principle is that heightened stakes should reduce overconfidence relative to low-stakes settings — whether this results in underconfidence, accuracy, or merely reduced overconfidence may depend on the specific environment and strength of incentives.

Several directions merit future investigation. First, examining how belief adjustment operates under varying stake levels could help identify whether conscious corrections emerge gradually or only above certain thresholds. A particularly promising approach would be laboratory experiments

where participants first provide beliefs under low incentives, then receive opportunities to revise their answers when stakes increase substantially. Testing whether revisions are systematically downward would provide cleaner evidence for conscious adjustment, though such experiments are beyond the scope of this paper. Second, it would be valuable to investigate which strategies individuals adopt when accuracy becomes important: do they attempt to correct their biased memories directly, switch to using objective information when available, or maintain biased inputs while adjusting final beliefs? The third strategy dominates in our setting, but understanding what determines this choice of strategy remains an open question. Third, further research should examine how individuals develop awareness of their own biases. In repeated settings, either laboratory or field, tracking how meta-awareness emerges through feedback could reveal whether people gradually learn the magnitude and sources of their biases, refine their correction strategies, or shift to different approaches altogether as experience accumulates.

REFERENCES

- Ariely, Dan, Uri Gneezy, George Loewenstein, and Nina Mazar**, “Large Stakes and Big Mistakes,” *Review of Economic Studies*, 2009, 76 (2), 451–469.
- Augenblick, Ned, Eben Lazarus, and Michael Thaler**, “Overinference from weak signals and underinference from strong signals,” *The Quarterly Journal of Economics*, 2025, 140 (1), 335–401.
- Ba, Cuimin, J Aislinn Bohren, and Alex Imas**, “Over- and underreaction to information,” *Available at SSRN 4274617*, 2024.
- Belot, Michèle, Bhaskar Bhaskar, and Jeroen van de Ven**, “Promises and Cooperation: Evidence from a TV Game Show,” *The Journal of Economic Behavior and Organization*, 2010, 73 (3), 396–405.
- Bénabou, Roland and Jean Tirole**, “Self-confidence and personal motivation,” *The quarterly journal of economics*, 2002, 117 (3), 871–915.
- and —, “Mindful economics: The production, consumption, and value of beliefs,” *Journal of Economic Perspectives*, 2016, 30 (3), 141–64.
- Benjamin, Daniel J**, “Errors in probabilistic reasoning and judgment biases,” *Handbook of Behavioral Economics: Applications and Foundations 1*, 2019, 2, 69–186.
- Berk, Jonathan B., Eric Hughson, and Kirk Vandezande**, “The Price is Right, But Are the Bids? An Investigation of Rational Decision Theory,” *The American Economic Review*, 1996, 86 (4), 954–970.
- Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer**, “Beliefs about gender,” *American Economic Review*, 2019, 109 (3), 739–73.
- Bosch-Rosa, Ciril, Bernhard Kassner, and Steffen Ahrens**, “Overconfidence and the Political and Financial Behavior of a Representative Sample,” *Working Paper*, 2024.
- Bracha, Anat and Donald J Brown**, “Affective decision making: A theory of optimism bias,” *Games and Economic Behavior*, 2012, 75 (1), 67–80.

- Brunnermeier, Markus K and Jonathan A Parker**, “Optimal expectations,” *American Economic Review*, 2005, 95 (4), 1092–1118.
- Camerer, Colin F**, “Do biases in probability judgment matter in markets? Experimental evidence,” *The American Economic Review*, 1987, 77 (5), 981–997.
- Camerer, Colin F. and Dan Lovo**, “Overconfidence and Excess Entry: An Experimental Approach,” *American Economic Review*, 1999, 89 (1), 306–318.
- Camerer, Colin F and Robin M Hogarth**, “The effects of financial incentives in experiments: A review and capital-labor-production framework,” *Journal of risk and uncertainty*, 1999, 19, 7–42.
- Caplin, Andrew and John V Leahy**, “Wishful thinking,” Technical Report, National Bureau of Economic Research 2019.
- Chambers, Christopher P and Paul J Healy**, “Updating toward the signal,” *Economic Theory*, 2012, 50 (3), 765–786.
- Chen, Daniel L., Tobias J. Moskowitz, and Kelly Shue**, “Decision-Making under the Gambler’s Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires,” *The Quarterly Journal of Economics*, 2016, 131 (3), 1181–1242.
- Coffman, Katherine, Manuela R Collis, and Leena Kulkarni**, “Stereotypes and belief updating,” *Journal of the European Economic Association*, 2024, 22 (3), 1011–1054.
- Compte, Olivier and Andrew Postlewaite**, “Confidence-enhanced performance,” *American Economic Review*, 2004, 94 (5), 1536–1557.
- Enke, Benjamin, Uri Gneezy, Brian Hall, David Martin, Vadim Nelidov, Theo Offerman, and Jeroen Van De Ven**, “Cognitive biases: Mistakes or missing stakes?,” *Review of Economics and Statistics*, 2023, 105 (4), 818–832.
- Exley, Christine L and Judd B Kessler**, “Motivated errors,” *American Economic Review*, 2024, 114 (4), 961–987.
- Gneezy, Uri, Yoram Halevy, Brian Hall, Theo Offerman, and Jeroen van de Ven**, “How Real is Hypothetical? A High-Stakes Test of the Allais Paradox,” Technical Report 2024.

- Gödker, Katrin, Peiran Jiao, and Paul Smeets**, “Investor memory,” *The Review of Financial Studies*, 2025, 38 (6), 1595–1640.
- Gossner, Olivier and Jakub Steiner**, “On the cost of misperception: General results and behavioral applications,” *Journal of Economic Theory*, 2018, 177, 816–847.
- Gottlieb, Daniel**, “Will you never learn? self deception and biases in information processing,” *Working Paper*, 2010.
- Graddy, Kathryn, Noah Horowitz, and Stefan Szymanski**, “A Study of Auction Prices in Impressionist and Contemporary Art Markets,” *The Review of Economics and Statistics*, 2014, 96 (4), 784–795.
- Graeber, Thomas, Christopher Roth, and Florian Zimmermann**, “Stories, statistics, and memory,” *The Quarterly Journal of Economics*, 2024, 139 (4), 2181–2225.
- Hagenbach, Jeanne and Charlotte Saucet**, “Motivated skepticism,” *Review of Economic Studies*, 2025, 92 (3), 1882–1919.
- **and Frédéric Koessler**, “Selective memory of a psychological agent,” *European Economic Review*, 2022, 142, 104012.
- **, Nicolas Jacquemet, and Philipp Sternal**, “The motivated memory of noise,” *Games and Economic Behavior*, 2025.
- Heck, Patrick R, Daniel J Benjamin, Daniel J Simons, and Christopher F Chabris**, “Overconfidence persists despite years of accurate, precise, public, and continuous feedback: Two studies of tournament chess players,” 2024.
- Huang, Wei, Soo Hong Chew, and Xiaojian Zhao**, “Motivated False Memory,” *Journal of Political Economy*, 2020.
- Huffman, David, Collin Raymond, and Julia Shvets**, “Persistent overconfidence and biased memory: Evidence from managers,” *American Economic Review*, 2022, 112 (10), 3141–75.
- Jetter, Michael and Jay K. Walker**, “Game, Set, and Match: Do Women and Men Perform Differently in Competitive Situations?,” *Journal of Economic Behavior & Organization*, 2017, 135, 362–372.

- Kőszegi, Botond**, “Ego utility, overconfidence, and task choice,” *Journal of the European Economic Association*, 2006, 4 (4), 673–707.
- Kruger, Justin and David Dunning**, “Unskilled and unaware of it: how difficulties in recognizing one’s own incompetence lead to inflated self-assessments.,” *Journal of personality and social psychology*, 1999, 77 (6), 1121.
- Levitt, Steven D.**, “Testing Theories of Discrimination: Evidence from ”The Weakest Link”,” *The Journal of Law and Economics*, 2004, 47 (2), 431–452.
- Malmendier, Ulrike and Geoffrey Tate**, “CEO overconfidence and corporate investment,” *The journal of finance*, 2005, 60 (6), 2661–2700.
- Metrick, Andrew**, “A Natural Experiment in ”Jeopardy!”,” *The American Economic Review*, 1995, 85 (1), 240–253.
- Moore, Don A and Paul J Healy**, “The trouble with overconfidence.,” *Psychological review*, 2008, 115 (2), 502.
- Mueller, Andreas I, Johannes Spinnewijn, and Giorgio Topa**, “Job seekers’ perceptions and employment prospects: Heterogeneity, duration dependence, and bias,” *American Economic Review*, 2021, 111 (1), 324–363.
- Orhun, Yeşim, Alain Cohn, and Collin B Raymond**, “Motivated optimism and workplace risk,” *The Economic Journal*, 2024, 134 (663), 2951–2981.
- Ortoleva, Pietro and Erik Snowberg**, “Overconfidence in political behavior,” *American Economic Review*, 2015, 105 (2), 504–535.
- Pires, Pedro**, “How Much Can You Make? Misprediction and Biased Memory in Gig Jobs,” *Working Paper*, 2025.
- Pope, Devin G. and Maurice E. Schweitzer**, “Is Tiger Woods Loss Averse? Persistent Bias in the Face of Experience, Competition, and High Stakes,” *The American Economic Review*, 2011, 101 (1), 129–157.
- Roy-Chowdhury, Vivek**, “Biased Recall and The Dynamics of Beliefs: Evidence from schools,” 2024.

Sial, Afras Y, Justin R Sydnor, and Dmitry Taubinsky, “Biased Memory and Perceptions of Self-Control,” Technical Report, National Bureau of Economic Research 2023.

Teeselink, Bouke Klein, Dennie van Dolder, Martijn J van den Assem, and Jason Dana, “High-stakes failures of backward induction,” *Available at SSRN 4130176*, 2024.

Zimmermann, Florian, “The dynamics of motivated beliefs,” *American Economic Review*, 2020, *110* (2), 337–363.

THE MALLEABILITY OF MOTIVATED BELIEFS

ONLINE APPENDIX

A.1 Appendix Tables

Table A.1: Sample comparison: Full sample vs. Recall sample

	Full Sample			Recall Sample			Diff
	n	mean	sd	n	mean	sd	
<i>Panel A: Exam Performance</i>							
Entrance exam score	1812	80.50	21.24	1478	79.92	21.66	-0.58**
Estimated entrance score	1812	80.00	22.12	1478	79.07	22.44	-0.93***
Mock exam score	1804	73.87	23.56	1478	73.55	24.02	-0.31
<i>Panel B: Student-Level Comparison</i>							
	n	mean	sd	n	mean	sd	
Total entrance score	302	545.48	78.52	251	542.19	82.98	-3.29
Total estimated score	302	542.57	84.21	251	537.60	87.73	-4.97**
Total mock score	302	441.24	98.93	251	439.90	102.65	-1.35
Male	302	0.53	0.50	251	0.53	0.50	0.01
Risk tolerance	296	5.85	2.67	249	5.87	2.65	0.02
Father education	290	3.77	0.98	247	3.75	0.99	-0.02
Mother education	289	3.81	1.05	246	3.81	1.01	0.00

Notes: This table compares students who provided recall data with the full sample. Panel A shows statistics at the observation level (student-subject), while Panel B shows student-level comparisons. The Diff column shows the difference in means. Risk tolerance is measured on a 0-10 scale, with higher values indicating greater willingness to take risks. Parental education is coded on a 1-6 scale: 1 = primary school or below, 2 = middle school, 3 = high school, 4 = vocational/technical college, 5 = bachelor's degree, 6 = graduate degree. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$, based on t-tests allowing for clustering at the student level (Panel A) or standard t-tests (Panel B).

Table A.2: Overconfidence in recall and performance by subject

Dependent variable	Overconfidence in recall					
	Math (1)	Chinese (2)	English (3)	Phy&Che (4)	His&Pol (5)	Geo&Bio (6)
Actual mock exam score	-0.252*** (0.032)	-0.233*** (0.039)	-0.056** (0.023)	-0.198*** (0.031)	-0.234*** (0.050)	-0.292*** (0.034)
Observations	249	244	248	245	246	246
R-squared	0.196	0.130	0.023	0.140	0.083	0.228

Notes: This table reports the relationship between overconfidence in recall and performance in the mock exam separately for each subject. The dependent variable “Overconfidence in recall” is defined as the difference between recalled and actual mock exam scores. The independent variable “Actual mock exam score” is the actual mock exam score in each subject. The observation unit is student. Columns (1) through (6) present results separately for Mathematics, Chinese, English, Physics & Chemistry, History & Politics, and Geography & Biology, respectively. Standard errors are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.3: Gender differences in recall bias by subject domain

Dependent variable	Recalled mock exam score		
	(1) Male Students	(2) Female Students	(3) Full Sample with Interaction
Actual mock exam score	0.620*** (0.055)	0.482*** (0.068)	0.558*** (0.043)
STEM subject	-0.019 (0.041)	-0.210*** (0.041)	-0.191*** (0.070)
Male \times STEM			0.187*** (0.057)
Observations	794	684	1478
R-squared	0.748	0.800	0.779

Notes: This table reports gender differences in recall bias across subject domains. The dependent variable is recalled mock exam score. Column (1) restricts the sample to male students, column (2) to female students, and column (3) includes the full sample with a male-STEM interaction term. All specifications include individual fixed effects. Standard errors, reported in parentheses, are clustered at the student level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.4: Predictors of not recalling mock exam scores

Dependent variable	Do not recall				
	Individual-Level			Within-Individual	
	(1)	(2)	(3)	(4)	(5)
Total estimated entrance score	0.002*** (0.001)	0.002*** (0.001)			
Total actual entrance score			0.003*** (0.001)		
Total mock score	-0.002*** (0.000)	-0.002*** (0.000)	-0.002*** (0.001)		
Male		-0.028 (0.041)	-0.007 (0.043)		
Risk tolerance		-0.004 (0.005)	-0.003 (0.004)		
Father education		0.026 (0.045)	0.034 (0.042)		
Mother education		-0.029 (0.059)	-0.020 (0.057)		
Estimated entrance score (subject)				0.572** (0.244)	
Actual entrance score (subject)					0.165** (0.082)
Mock score (subject)				-0.642** (0.281)	-0.145 (0.130)
Subject FE	No	No	No	Yes	Yes
Individual FE	No	No	No	Yes	Yes
Observations	302	284	284	40	40

Notes: This table reports logit regression marginal effects for the probability of not providing recall data. Columns (1)-(3) show individual-level analysis, while columns (4)-(5) show within-individual analysis with fixed effects. “Estimated” columns use students’ estimated scores (what they knew at the time), while “Actual” columns use realized scores to address potential endogeneity. Standard errors in parentheses are clustered at the class level for individual-level regressions and at the student level for within-individual regressions. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.5: Overconfidence in estimation and performance by subject

Dependent variable	Overconfidence in estimation					
	Math (1)	Chinese (2)	English (3)	Phy&Che (4)	His&Pol (5)	Geo&Bio (6)
Actual entrance exam score	-0.017 (0.031)	-0.251*** (0.041)	-0.119*** (0.020)	0.014 (0.024)	-0.029 (0.047)	-0.213*** (0.041)
Observations	302	302	302	302	302	302
R-squared	0.001	0.113	0.103	0.001	0.001	0.084

Notes: This table reports the relationship between overconfidence in estimation and performance in the entrance exam separately for each subject. The dependent variable “Overconfidence in estimation” is defined as the difference between estimated and actual entrance exam scores. The independent variable “Actual entrance exam score” is the actual entrance exam score in each subject. The observation unit is student. Columns (1) through (6) present results separately for Mathematics, Chinese, English, Physics & Chemistry, History & Politics, and Geography & Biology, respectively. Standard errors are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

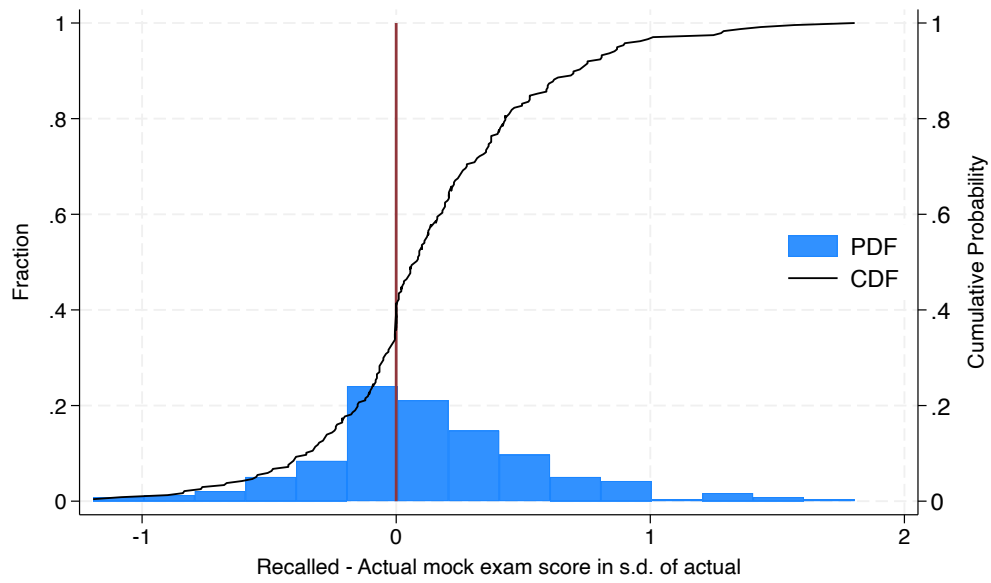
Table A.6: Gender differences in estimation bias by subject domain

Dependent variable	Estimated entrance exam score		
	(1) Male Students	(2) Female Students	(3) Full Sample with Interaction
Actual entrance exam score	0.533*** (0.055)	0.641*** (0.055)	0.575*** (0.038)
STEM subjects	0.081** (0.036)	-0.147*** (0.037)	-0.158*** (0.036)
Male \times STEM			0.233*** (0.051)
Observations	954	858	1812
R-squared	0.762	0.806	0.840

Notes: This table reports gender differences in estimation bias across subject domains. The dependent variable is estimated entrance exam score. Column (1) restricts the sample to male students, column (2) to female students, and column (3) includes the full sample with a male-STEM interaction term. All specifications include individual fixed effects. Standard errors, reported in parentheses, are clustered at the student level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

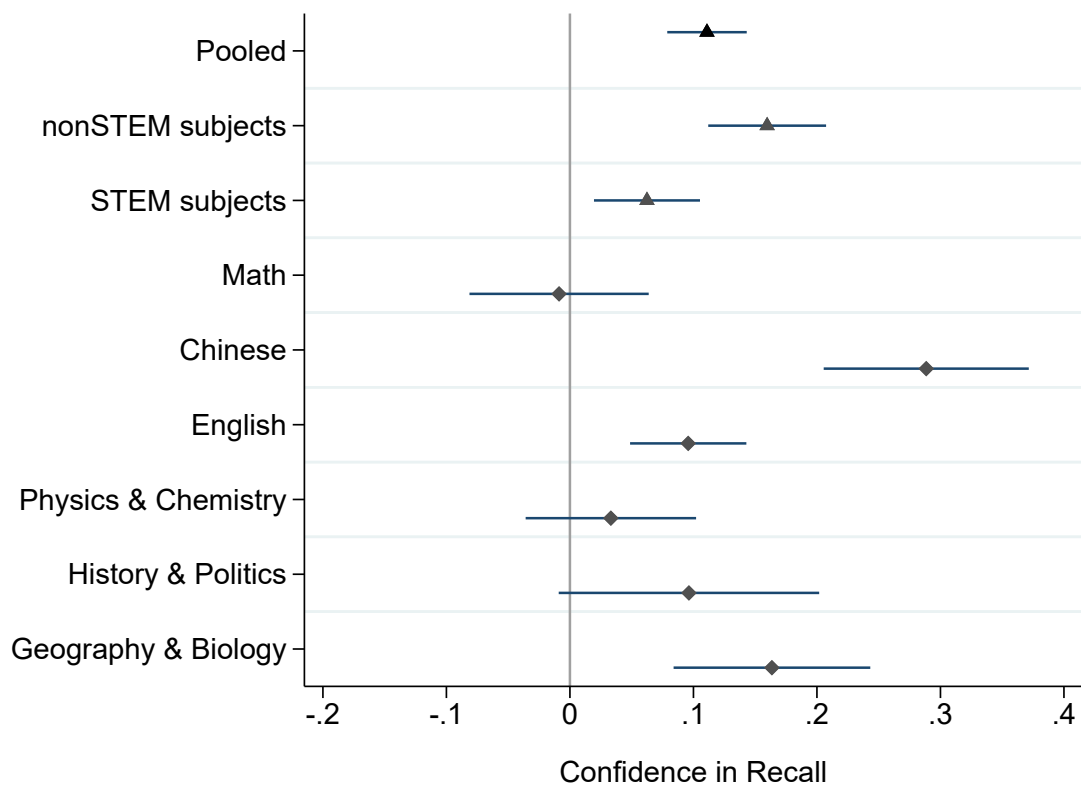
A.2 Appendix Figures

Figure A.1: Confidence in recall: Distribution



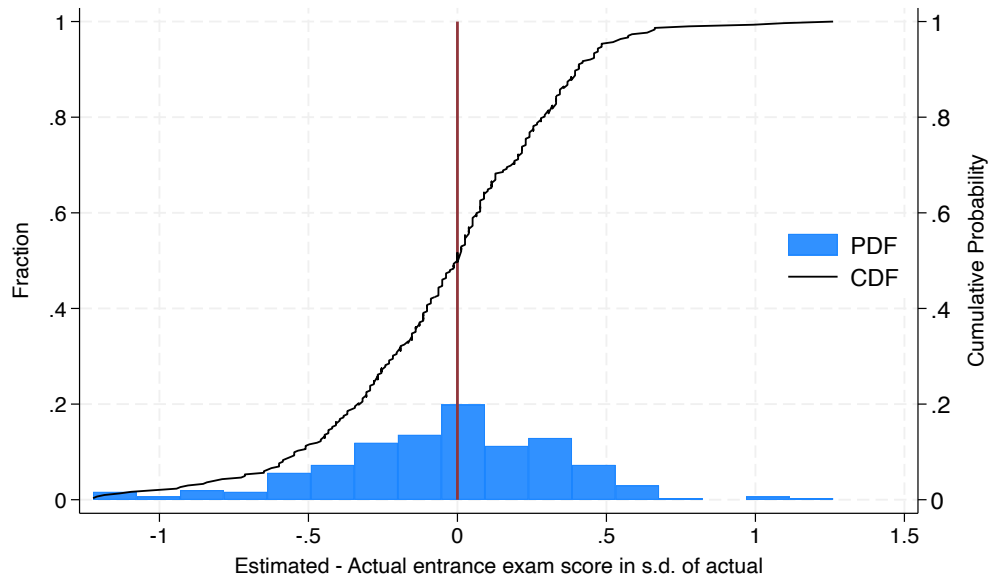
Notes: This graph reports the distribution of confidence in recall at the individual level. A positive number represents overconfidence in recall, a negative number represents underconfidence in recall, and 0 stands for accurate recall.

Figure A.2: Confidence in recall: Subjects



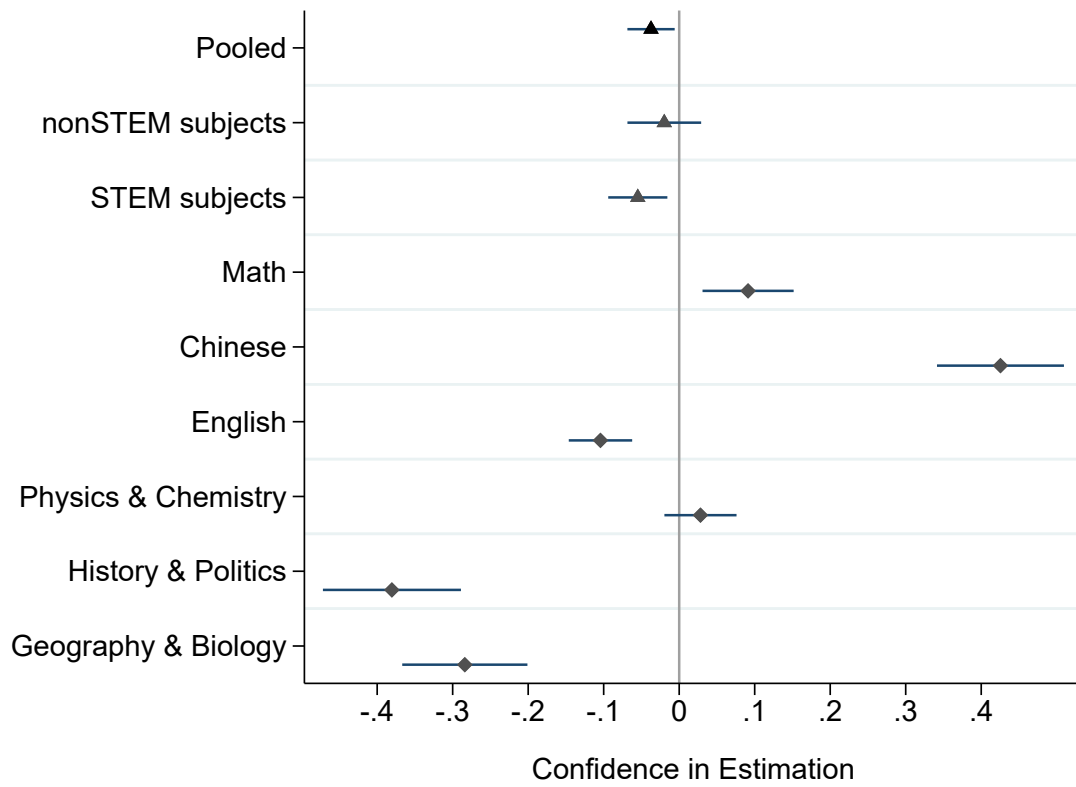
Notes: This graph reports the confidence in recall at the subject level. Here “Pooled” stands for pooling data of all six subjects; “non-STEM subjects” refers to pooling data of three non-STEM subjects, namely Chinese, English, History & Politics; “STEM subjects” refers to pooling data of three STEM subjects, namely Mathematics, Physics & Chemistry, and Geography & Biology.

Figure A.3: Confidence in estimation: Distribution



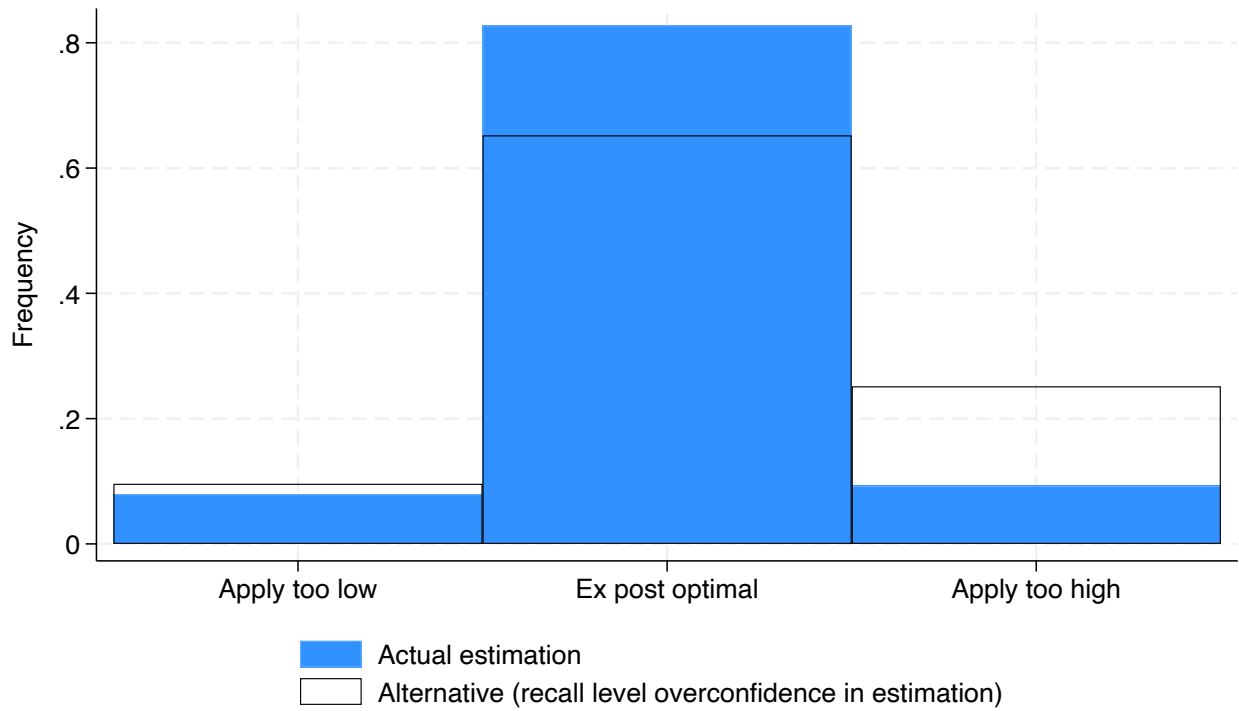
Notes: This graph reports the distribution of confidence in estimation at the individual level. A positive number represents overconfidence in estimation, a negative number represents underconfidence in estimation, and 0 stands for accurate estimation.

Figure A.4: Confidence in estimation: Subjects



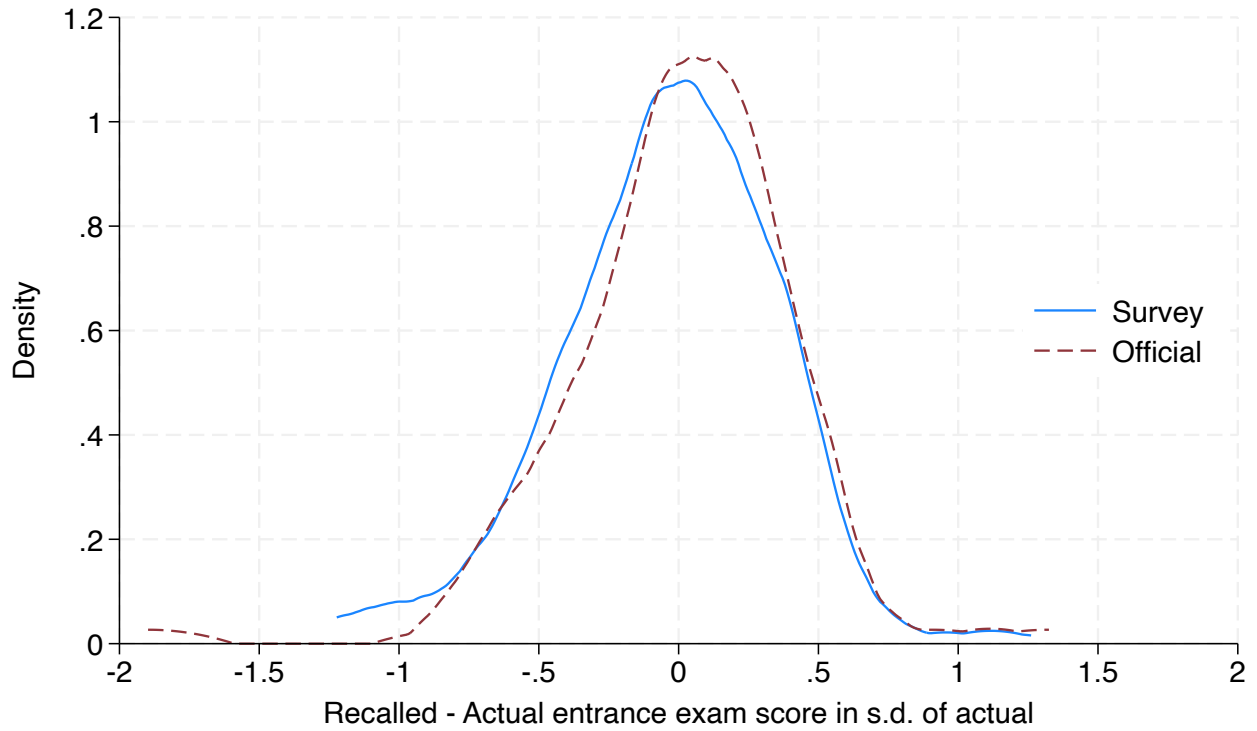
Notes: This graph reports the confidence in estimation at the subject level. Here “Pooled” stands for pooling data of all six subjects; “nonSTEM subjects” refers to pooling data of three non-STEM subjects, namely Chinese, English, and History & Politics; “STEM subjects” refers to pooling data of three STEM subjects, namely Mathematics, Physics & Chemistry, and Geography & Biology.

Figure A.5: Distribution of mistakes in school choice



Notes: This figure reports the distribution of potential mistakes in school choice. We calculate mistakes by comparing the highest-ranked school a student could attend based on their actual score (ex-post optimal) with the highest-ranked school they could attend based on their estimated score, assuming students apply sincerely according to their estimates. “Apply too low” indicates that a student’s estimated score would lead them to apply to a less selective school than their ex-post optimal choice, while “Apply too high” indicates they would apply to a more selective school. The blue bars show the distribution based on students’ actual estimations. The white bars show a counterfactual distribution where we assume students exhibit the same degree of overconfidence in estimation as observed in their recall of mock exam performance.

Figure A.6: Comparison of estimation errors: Anonymous survey vs. Official reporting



Notes: This figure compares the distribution of estimation errors between students' anonymous survey responses and their official reports to teachers. The blue solid line shows the kernel density of overconfidence in our survey. The maroon dashed line shows the kernel density of overconfidence in official estimations reported to their school. Overconfidence is defined as the difference between estimated and actual entrance exam scores, standardized by the standard deviation of actual scores.

B Conceptual framework: Further details

Proof of Comparative Statics: To derive the relationship between overconfidence in recall b_j^* and the ego-relevance, the cognitive cost, and the actual past performance, we firstly define the optimal bias b_j^* implicitly by $F(\cdot)$:

$$F(b_j^*, w_j, \kappa, P_{past,j}) = w_j v'(P_{past,j} + b_j^*) - 2\kappa b_j^* = 0$$

By the implicit function theorem, $\frac{\partial b_j^*}{\partial w_j}$, for example, equals to

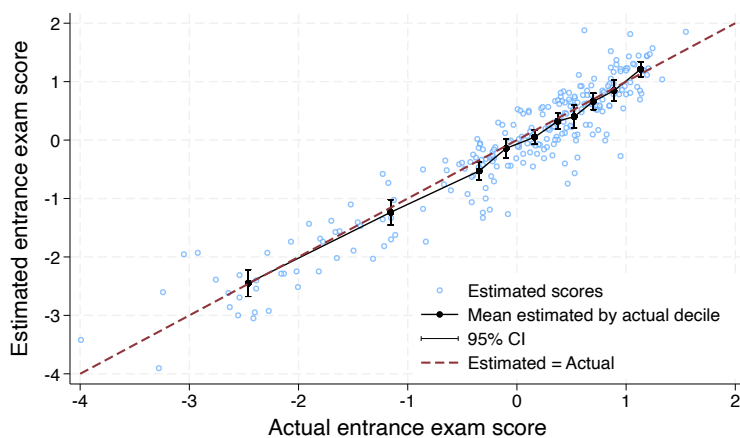
$$\frac{\partial b_j^*}{\partial w_j} = -\frac{\frac{\partial F}{\partial w_j}}{\frac{\partial F}{\partial b_j^*}},$$

here $\frac{\partial F}{\partial b_j^*} = w_j v''(P_{past,j} + b_j^*) - 2\kappa < 0$ because $v''(\cdot) < 0$ (ego-utility is assumed to be concave), and combined with $w_j > 0$ and $\kappa > 0$. And $\frac{\partial F}{\partial w_j} = v'(P_{past,j} + b_j^*) > 0$, because ego-utility is increasing in the prior belief. Thus $\frac{\partial b_j^*}{\partial w_j} > 0$, indicating that stronger ego-relevance leads to more overconfidence in recall.

As for the effect of Cognitive Costs (κ) and the effect of Past Performance ($P_{past,j}$), we can similarly derive that $\frac{\partial b_j^*}{\partial \kappa} = -\frac{-2b_j^*}{w_j v''(P_{past,j} + b_j^*) - 2\kappa} < 0$, and $\frac{\partial b_j^*}{\partial P_{past,j}} = -\frac{w_j v''(P_{past,j} + b_j^*)}{w_j v''(P_{past,j} + b_j^*) - 2\kappa} < 0$. Thus, overconfidence in recall is increasing in ego relevance, and decreasing in cognitive cost and actual past performance.

C Overconfidence in estimation: Recall sample

Figure C.1: Estimated vs. actual entrance exam scores: Recall sample



Notes: This graph reports the actual and estimated total entrance exam scores in the Recall sample. The 45-degree line represents accurate estimation of entrance exam scores; points above the 45-degree line represent overconfidence in the estimation and points below it represent underconfidence. The average estimated scores by decile of the actual scores are also shown with 95% confidence interval. Both actual and estimated scores are standardized with the mean and s.d. of actual entrance exam scores. Thus 0 on the x-axis represents average performance in the entrance exam and positive on the x-axis represents performing better than the average.

Table C.1: Overconfidence in estimation and performance by subject: Recall sample

Dependent variable	Overconfidence in estimation					
	Math (1)	Chinese (2)	English (3)	Phy&Che (4)	His&Pol (5)	Geo&Bio (6)
Actual entrance exam score	-0.012 (0.038)	-0.210*** (0.060)	-0.115*** (0.032)	0.009 (0.027)	-0.038 (0.048)	-0.228*** (0.061)
Observations	249	244	248	245	246	246
R-squared	0.001	0.088	0.096	0.001	0.003	0.102

Notes: This table reports the relationship between overconfidence in estimation and performance in the entrance exam separately for each subject in the Recall sample. The dependent variable “Overconfidence in estimation” is defined as the difference between estimated and actual entrance exam scores. The independent variable “Actual entrance exam score” is the actual entrance exam score in each subject. The observation unit is student. Columns (1) through (6) present results separately for Mathematics, Chinese, English, Physics & Chemistry, History & Politics, and Geography & Biology, respectively. Standard errors are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table C.2: Performance and overconfidence in estimation: Recall sample

Dependent Variable	Overconfidence in estimation		
	(1) Full	(2) High performers	(3) Low performers
Actual entrance exam score	-0.216*** (0.031)	-0.434*** (0.053)	-0.230*** (0.038)
Observations	1478	733	745
R-squared	0.516	0.638	0.473

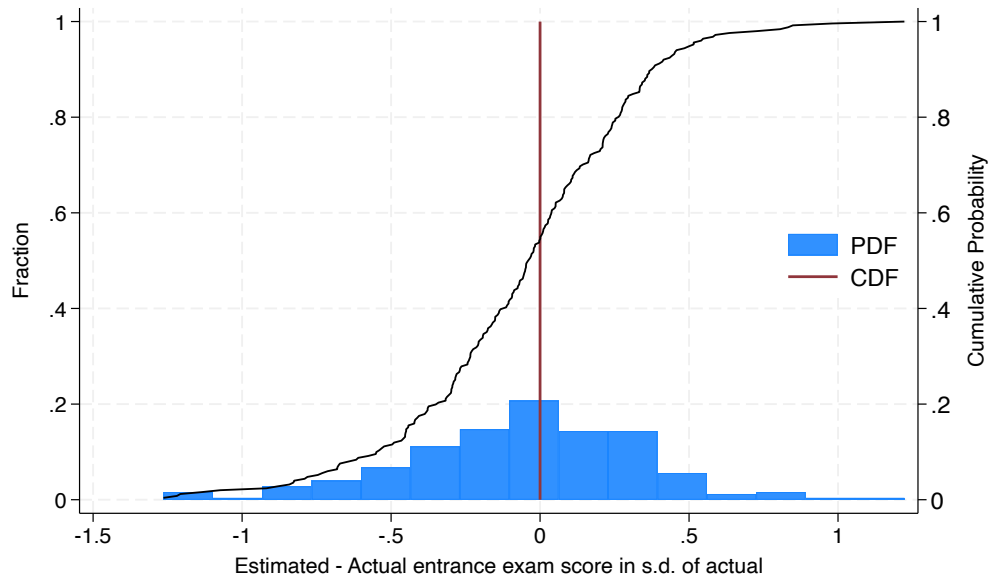
Notes: This table reports the relationship between overconfidence in estimation and performance in the entrance exam in the Recall sample. The dependent variable “Overconfidence in estimation” is defined as the difference between estimated and actual entrance exam scores. The independent variable “Actual entrance exam score” is the actual entrance exam score in a subject. The observation unit is student-subject. All specifications include individual and subject fixed effects to control for student-level and subject-specific factors that might affect estimation errors. Column (1) presents results for the full sample, while columns (2) and (3) present results for students who scored above and below the median in the entrance exam, respectively. Standard errors, reported in parentheses, are clustered at the student level to accommodate multiple subjects per student. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table C.3: Gender differences in estimation bias by subject domain: Recall sample

Dependent variable	Estimated entrance exam score		
	(1) Male Students	(2) Female Students	(3) Full Sample with Interaction
Actual entrance exam score	0.563*** (0.060)	0.663*** (0.062)	0.601*** (0.044)
STEM subjects	0.089** (0.037)	-0.129*** (0.041)	
Male \times STEM			0.223*** (0.055)
Observations	794	684	1478
R-squared	0.776	0.818	0.797

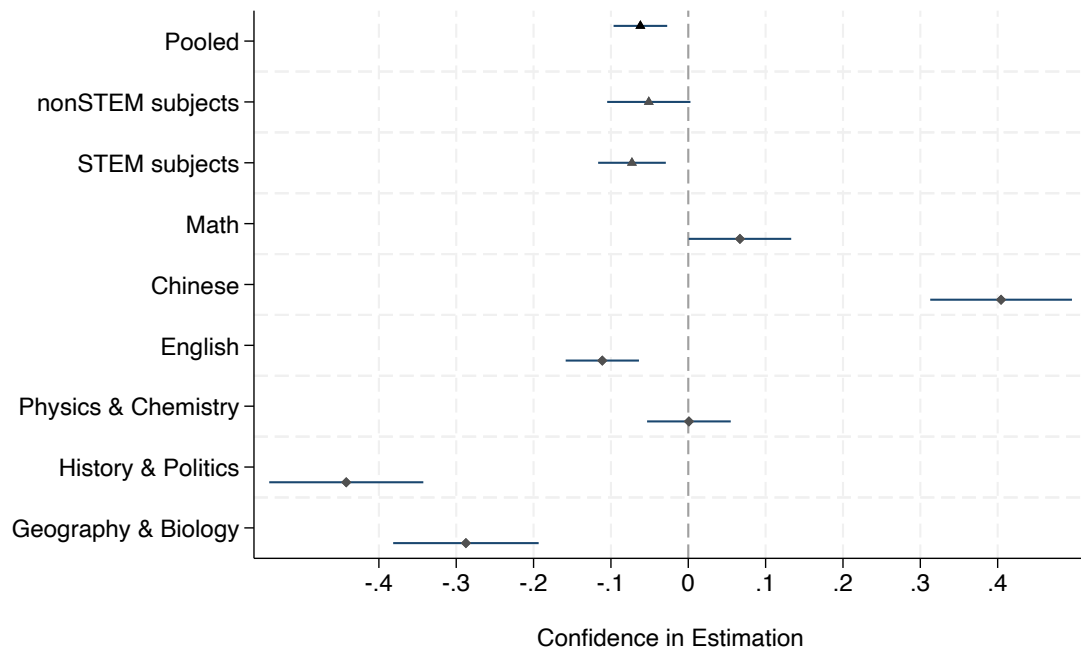
Notes: This table reports gender differences in estimation bias across subject domains in the Recall sample. The dependent variable is estimated entrance exam score. Column (1) restricts the sample to male students, column (2) to female students, and column (3) includes the full sample with a male-STEM interaction term. All specifications include individual fixed effects. Standard errors, reported in parentheses, are clustered at the student level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Figure C.2: Confidence in estimation: Distribution (Recall sample)



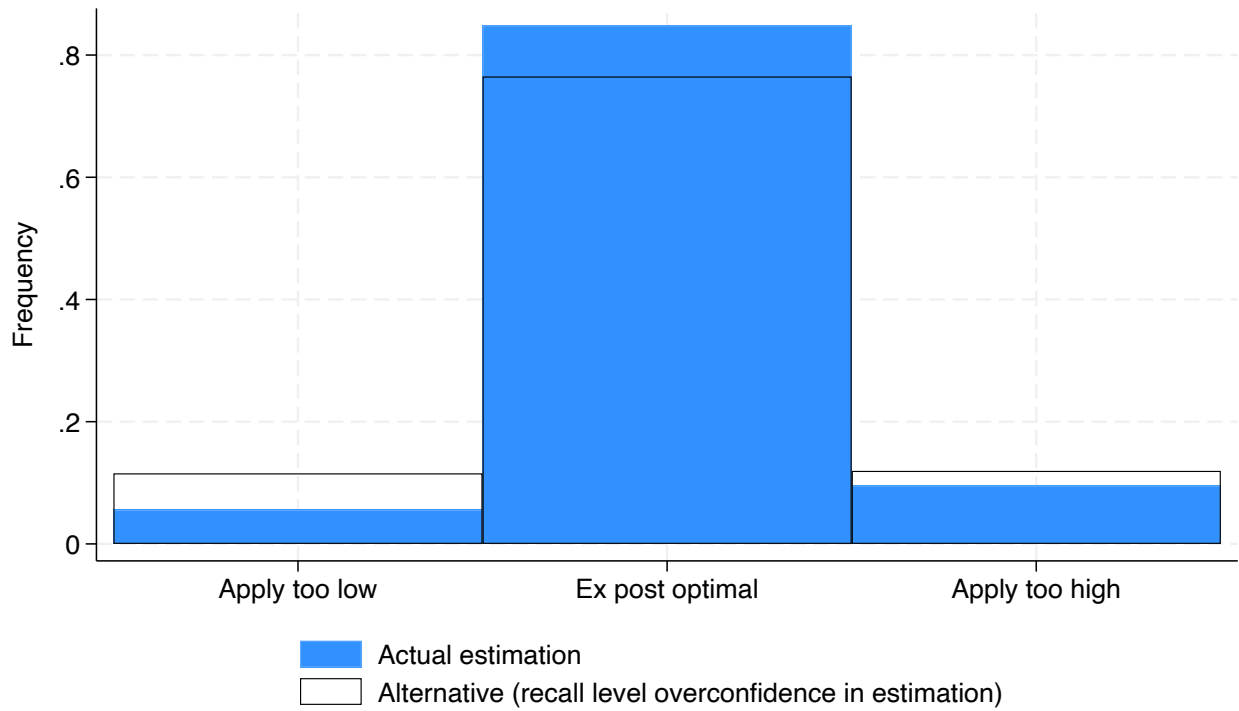
Notes: This graph reports the distribution of confidence in estimation at the individual level in the Recall sample. A positive number represents overconfidence in estimation, a negative number represents underconfidence in estimation, and 0 stands for accurate estimation.

Figure C.3: Confidence in estimation: Subjects (Recall sample)



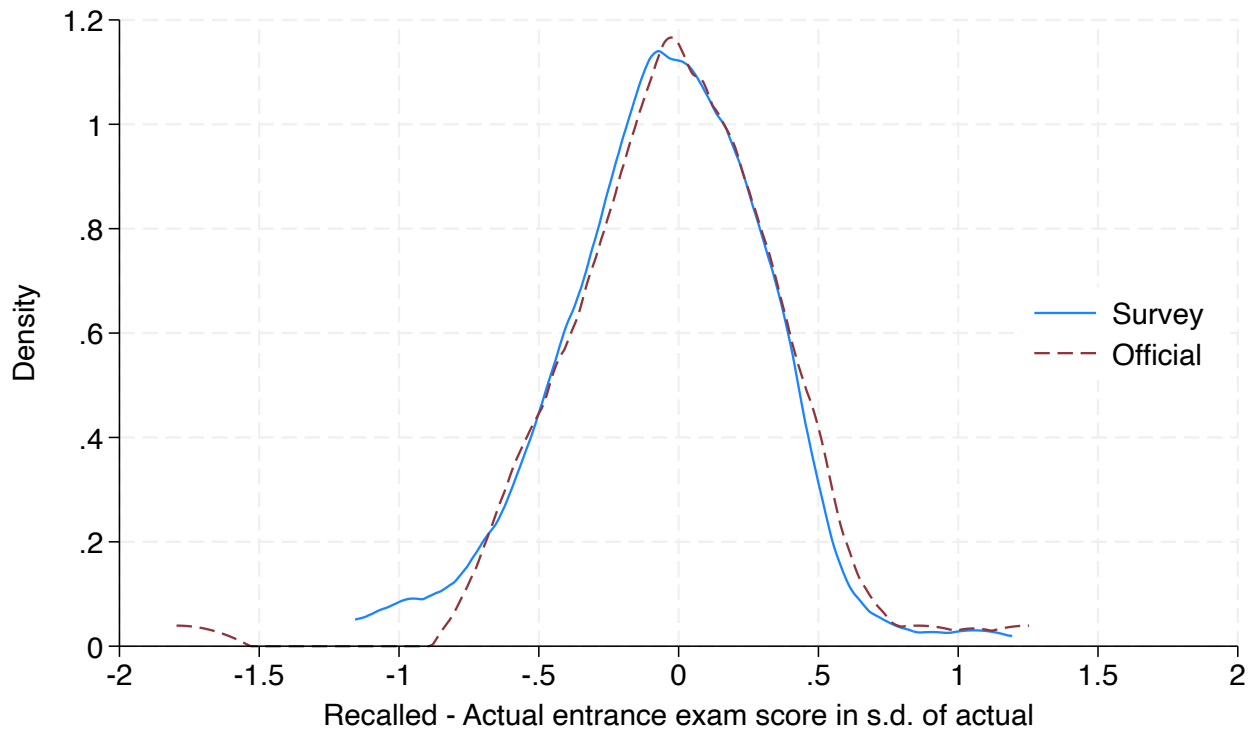
Notes: This graph reports the confidence in estimation at the subject level in the Recall sample. Here “Pooled” stands for pooling data of all six subjects; “nonSTEM subjects” refers to pooling data of three non-STEM subjects, namely Chinese, English, and History & Politics; “STEM subjects” refers to pooling data of three STEM subjects, namely Mathematics, Physics & Chemistry, and Geography & Biology.

Figure C.4: Distribution of mistakes in school choice: Recall sample



Notes: This figure reports the distribution of potential mistakes in school choice in the Recall sample. We calculate mistakes by comparing the highest-ranked school a student could attend based on their actual score (ex-post optimal) with the highest-ranked school they could attend based on their estimated score, assuming students apply sincerely according to their estimates. “Apply too low” indicates that a student’s estimated score would lead them to apply to a less selective school than their ex-post optimal choice, while “Apply too high” indicates they would apply to a more selective school. The blue bars show the distribution based on students’ actual estimations. The white bars show a counterfactual distribution where we assume students exhibit the same degree of overconfidence in estimation as observed in their recall of mock exam performance.

Figure C.5: Comparison of estimation errors: Anonymous survey vs. Official reporting (Recall sample)



Notes: This figure compares the distribution of estimation errors between students' anonymous survey responses and their official reports to teachers in the Recall sample. The blue solid line shows the kernel density of overconfidence in our survey. The maroon dashed line shows the kernel density of overconfidence in official estimations reported to their school. Overconfidence is defined as the difference between estimated and actual entrance exam scores, standardized by the standard deviation of actual scores.