Beliefs in a High-Stakes Environment

Yiming Liu* and Stephanie W. Wang†

Abstract

We investigate whether biased beliefs persist in a high-stakes environment. Students have strong incentives to accurately predict their entrance exam scores because they must submit their school choice before receiving their scores under an immediate acceptance mechanism. Consistent with models of motivated beliefs, students are not overconfident in high-stakes score estimation on average, but are overconfident in low-stakes recall of mock exam performance. The systematic patterns in how bias varies with performance, weight is placed on biased memory, and asymmetric updating suggest that even in high-stakes environments, belief formation remains subject to motivated forces.

1 Introduction

Biased beliefs about ability and performance, such as overconfidence or optimism bias, have been documented in many domains (Moore and Healy, 2008; Bandiera et al., 2022) and affect economic and political behavior (Camerer and Lovallo, 1999; Malmendier and Tate, 2005; Ortoleva and Snowberg, 2015; Bosch-Rosa et al., 2024). 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 couldn't change it without changing many other things."[1] On the other hand, models such as Bénabou and Tirole (2002), Compte and Postlewaite (2004), Brunnermeier and Parker (2005), and Köszegi (2006)

*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.

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

suggest that overconfidence is a motivated belief in that people enjoy the consumption utility or motivational effects of overconfident beliefs as long as the cost of the decision mistakes resulting from these biased beliefs do not outweigh the benefits. However, if these biased beliefs are ingrained rather than chosen, they can be harmful in situations such as overreaching when applying for universities or choosing a math-intensive track without sufficient proficiency. We ask whether the apparent overconfidence is an innate trait or optimally chosen by exploring whether individuals exhibit biased beliefs in both low-stakes but ego-relevant recall and high-stakes decision-making where mistakes can be very costly.

The school admission system in China provides an ideal environment to study this question. Specifically, we examine admissions to high school under the immediate acceptance algorithm in a city in Hebei province. After taking the high school entrance exam and receiving the answer key and scoring rubric, but before knowing their exact exam scores, students apply to one of several elite high schools. The schools' preferences for students are entirely defined by the high school entrance exam scores. All students not accepted in the first round are pooled together, and the system applies on their behalf for all the remaining slots using an identical ordering of the schools. Only around 30% of students can get into an elite high school, and the chance of attending a good university from a non-elite high school is slim if not zero. As a result, students need to accurately estimate their exam performance and make high-stakes decisions based on their estimation.

We conducted a survey to collect the students' estimations of their performance in all six subject exams before the students knew the actual scores and admissions outcomes. In total, we have 1812 score estimations from 302 students. We also asked them to recall their performance in a past mock exam that aimed to simulate the entrance exam in all aspects. These estimates are matched to administrative data of their actual entrance and mock exam scores. We also gathered information on student demographics using administrative data (gender and age) and the survey (e.g., family background).

In the low stakes environment of recalling mock exam performance, students exhibit overconfidence, with recalled scores being, on average, 0.1 standard deviation higher than actual scores. This overconfidence in the recall is observed across almost all subject exams (math, Chinese, English, natural sciences, and social sciences). We also observe more overconfidence when performance is worse: students who performed worse in the mock exam show greater overconfidence in their recall, and within individual, the tendency to inflate recalled scores is stronger for subjects where they

performed relatively poorly. These patterns in recall bias suggest that memory distortion may serve a motivated purpose of maintaining positive self-image rather than reflecting pure cognitive limitations.

However, in the high stakes environment of estimating their actual exam performance, students are remarkably accurate on average. The average estimated score is 0.04 standard deviation lower than the actual score. We continue to find no evidence of biased beliefs when we look at the estimates by subject. Moreover, strong overconfidence or underconfidence is rare, with more than 97% of students' estimated score deviate by less than one standard deviation from their actual score. While the stakes are high overall, the incentives to accurately estimate performance are low for students who either ranked at top or bottom. We indeed observe higher levels of overconfidence among these two groups, with the top performing students showing significant overconfidence.

We go on to examine the supply side of overconfidence, specifically biased memory as a tool for generating biased belief. After receiving answer keys and scoring rubrics, students obtain noisy signals about their performance in the entrance exam. They may also consult their recalled mock exam score as their prior. We find that students largely follow Bayes' rule in forming their beliefs, placing a larger weight on the signal obtaining from checking the answer keys and scoring rubrics than on the prior, the recalled mock exam score. Despite the fact that the more accurate actual mock exam score is readily available when estimating scores, students rely more on their potentially biased recalled score informing their estimation. Thus, bias in recall and bias in estimation are correlated. However, students also respond to the high stakes by limiting the transmission of the bias from recall to estimation: only around 10% of overconfidence in recall is transmitted to overconfidence in estimation, which explains why we observe strong average level overconfidence in recall but not estimation. Notably, this transmission of bias is nearly five times larger in subjects where students were overconfident in recall compared to subjects where they were accurate or underconfident, suggesting that students are motivated to inflate their recalled score in order to inflate their estimated score.

In addition to biased memory, we also find suggestive evidence for asymmetric updating as a supply side channel of overconfident beliefs. We define good news as performing noticeably better in the entrance exam than in the mock exam, bad news as performing noticeably worse, and neutral news as performing similar in the two exam. Students place more weight on actual entrance exam performance when receiving informative signals (either good or bad news) compared to neutral news, suggesting they can distinguish signal precision. However, they respond asymmetrically to these

signals: while they significantly increase their weight on entrance exam performance when receiving good news relative to neutral news, they show no such increase when receiving bad news. This asymmetric updating pattern in a field setting with non-binary states provides evidence consistent with studies documenting greater responsiveness to positive than negative feedback.

Lastly, we find substantial heterogeneity in overconfidence by gender. While we find no overconfidence on average in high-stakes estimation, male students show significantly greater overconfidence than female students, with this gap being particularly pronounced in STEM subjects. We further show that this gender difference in overconfidence is due to differences in priors. While male students are significantly more overconfident than female students in the low-stakes recall, especially among STEM subjects, they are similar in the updating process—they place similar weights on the prior versus the signal. Interestingly, only part of the gender gap in overconfidence in recall is transmitted into a gender gap in overconfidence in estimation, suggesting that the gender gap in overconfidence is motivated and can be altered by the cost of decision mistakes.

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; Huang et al., 2020; Zimmermann, 2020; Huffman et al., 2022). We provide some of the first field evidence on a key prediction of motivated beliefs: overconfidence decreases when the costs of holding biased beliefs increase. We show that the same students exhibit overconfidence in low-stakes recall tasks but no overconfidence in high-stakes estimations. This suggests that overconfidence is malleable, and can be influenced by the potential costs of decision mistakes.

Our study also adds to the literature on the supply-side mechanisms that produce overconfident beliefs. Our within-individual analysis indicates that students' overconfidence in low-stakes recall still predicts overconfidence in high-stakes estimation, replicating findings from Huffman et al. (2022), Sial et al. (2023), and Roy-Chowdhury (2024), who show that individuals exhibit overconfidence in recalling their past performance, and overconfidence in recall is correlated with overconfidence in prediction. We also document asymmetric updating: students place significantly more weight on the informative signal only when it is positive, echoing findings from Eil and Rao (2011), Wiswall and Zafar (2015), Charness and Dave (2017), and Möbius et al. (2022), who show that people update their beliefs asymmetrically depending on whether a signal is good or bad news to their self-view. We contribute to this literature by providing field evidence that the mechanisms producing overconfident

beliefs remain robust even when the cost of decision mistakes is high. Individuals do not completely "turn off" these underlying forces, even as their overall level of confidence becomes more accurate.

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 comparing the same individuals in both low- and high-stakes decisions (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; Klein Teeselink et al., 2024). Meanwhile, experimental work shows that higher incentives can mitigate but seldom eliminate cognitive biases (Camerer, 1987; Camerer and Hogarth, 1999; Ariely et al., 2009; Enke et al., 2023; Gneezy et al., 2024). By examining how the same individuals respond to a substantial change in stakes within the same decision environment, our paper complements this literature and provides insight into whether the supply-side mechanisms that generate biased beliefs remain operative, even when the potential costs of holding biased beliefs become very high.

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 whether students exhibit overconfidence in both low-stakes recall and high-stakes estimation environments. Section 5 investigates the supply side of biased beliefs by testing the role of biased memory and asymmetric updating in belief formation. Section 6 discusses heterogeneity results on gender and external influence. Section 7 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: math, English, Chinese, physics and chemistry, history and political science, and geography and 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.

Each Tier-1 high school assigns admission quotas to middle schools, thus students compete with their fellow students at the same middle school for Tier-1 seats. The quota is announced before the high school entrance exam. The schools' priority over students is fully defined by the students' entrance exam total score, which is unknown at the time of application. Students can apply for one Tier-1 high school only no matter how many Tier-1 schools are in a district.

The entrance exam score is announced one week after all school choices are submitted. The matching algorithm is then implemented. First, only students who scored above the median score among all students in the city are considered for Tier-1 high schools. Next, the choice of the qualified students are implemented. The Tier-1 schools accept the highest-ranked applicants at each middle school up to its quota. Unaccepted students become unmatched.

There is a "aftermarket" round for unmatched students and unassigned seats. Students who volunteered to participate in this round are assigned to a seat at a Tier-1 high school. Students with higher scores are assigned to schools that are deemed to be of higher quality according to the city education committee.

This matching mechanism's distinctive feature is that students must submit their school choice before they know their exam scores. As shown in Figure 1, the timeline of the admission process goes as follows. At the end of the third year of middle school, all students take the high school entrance exam simultaneously. Right after the 2-day exam, they are given the scoring rubric and answer keys to help them estimate their scores. Students are asked to submit their first choices seven days after the

exam without knowing the actual exam score.

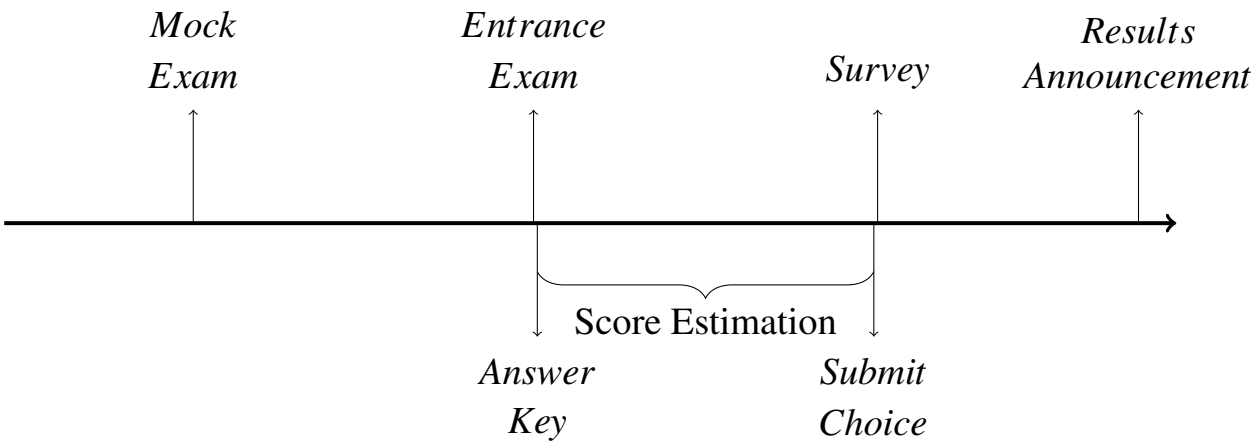

Figure 1: Timeline

Because the matching mechanism is Immediate Acceptance (IA), students have incentives to "game" the system. Importantly, this motivates them to estimate their scores accurately. Under a strategy-proof mechanism, a student only needs to report her preferences truthfully. Under IA, she must accurately estimate her score to "game" the system because the optimal strategy is contingent on the score. Consider the following example: student i prefers school a to school b. Based on past years, it is necessary to score 580 on the exam to get into school a, whereas 550 is enough for school b. Under a truth-telling matching algorithm, student i will list school a as her first choice. Under IA, she may do the same thing if her score exceeds 580. However, if her score is lower than 580, listing school b as the first choice might be optimal. Whether she scores above 580 or not becomes crucial, and she has strong incentives to estimate her score as accurately as possible.

Two groups of students are exempted from these high stakes. First, top-performing students do not need to estimate their scores accurately. This situation is reminiscent of the "Texas top 10% rule." For instance, the most prestigious high school in our study enrolls about 15% of students from each middle school. Therefore, students who rank in the top 10% of their cohort essentially have a guaranteed spot in any high school, which would reduce their motivation to accurately estimate their scores. Similarly, low-performing students who will not meet the threshold for Tier-2 high schools might also be unmotivated to produce accurate score estimates. Those falling below this benchmark have to retake the exam the following year to continue with their high school education.

2.1 Data

We use various data sources to analyze how students estimate their high school entrance exam scores. We surveyed two middle schools on the day students went back to school to submit their high school choice. The survey was given right after students submitted their high school choices. At this point, they had finished estimating their scores without knowing their final match. In the survey, we requested that students report their best-estimated scores for each subject exam and the possible score ranges (their highest and lowest possible scores). Additionally, we gathered 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.[2]

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 the middle schools for a subset of the students including their actual high school entrance exam scores on all six subjects, actual mock exam scores on all subjects, and basic demographic information, including gender and age.

We collected 302 valid survey responses.[3] We treat a student's score estimation for a subject exam as one observation. Because each student made six estimations, 302 students gave us 1812 observations on estimations.

[2]The 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.

[3]We sent out 321 questionnaires in total. Only 16 students failed to reported estimated scores for all six subjects in the questionnaire and are dropped from the analysis. In addition, three responses are dropped because their entrance exam scores are missing.

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 all 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 the majority of our sample (202 out of 302 students), schools explicitly collect the students' estimated scores and provide a predicted ranking based on all submitted estimates. Moreover, even for those students whose schools do not collect estimated scores, teachers typically have a stable 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 last 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.

3 Conceptual Framework

We first present a conceptual framework to describe the students' decision environment. After taking the entrance exam, student i estimates her performance τ_i in the exam based on her prior belief about her exam performance and the signal she receives from taking the exam. We assume that she forms her prior from her mock exam score, and updates her beliefs after getting the answer key and the scoring rubric. Formally, we define the prior and the signal as follow:

- True prior: $\tau_i \sim N(\tau_{mock,i} + \delta, \sigma_0^2)$, where $\tau_{mock,i}$ is the student i's actual score in the mock exam, δ is a measure of the difficulty of the mock exam relative to the entrance exam. The sum of the two serves as the mean of the true prior distribution. Here we introduce the δ term because the mock exam can be harder or easier than the entrance exam even though it is designed to be as similar as possible.

- Recalled prior: $\tilde{\tau}_i \sim N(\tilde{\tau}_{mock,i} + \delta, \sigma_0^2)$, where $\tilde{\tau}_{mock,i}$ is student i's recalled score in the mock

exam. It may or may not equals to the true score. We assume that students use their recalled prior instead of their actual prior when estimating their scores.

- Signal: $\tau_{exam,i} = \tau_i + \varepsilon_i$, where τ_i is the student's true performance in the entrance exam and $\varepsilon_i \sim N(0, \sigma^2_{\varepsilon_i})$ is the noise in the signal. As researchers, we can observe the actual entrance exam score τ_i ex-post, but the student can only observe a noisy signal $\tau_{exam,i}$ when they estimate their scores. While the student has access to the answer key, there is still uncertainty about their performance due to factors such as subjective grading in certain subjects and the accuracy of answer recall.

Since the student's prior belief and the signal error are normally distributed, and the error term is independent of the prior, we can use the properties of conjugate priors for normal distributions. In this case, the posterior distribution will also be normal. Bayes' rule implies that the posterior distribution of τ_i given $\tau_{exam,i}$ is: $\tau_i|\tau_{exam,i} \sim N(\hat{\tau}_i, \sigma^2_{1_i})$, where

$$\hat{\tau}_i = a_i\, \tilde{\tau}_{mock,i} + (1 - a_i)\tau_{exam,i} + a_i\delta. \tag{1}$$

$$\sigma_{1_i} = \sqrt{\frac{\sigma_0^2\sigma^2_{\varepsilon_i}}{\sigma_0^2 + \sigma^2_{\varepsilon_i}}} \tag{2}$$

, where

$$a_i = \frac{\sigma^2_{\varepsilon_i}}{\sigma_0^2 + \sigma^2_{\varepsilon_i}},\ 1 - a_i = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2_{\varepsilon_i}}.$$

We interpret $\hat{\tau}_i$ as the estimated score reported by the student. This estimated score $\hat{\tau}_i$ is a linear combination of the recalled mock exam score and the entrance exam score. The coefficients a_i and $1 - a_i$ indicate the relative weights the student assigns to their prior belief and the signal when estimating their performance.

To estimate the weight of the prior (mock exam score) and the signal (entrance exam score) on the estimated score, we use the following regression specification:

$$\hat{\tau}_{is} = \beta_1\, \tilde{\tau}_{mock,is} + \beta_2\, \tau_{exam,is} + \beta_1\, \delta_s + \theta_i + \varepsilon_{is} \tag{3}$$

, where $\hat{\tau}_{is}$ is student i's estimated score for subject s, δ_s is the difficulty of subject s in the mock exam relative to the entrance exam, θ_i is the individual fixed effect, and ε_{is} is the error term.

Bias in recall and estimation. To further investigate the role of biased recall in belief formation, we examine how the bias in mock exam score recall (i.e., the difference between the recalled and actual mock exam scores) contributes to the bias in the estimated entrance exam score (i.e., the difference between the estimated and actual entrance exam scores). We hypothesize that when students are more biased in their recall of the mock exam score for a particular subject, they will also be more biased in their estimation of the entrance exam score for that subject.

Let $\Delta\tau_{mock,is} = \tilde{\tau}_{mock,is} - \tau_{mock,is}$ denote the bias in recall for student i in subject s, and let $\Delta\tau_{is} = \hat{\tau}_{is} - \tau_{exam,is}$ denote the bias in estimation. We can model the relationship between these two biases as follows:

$$\Delta\tau_{is} = \gamma\Delta\tau_{mock,is} + \theta_i + \sigma_s + \varepsilon_{is} \tag{4}$$

where γ captures the effect of the bias in the recalled mock exam score on the bias in the estimated entrance exam score, θ_i and σ_s are individual and subject fixed effects, respectively, and ε_{is} is the error term. By including individual fixed effects (θ_i), we control for any time-invariant individual characteristics that may affect both the recall bias and the estimation bias, such as general overconfidence or memory capabilities. This allows us to focus on the within-person variation in recall bias across subjects and its impact on the estimation bias. If the coefficient γ is significantly different from zero, it suggests that when a student is more biased in recalling their mock exam score for a particular subject (relative to their average recall bias across all subjects), they are also more biased in estimating their entrance exam score for that subject (relative to their average estimation bias across all subjects). This would provide further evidence that the recalled scores are indeed used in the estimation process, and that biases in recall can contribute to biases in belief formation.

4 Demand for Biased Beliefs

To test how demand side forces in the form of stakes affect confidence, we look at both confidence in recalling a past performance, a low stake environment, and confidence in estimating performance on the actual entrance exam, a high stake environment. To understand to what extent our results were expected by experts, we run a survey with mostly Chinese economists who are familiar with the high school admission system and had some experience in estimating their exam scores in the past. We will first present the predictions by the experts and then examine the results on accuracy of both recall

and estimations.

Expert Predictions. We ask experts to predict three primary outcomes of our study: estimation error, recall error, and the correlation coefficient between the recall and the estimate. We provided the mean and standard deviations of the exam scores in both the entrance and mock exams. We posted the survey on the Social Science Prediction Platform and the Chinese experimental economists WeChat group (481 members) and received 51 valid responses.

Experts predicted overconfidence in the estimation and no overconfidence in the recall. The average predicted estimation error (estimated entrance exam score minus actual entrance exam score) is 0.13 standard deviations of the actual score, significantly different from 0 at the 1% level (Wilcoxon signed-rank test). The average predicted recall error (recalled mock exam score minus actual mock exam score) is 0.03 standard deviations of the actual mock exam score, which is not significantly different from 0 (p=0.119, Wilcoxon signed-rank test).

Experts also predicted a strong correlation between recall and estimation. On average, they predicted that a one standard deviation increase in recalled mock score is associated with a 0.35 standard deviation increase in estimated score after controlling for actual mock score (significantly different from 0 at the 1% level). This high level of expected correlation could be due to not accounting enough for the fact that we have controlled for the actual mock score when asking about the correlation coefficient between the two variables.

Error in Recall. We first examine whether students exhibit overconfidence in the low-stakes environment of recalling their mock exam performance. Error in recall is defined as the difference between the recalled and actual mock exam scores. 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 data 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 deviations 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 ones (see Appendix Figure A1 for details). The substantially higher proportion of upward versus downward

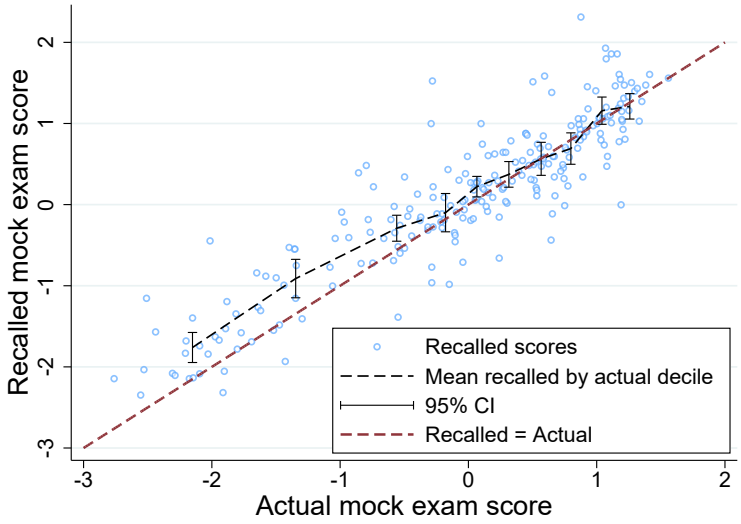

Figure 2: Recalled vs. actual mock exam scores

Note: 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 interval. 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.

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.

We also observe a consistent pattern of overconfidence at the subject level. As shown in Appendix Figure A2, students are overconfident at the subject level if we pool data for all six subjects, or if we look separately at STEM and non-STEM subjects. Interestingly, the difference between the recalled score and the actual score is larger among non-STEM subjects than among STEM subjects.[4] Looking at individual subjects, we find significant recall overconfidence in all subjects except Math, thus recall overconfidence is not driven by overconfidence in a specific subject.

Error in Recall and Mock Exam Performance. After establishing that there is widespread overconfidence in recall, we study whether this bias is motivated by testing the relationship between error in recall and the actual mock exam performance. Figure 2 illustrates this through decile analysis of actual scores, with 95% confidence intervals. Students in the bottom five deciles consistently recall

[4]This suggests that when recalling their past performance, students engage in a process of memory reconstruction rather than simply retrieving a stored number (Schacter and Addis, 2007). The reconstruction process is influenced by subject-specific schemas, with STEM subjects' objective nature potentially leading to more structured schemas and precise recall, while non-STEM subjects' subjective nature may result in less precise schemas and more variable recall.

a higher score than their actual one, with the magnitude of overconfidence being largest among the lowest performers and gradually decreasing as performance improves. In contrast, students in the top five deciles demonstrate notably accurate recall of their scores. One potential explanation for the correlation between performance and confidence in recall is that students who perform worse in the mock exam are more motivated to bias their beliefs upward to maintain a positive self-image about ability. An alternative explanation for the observed correlation between performance and confidence, based on the Dunning-Kruger effect (Kruger and Dunning, 1999), is that people who are incompetent in certain tasks lack the meta-awareness of their incompetence.

To distinguish between these explanations, we examine how recall accuracy varies within individuals across different subjects. This approach allows us to control for student-specific factors like sophistication in memory or meta-awareness by including individual fixed effects. If motivated beliefs drive recall bias, we should observe stronger upward bias in subjects where students perform relatively poorly, as these are the instances where maintaining positive self-image is most valuable.

Table 1: Performance and error in recall

	(1)	(2)	(3)
Dependent variable	Overconfidence in recall		
	Full	High performers	Low performers
Actual mock score	-0.288***	-0.439***	-0.277***
	(0.033)	(0.054)	(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

Note: This table reports the relationship between error 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, a one point increase in actual performance is associated with a 0.288 point decrease in recall overconfidence. This relationship is particularly strong among high-performing students (coefficient = -0.439) compared to low-performing ones (coefficient = -0.277), suggesting that

high performers exhibit stronger motivated memory when confronting subjects where they perform relatively poorly. This heightened sensitivity to poor performance among high performers is inconsistent with the Dunning-Kruger effect, because the high performers should be more aware of their competence across subjects.

Taken together, we find significant overconfidence in students' recall of mock exam scores, a pattern that starkly contrasts with experts' prediction of unbiased recall. We further show that low-performing students are more overconfident and, in general, students are more overconfident in subjects wherethey performed worse, suggesting that the overconfidence in memory serves a purpose of maintaining positive self-image rather than reflecting pure cognitive limitations.

Error in Estimation. We have shown that students are overconfident in the low-stakes environment of recalling mock exam scores. We now test whether they are overconfident when the stakes are high in estimating the entrance exam scores.

Error in estimation is defined as the difference between the estimated and the actual entrance exam scores, where a positive value indicates overconfidence. Figure 3 plots the relationship between estimated and actual entrance exam scores. In this high-stakes environment, we find no evidence of overconfidence. On average, the students' estimated scores are 0.04 standard deviations 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).

The distribution of error in estimation is also more balanced than error in recall. Only 2% of students accurately predict their score in the entrance exam. 146 (48%) students over-estimated their score and 149 (49%) students under-estimated their score. There is no clear tendency towards over-estimation. Appendix Figure A3 further shows that students' estimation mistakes are often small. In fact, more than 97% of students' estimated score deviate by less than one standard deviation from their actual 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. Appendix Figure A4 shows that pooling all subject exam estimations together, the estimated scores are, on average, 0.02 standard deviations lower than the actual scores. Students 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 the

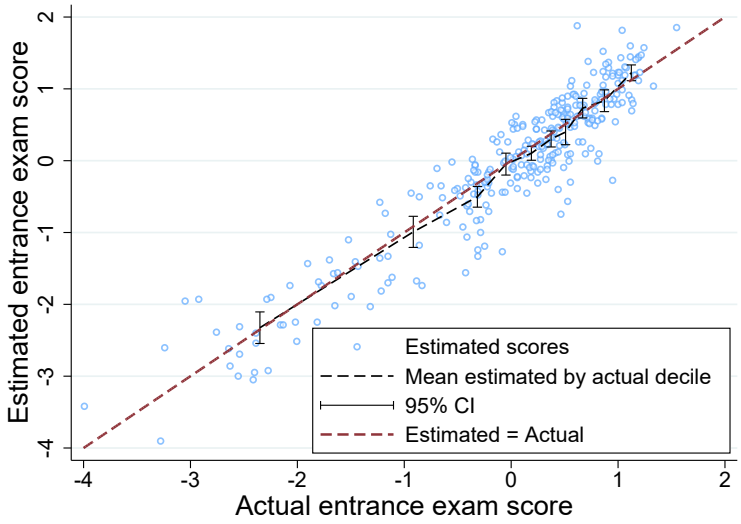

Figure 3: Estimated vs. actual entrance exam scores

Note: 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.

difference between estimated and actual scores never exceeding 0.25 standard deviation of the actual score.

Officially Reported Estimations. In one district, middle schools asked students to report their estimated exam scores to the school to better coordinate their school choices. We have access to the reported scores and can check whether officially reported estimations are accurate or not. In total, 202 students submitted their estimations to the school.

One concern about using officially reported estimations is that students may have incentives to over-report their (potentially accurately) estimated scores. The teacher could rank students based on the reported scores and the higher a student ranks, the better is the school recommended as her choice. However, students do not have to follow their teacher's recommendations. Thus, over-reporting estimated scores could discourage others from applying to a good school. The overconfidence as measured by this administrative data could be seen as an upper bound.

On average, the official estimated score is only higher by 0.01 standard deviation than the actual score. This difference is not only tiny in size but also not significantly different from 0 (p-value=0.680, Wilcoxon ranksum test). Consistent with the survey data, students' estimations are very close to the actual scores.

Error in Estimation and Exam Performance. We find that lower performance is associated with more overconfidence in recall, here we test whether we also see the same pattern with error in estimation. Figure 3 shows that low-performing students in the bottom five deciles are, on average, either accurate or slightly underconfident in estimation, which contrast sharply with their overconfidence in recall. We do not find a relationship between performance and error in estimation at the individual level.

We then test for the relationship between performance and estimation accuracy at the subject level within individual by regressing error in estimation on actual entrance exam scores while controlling for individual fixed effects. Table 2 shows that within individual, a one point increase in actual performance is associated with a 0.234 point decrease in error in estimation. 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). The fact that students are well-calibrated on average but more overconfident in subjects with worse performance means that they are overconfident in low-performing subjects but underconfident in high-performing subjects. One interpretation of this motivated bias pattern persisting in high-stakes estimation is that while students can overcome their tendency for overconfidence on average, they remain susceptible to motivational forces when forming beliefs about specific subjects. However, this pattern could also be consistent with a Bayesian model as specified in our conceptual framework if students place positive weight on their prior beliefs in forming their posterior. We examine this possibility in Section 5.

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.

We define a mistake as the difference between the rank of the highest-ranked school for which a student qualifies based on the actual score and the rank of the highest-ranked school for which they would qualify based on their estimated score. Figure A5 presents the distribution of mistakes. Negative values correspond to overconfidence, potentially leading to applications to overly competitive schools, while positive values correspond to underconfidence. To assess the role of stakes in estimation accuracy, we compare this distribution to the case where students exhibit the same degree of

Table 2: Performance and overconfidence in estimation

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

Note: 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 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 entrance exam, respectively. Standard errors, reported in parentheses, are clustered at the student level to accommodate multiple subjects per student. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

overconfidence in estimation as observed in their recall.

Our analysis reveals three findings. First, consistent with the no average overconfidence result, students demonstrate a high degree of accuracy in their estimations, with 82% making choices consistent with their actual scores. Second, the distribution of mistakes is relatively symmetric, with similar frequencies of over- and underestimation. This balance indicates no systematic bias in either direction under the baseline estimation model. Third, the alternative model, which assumes recall-level overconfidence, shows a marked increase in mistakes of shooting too high in the preference rankings. Notably, it predicts a higher frequency of severe overestimations (more than 3 ranks above qualification), which could result in non-assignment under the immediate acceptance system. The contrast between these models—with estimation errors being milder and more balanced—suggests that students adjust their behavior in response to the high stakes, exhibiting more cautious and accurate estimations than their recall performance would predict.

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 favorite school. We define this group as the top decile students at each school. As shown in Figure 3, the top decile students overestimated their scores by 0.1 standard deviations 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 deviations. However, this difference is not significantly different from 0 (P=0.861).

5 The Supply Side of Biased Beliefs

In the previous section, we test how overconfident beliefs react to demand side factors. Our findings indicate that students are overconfident when stakes are low in recall, but are not overconfident on average when stakes are high in estimation. Meanwhile, we observe substantial variations in confidence, even when stakes are high. In this section, we study the formation of (overconfident) beliefs and whether supply side factors can explain variations in confidence. In particular, we test the role of potentially biased recalls in belief formation, as in Bénabou and Tirole (2002), and the asymmetry in updating upon receiving good news and bad news.

Belief Formation. To test the effect of potentially biased recalls on belief formation, we run the regression, as specified in Equation 3, of estimated score in a subject on the actual score in that subject as the signal and the recalled mock exam score as the prior. We control for the individual fixed effect to account for individual-level tendencies to be over or under-confident in both estimation and recall.

We report the results in Table 3. First, we observe a large weight of the actual score on the estimated score. A one point increase in the actual score is associated with a 0.59 points increase in the estimated score. This suggests that students obtain a rather accurate signal about their performance in the exam after looking at the answer key and the scoring rubric. However, the fact that the coefficient is smaller than one indicates that the signal is still noisy.

Second, the weight of the recalled mock exam score is 0.27, which means a one point increase in the recalled mock exam score is associated with an around 0.27 increase in the estimated score. The total weights of the prior and the signal is close to one, which lends support to our model's implication

[5]The 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 the highest-ranked school. The result is qualitatively consistent if we instead look at the top 15% of students.

in Equation 1, and suggests that students' estimation process approximately follows Bayes' rule.

Third, we add the actual mock exam score as a control variable and show the results in Column (2). The recalled mock exam score remains a significant predictor of estimated scores . This indicates that students rely on their potentially biased memory when forming beliefs, despite having easy access to their actual mock exam scores. The relationship is also economically meaningful: given the same actual score, a one point increase in the recalled score in a subject leads to a 0.23 point increase in the estimated score.

Table 3: The supply side of overconfident beliefs

Dependent variable	(1)	(2)	(3)	(4)	(5)	(6)
	Estimated score					
	Full sample		STEM	NonSTEM	High performers	Low performers
Entrance exam score	0.586***	0.536***	0.724***	0.536***	0.490***	0.544***
	(0.037)	(0.037)	(0.057)	(0.055)	(0.051)	(0.047)
Recalled mock score	0.266***	0.226***	0.236***	0.276***	0.176***	0.310***
	(0.033)	(0.034)	(0.047)	(0.056)	(0.041)	(0.042)
Mock exam score		0.112***				
		(0.030)				
Individual fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Subject fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1478	1478	740	738	733	745
R-squared	0.940	0.941	0.955	0.915	0.948	0.910

Note: This table reports how students combine their recalled mock exam scores and the perceived entrance exam performance when forming their estimated scores. The dependent variable is the estimated entrance exam score. 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 accuracy. Column (1) presents results for the full sample without controlling for actual mock exam scores, column (2) adds actual mock exam scores as a control, columns (3) and (4) present results separately for STEM and non-STEM subjects, while columns (5) and (6) present results separately for students who scored above and below the median in the entrance exam. Standard errors, reported in parentheses, are clustered at the student level to accommodate multiple subjects per student. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Signal Accuracy. When the signal becomes less noisy, our model predicts that the student places more weight on the signal (τ_{exam}) and less weight on their prior belief ($\tilde{\tau}_{mock}$) when estimating performance in the entrance exam (τ). To test this prediction, we divide the sample into STEM subjects and non-STEM subjects and estimate the weights in each sub-sample. Thanks to the more objective nature of STEM subjects, the signals students get after taking STEM subject exams are less noisy than signals they get after taking non-STEM ones, which can be heavily influenced by the grader.

Thus,the model predicts that students put a larger weight on the signal (the entrance exam score) and a smaller weight on the prior (recalled mock exam score) in STEM subjects than in non-STEM subjects. Column (3) and (4) of Table 3 show that students indeed place more weight on the signal when the signal is less noisy. The weight of the signal on estimated score is almost doubled in STEM subjects than non-STEM subjects, and the difference is significant at the 1% level. At the meantime, students rely more on their recalled mock exam scores in non-STEM subjects than in STEM subjects when estimating their scores. The weight of recalled mock exam scores in non-STEM subjects is 0.11 higher than in STEM subjects.

Performance and Belief Formation. The relationship between recalled scores and estimation differs notably between high and low performers. As shown in column (5) and (6) of Table 3, low-performing students place substantially more weight on their recalled mock exam scores compared to high-performing students (0.310 versus 0.176, significant at 5% level), while the weights they place on the actual entrance exam score are similar (0.544 versus 0.490, $P = 0.438$). This pattern indicates that low performers rely more heavily on their potentially biased memory when forming beliefs about their performance. One interpretation is that low-performing students, who tend to have more positively biased recall as shown earlier, might derive greater psychological benefit from maintaining these biased beliefs even in their high-stakes estimation.

Bias in Recall and Bias in Estimation. Next, we directly test the role of biased recall as a supply side factor leading to a bias in belief. Following Equation 4, we regress overconfidence in estimation, which is the difference between the estimated and the actual exam score, on overconfidence in recall, which is the difference between the recalled and the actual mock exam score. To account for factors that contribute both to overconfidence in recall and overconfidence in estimation, we control for the individual fixed effects and cluster the standard errors at the individual level. The regression results, as shown in Table 4, indicate that within individual, a one point increase in recall bias is associated with a 0.1 point increase in estimation bias, and this relationship is significantly different from zero at the 1% level.

Low-performing students exhibit significantly stronger transmission of recall bias to estimation bias compared to high-performing ones (0.129 vs 0.060, difference significant at 5% level), which is consistent with motivated bias because low performers have stronger incentive to maintain a high self-image and potentially face a lower cost of decision mistakes. The most direct evidence supporting

motivated bias is that we find complementarity between bias in recall and the transmission of the bias to estimation. In columns (6) and (7) of Table 4, we find transmission when there is overconfidence in recall, but no significant relationship between the two biases when there is no overconfidence (0.201 vs 0.043, difference significant at 1% level)[6]. This result is consistent with students inflating their recalled performance in the past more when they place more weight on the recalled performance in estimating their actual exam performance.

Table 4: Overconfidence in estimation and overconfidence in recall

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Overconfidence in estimation						
	Full sample	STEM	Non-STEM	High performers	Low performers	Recall Over-confident	Not recall over-confident
Overconfidence in recall	0.092***	0.094**	0.111**	0.060*	0.129***	0.201***	0.043
	(0.027)	(0.040)	(0.053)	(0.032)	(0.036)	(0.071)	(0.056)
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subject FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1478	740	738	733	745	767	711
R-squared	0.476	0.509	0.637	0.571	0.430	0.561	0.574

Note: This table reports the relationship between overconfidence in recall and overconfidence in estimation. Overconfidence in estimation, the dependent variable, is measured by the difference between the estimated entrance exam score in a subject and the actual entrance exam score in that subject. Overconfidence in recall is measured by the difference between the recalled mock exam score in a subject and the actual mock exam score in that 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 accuracy. Column (1) presents results for the full sample, columns (2) and (3) present results separately for STEM and nonSTEM subjects, columns (4) and (5) present results separately for students who scored above and below the median in the entrance exam, and columns (6) and (7) present results separately for subjects where the recalled score is above the actual mock exam score and subjects where the recalled score is not above the actual mock exam score. Standard errors, reported in parentheses, are clustered at the student level to accommodate multiple subjects per student. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Lastly, despite the significant correlation between the two biases, only around 10% of the bias in recall is transmitted into the bias in estimation, the posterior belief. This transmission rate of bias is notably smaller than the weight of the recalled score on the estimated score, which is around 0.3. This low rate of transmission enables the estimation to be accurate on average when the recall is

[6]The results are similar if we instead look at underconfidence in recall (recalled score $<$ actual score), which suggests that this difference is not mechanical.

overconfident. One interpretation is that people have some mega-awareness of their tendency to be overconfident, and try to limit the role of biased recall in forming their beliefs when the stakes are high.

The Good News/Bad News Asymmetry. Another channel of supplying biased beliefs is asymmetrical updating depending on whether the news is good or bad (Bénabou and Tirole, 2016). We test the relevance of this channel in our environment by studying how students' score estimations respond to good news and bad news. The extensive empirical literature on the good news/bad news asymmetry in belief updating literature is mixed. While Eil and Rao (2011), Wiswall and Zafar (2015), Charness and Dave (2017) and Möbius, Niederle, Niehaus and Rosenblat (2022) find that people are more responsive to positive feedback than to negative feedback, other studies find either no asymmetry in belief updating (Grossman and Owens, 2012; Buser, Gerhards and Van Der Weele, 2018; Barron, 2020) or exactly the opposite: people react more to bad news than to good news (Ertac, 2011; Coutts, 2019).

In a lab experiment, the good news (bad news) is defined as signals that leads the decision maker to update their belief up (down) towards the good state relative to the prior. In our environment, good news (bad news) can be defined as higher (lower) performance in the entrance exam (the signal) than in the mock exam (the prior). We also define neutral news as similar performance in the entrance exam as in the mock exam, which serves as a Bayesian benchmark. Even though students did not know whether they performed better in the entrance exam than in the mock exam when estimating their scores, they should have received the noisy signal in the right direction of good or bad news after receiving the answer keys and scoring rubric. To compare students' performances in the entrance and the mock exam, we standardize these two scores by ranking students' subject scores within their school. We define good news as performing 10 ranks higher in the entrance exam, bad news as performing 10 ranks lower in the entrance exam, and neutral news otherwise. Our results are robust to alternative cutoffs (see Appendix Table A5 and Appendix Table A6 for details).

To test whether students update differently upon receiving good versus bad news, we regress the estimated score on the entrance exam score and the recalled mock exam score under different news. Again, we control for the individual fixed effect and subject fixed effect in the regression. Thus, the comparison is within individual between subjects the student performed well relative to the mock exam and subjects they performed poorly relative to the mock exam. Controlling for the individual

fixed effects addresses the concern that students who receive the good news and students who receive the bad news are fundamentally different.

Results in Table 5 indicate that students place a larger weight on the signal (the entrance exam score) when the news is more informative regardless of whether the news is good (0.622) or bad (0.573) than when the news is less informative (neutral, 0.539). They also place a lower weight on the prior (the recalled mock exam score) when receiving good or bad news compared to neutral news. This differential response to informative and uninformative news suggests that students are able to discern the strength of the signal and update accordingly.

Meanwhile, the weight on the signal is significantly larger when the news is good than when it is neutral at 5% level, but the weight is not significantly different when the news is bad versus neutral. The weight on the signal is larger when the news is good than when it is bad, but the difference is not significant.

Taken together, these results suggest that students in our environment update more when the signal is more informative and update asymmetrically: they weight the signal more when the news is good than when the news is uninformative, but not more when the news is bad. Thus in a field setting with non-binary states, we find directional evidence consistent with Eil and Rao (2011), Wiswall and Zafar (2015), Charness and Dave (2017), and Möbius, Niederle, Niehaus and Rosenblat (2022).

6 Discussion

Gender and Overconfidence. Previous research suggests that men tend to be more overconfident than women (Niederle and Vesterlund, 2007; Bordalo et al., 2019; Iriberri and Rey-Biel, 2021; Exley and Kessler, 2022; Coffman et al., 2024). Our setting allows us to examine gender differences in both low-stakes recall and high-stakes estimation. Panel A of Table 6 shows that male students exhibit significantly more overconfidence in recall compared to female students, with their recalled scores being 0.09 standard deviations (2.186 points) higher than female students' recalled scores given the same mock exam score (significant at 1% level). This gender difference is robust to controls for age, parental education, and self-reported risk preference. This gender gap in overconfidence in recall is particularly pronounced in STEM subjects (2.796 points) compared to non-STEM subjects (1.570 points), indicating that male students are more overconfident in their gender-congruent fields (Bordalo

Table 5: Good news, bad news and the supply side

	(1)	(2)	(3)
Dependent	Estimated score		
Variable	Good news	Bad news	Neutral news
Actual entrance exam score	0.622***	0.573***	0.539***
	(0.127)	(0.126)	(0.084)
Recalled mock exam score	0.216***	0.199*	0.343***
	(0.074)	(0.079)	(0.073)
Individual fixed effect	Yes	Yes	Yes
Subject fixed effect	Yes	Yes	Yes
Observations	453	496	529
R-squared	0.969	0.961	0.959

Note: This table reports how students' score estimation respond to good news, bad news and neutral news. The dependent variable is the estimated entrance exam score. 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 accuracy. Column (1) presents results for good news, which is defined as scored 10 ranks higher in the entrance exam than in the mock exam in a subject. Column (2) presents results for bad news, which is defined as scored 10 ranks lower in the entrance exam than in the mock exam in a subject, and columns (3) present results for neutral news, which is defined as scored within 10 ranks in the two exams. Standard errors, reported in parentheses, are clustered at the student level to accommodate multiple subjects per student. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

et al., 2019; Coffman et al., 2024; Li and Zhang, 2024). Interestingly, the gender gap is larger among high performing students than among low performing students, potentially caused by low-performing female students' stronger motivation to be overconfident in recall as well.

Table 6: Gender differences in overconfidence

Panel A: Overconfidence in recall

	(1)	(2)	(3)	(4)	(5)	(6)
	Full sample		STEM	Non-STEM	High performers	Low performers
Male	2.186***	2.309***	2.796***	1.570*	2.548**	1.834*
	(0.709)	(0.677)	(0.844)	(0.880)	(0.987)	(1.079)
Class fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Subject fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Demographic controls	No	Yes	No	No	No	No
Observations	1478	1430	740	738	733	745
R-squared	0.095	0.119	0.164	0.049	0.104	0.103

Panel B: Overconfidence in estimation

	(1)	(2)	(3)	(4)	(5)	(6)
	Full sample		STEM	Non-STEM	High performers	Low performers
Male	1.108**	1.318**	1.600***	0.616	1.059	1.641*
	(0.560)	(0.566)	(0.584)	(0.696)	(0.808)	(0.869)
Class fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Subject fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Demographic controls	No	Yes	No	No	No	No
Observations	1812	1704	906	906	900	912
R-squared	0.152	0.172	0.082	0.211	0.173	0.171

Note: This table reports gender differences in overconfidence in recall (Panel A) and overconfidence in estimation (Panel B). In Panel A, the dependent variable "Overconfidence in recall" is defined as the difference between recalled and actual mock exam scores. In Panel B, the dependent variable "Overconfidence in estimation" is defined as the difference between estimated and actual entrance exam scores. The independent variable "Male" is a dummy variable that equals 1 for male students and 0 for female students. All specifications include class and subject fixed effects. Demographic controls include age, parental education, and self-reported risk preference. Columns (1) and (2) present results for the full sample without and with demographic controls. Columns (3) and (4) present results separately for STEM and NonSTEM subjects. Columns (5) and (6) present results separately for students who scored above and below the median in the entrance exam. Standard errors, reported in parentheses, are clustered at the student level to accommodate multiple subjects per student. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Results in Panel B of Table 6 indicates that while overconfidence does not carry over to the high-stakes estimation decisions, the gender gap in overconfidence is persistent despite high stakes. Male students' estimated scores are 0.05 standard deviations (1.108 points) higher than those of female students with the same entrance exam score, and the difference is significant at the 5% level. However, this gap is also notably smaller than the gap in overconfidence in recall. Moreover, the gender gap in estimation is only significant in STEM subjects (1.600 points) but not in non-STEM subjects (0.616 points). Thus, male students are only overconfident in their gender-congruent fields when stakes

are high. The gender gap in confidence is larger among low-performing students than among high-performing students (1.641 versus 1.059), even though the difference is not significant. This result of a more pronounced gender confidence gap at low performance levels is consistent with previous findings (Exley and Kessler, 2022).

Our environment also allows us to dig deeper into the gender gap in confidence and ask whether the gap is due differences in prior beliefs or differences in updating (Coffman et al., 2024). We find evidence for the former but not the latter. While male and female students differ in their recalled score in the mock exam, they are strikingly similar in their updating process. Appendix Table A4 reports the weights of the prior belief (recalled mock exam score) and the signal (actual entrance exam score) on the estimated score by gender. The weights placed on the prior and the signal are not significantly different between male and female students. Both male and female students place a larger weight on the signal in STEM subjects versus non-STEM subjects, reflecting the smaller noise associated with the signal in STEM subjects.

Taken together, we find that male students are more confident than female students in our high-stakes environment, but the size of the gap is moderated by stakes. The gap is more pronounced in the male-congruent STEM subjects. The gender gap is due to male students holding more overconfident priors, not by differences in how they updating their beliefs.

External Influence. Another explanation for the lack of overconfidence in score estimation is that students in our sample are heavily influenced by external forces, such as parents, teachers, or their peers, during the estimation process. These external influences might help them correct the overconfidence bias and thus enable them to estimate their scores more accurately. We asked students in the survey whether they received assistance from their parents, teachers, or peers while estimating their scores. In our sample, 39.7% of students received help from their parents, 33.4% were aided by their teachers, and 45.4% collaborated with their peers in estimating their scores.

We regress the estimated subject scores on dummy variables representing different types of external help, with the results presented in Appendix Table A3. Contrary to what one might expect, we do not find that external help is correlated with less overconfidence. In fact, receiving external help seems to be associated with a slight increase in overconfidence, although this effect is small and not statistically significant in most specifications. However, we must be cautious in interpreting these results, as the choice to seek external help is not exogenous; it is influenced by other factors that could

potentially introduce selection bias, thereby obscuring the true effect of external help.

7 Conclusion

Our study contributes to understanding whether biased beliefs are innate or optimally chosen by examining them in the high-stakes environment of Chinese high school admissions. We report several key findings. First, while students exhibit significant overconfidence in the low-stakes context of recalling their mock exam performance, they show no systematic bias in the high-stake environment of estimating their entrance exam scores. Second, within individual, both overconfidence in recall and overconfidence in estimation increase in subjects with worse performance, suggesting a motivated component to belief formation. Third, students rely on their biased memory when forming beliefs about their entrance exam performance, even though accurate information about past performance is readily available. Fourth, students update their beliefs asymmetrically, being more responsive to good news than bad news, consistent with theories of motivated information processing. Fifth, while we observe substantial gender differences in confidence under low-stakes recall and high-stakes estimation, with males being more overconfident especially in STEM subjects, the gender gap in overconfidence becomes notably smaller in high-stakes estimation.

These findings help reconcile two views of biased beliefs in the literature. On one hand, the absence of average overconfidence in high-stakes estimation supports models of motivated bias in beliefs where beliefs are optimally chosen after weighing costs and benefits. People will seek the least costly distortions of their beliefs while still maintaining positive self-image, in this case, being overconfident about past performance. On the other hand, the systematic patterns in how bias varies with performance, the role of biased memory, and asymmetric updating suggest that even in high-stakes environments, belief formation remains subject to motivated forces.

Our results have broader implications for policy design in contexts where accurate belief formation is crucial. While high stakes can discipline average overconfidence, they do not eliminate motivated biases in forming relative beliefs across domains. This suggests that mechanisms designed to reduce overconfidence should focus not just on raising stakes but also on mitigating specific channels through which motivated beliefs persist, such as selective memory and asymmetric updating. These insights are particularly relevant for contexts like financial markets, where investors may maintain accurate beliefs about overall market conditions while exhibiting motivated biases about relative

performance across different investments.

REFERENCES

Ariely, Dan, Uri Gneezy, George Loewenstein, and Nina Mazar, "Large Stakes and Big Mistakes," *Review of Economic Studies*, 2009, *76* (2), 451–469.

Bandiera, Oriana, Nidhi Parekh, Barbara Petrongolo, and Michelle Rao, "Men are from Mars, and Women Too: A Bayesian Meta-analysis of Overconfidence Experiments," *Economica*, 2022, *89*, S38–S70.

Barron, Kai, "Belief updating: does the 'good-news, bad-news' asymmetry extend to purely financial domains?," *Experimental Economics*, 2020, pp. 1–28.

Belot, Michèle, Bhaskar Bhaskar, and Jeroen van de Ven, "Promises and Cooperation: Evidence from a TV Game Show," *The Journal of Economic Behavior 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.

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.

Brunnermeier, Markus K and Jonathan A Parker, "Optimal expectations," *American Economic Review*, 2005, *95* (4), 1092–1118.

Buser, Thomas, Leonie Gerhards, and Joël Van Der Weele, "Responsiveness to feedback as a personal trait," *Journal of Risk and Uncertainty*, 2018, *56* (2), 165–192.

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 Lovallo, "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.

Charness, Gary and Chetan Dave, "Confirmation bias with motivated beliefs," *Games and Economic Behavior*, 2017, *104*, 1–23.

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.

Coutts, Alexander, "Good news and bad news are still news: Experimental evidence on belief updating," *Experimental Economics*, 2019, *22* (2), 369–395.

Eil, David and Justin M Rao, "The good news-bad news effect: asymmetric processing of objective information about yourself," *American Economic Journal: Microeconomics*, 2011, *3* (2), 114–38.

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.

Ertac, Seda, "Does self-relevance affect information processing? Experimental evidence on the response to performance and non-performance feedback," *Journal of Economic Behavior & Organization*, 2011, *80* (3), 532–545.

Exley, Christine L and Judd B Kessler, "The gender gap in self-promotion," *The Quarterly Journal of Economics*, 2022, *137* (3), 1345–1381.

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.

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.

Grossman, Zachary and David Owens, "An unlucky feeling: Overconfidence and noisy feedback," *Journal of Economic Behavior & Organization*, 2012, *84* (2), 510–524.

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.

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

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.

Köszegi, 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.

Li, Fanghua and Y. Jane Zhang, "Response to Competition: Gender, Domains, and STEM choice," *Working Paper*, 2024.

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.

Möbius, Markus M, Muriel Niederle, Paul Niehaus, and Tanya S Rosenblat, "Managing self-confidence: Theory and experimental evidence," *Management Science*, 2022, *68* (11), 7793–7817.

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

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

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

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.

Schacter, Daniel L and Donna Rose Addis, "The cognitive neuroscience of constructive memory: remembering the past and imagining the future," *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2007, *362* (1481), 773–786.

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.

Wiswall, Matthew and Basit Zafar, "How do college students respond to public information about earnings?," *Journal of Human Capital*, 2015, *9* (2), 117–169.

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

Appendix

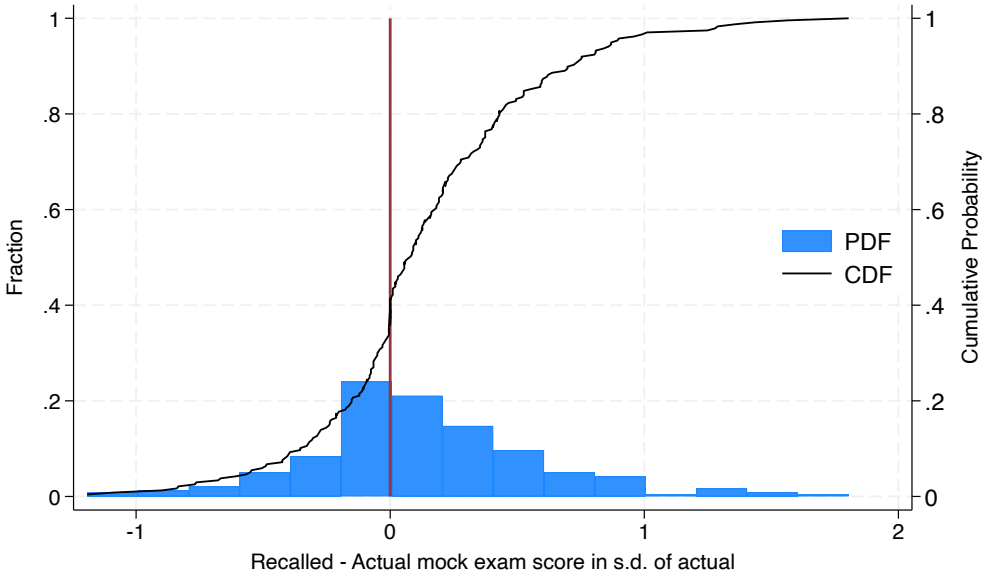

Figure A1: Confidence in recall: distribution

Note: 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.

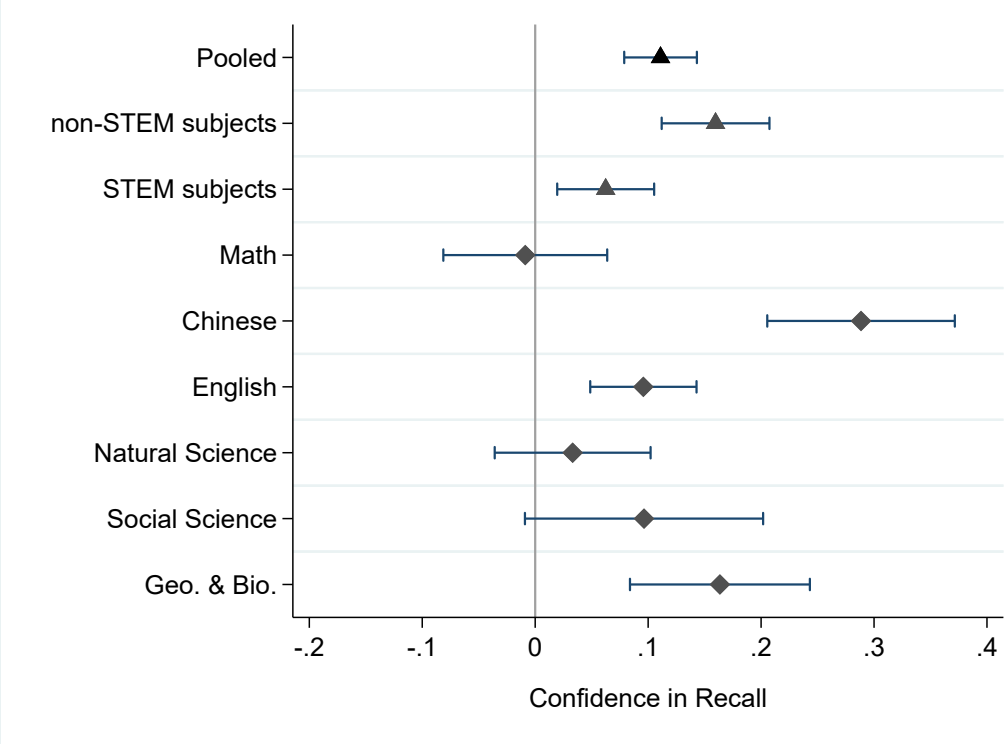

Figure A2: Confidence in recall: Subjects

Note: 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 and Social Science; "STEM subjects" refers to pooling data of three STEM subjects, namely Math, Natural Science and Geology and Biology.

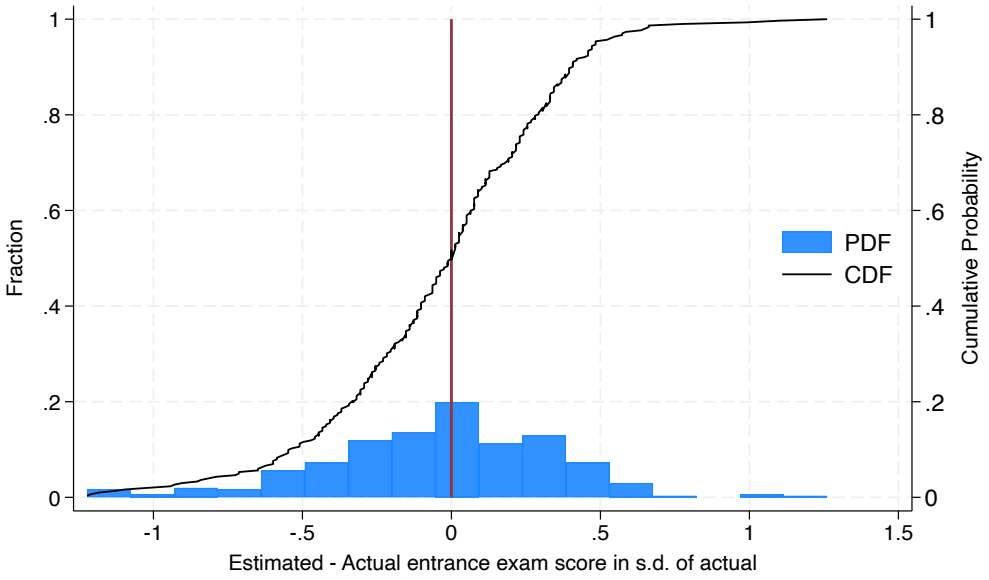

Figure A3: Confidence in estimation: distribution

Note: 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.

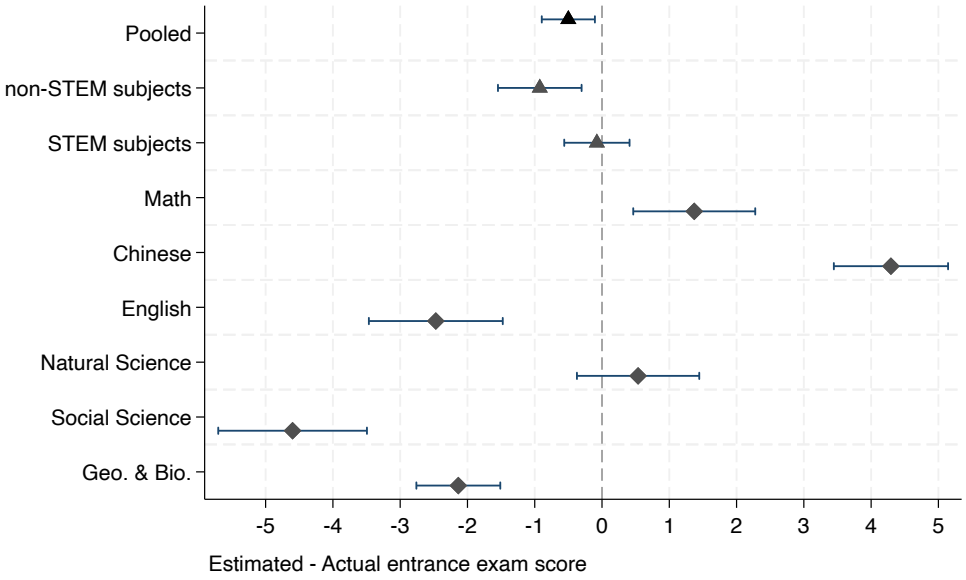

Figure A4: Confidence in estimation: Subjects

Note: This graph reports the confidence in estimation 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 and Social Science; "STEM subjects" refers to pooling data of three STEM subjects, namely Math, Natural Science and Geology and Biology.

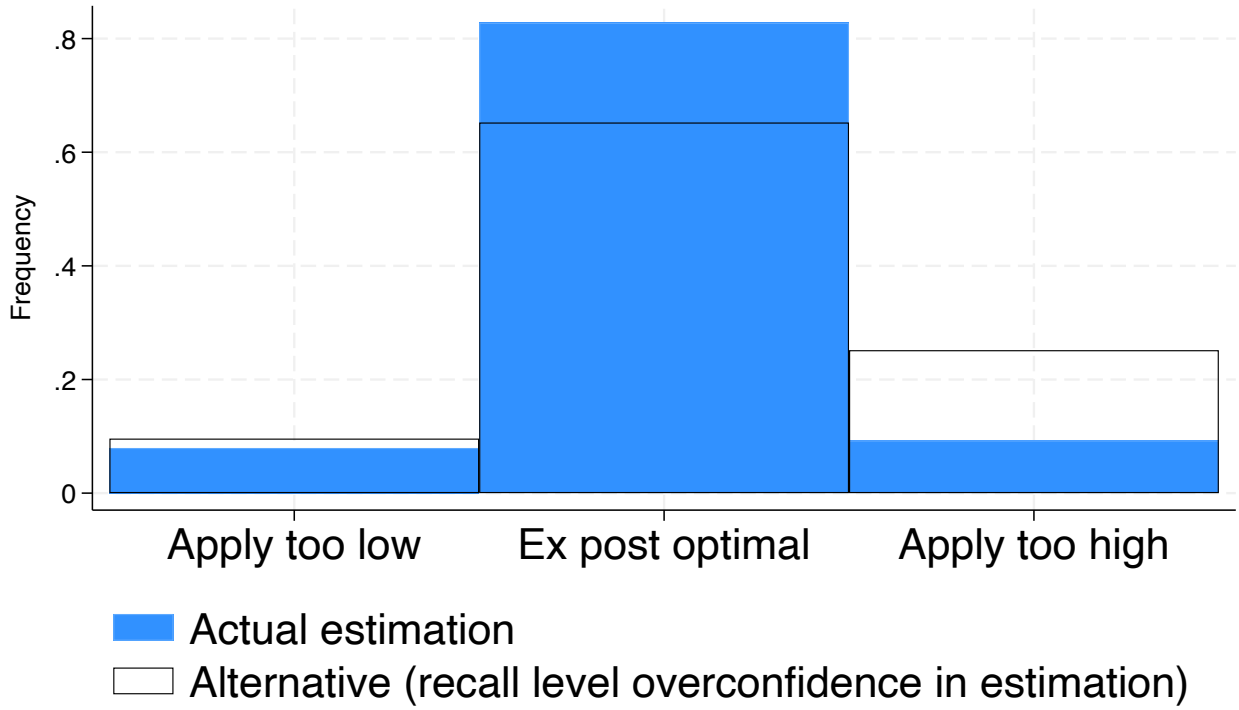

Figure A5: Distribution of mistakes in school choice

Note: Negative values indicate applying to a school too good for the actual score and positive values indicate applying to a lower-ranked school than necessary.

Table A1: Summary Statistics

Full Sample			
Variable	Obs	Mean	Std. Dev.
Entrance exam score	1812	80.50	21.24
Estimated entrance exam score	1812	80.00	22.12
Mock exam score	1804	73.87	23.56
Recalled mock exam score	1485	75.21	23.67
"Confidence in the Recall" Sample			
Variable	Obs	Mean	Std. Dev.
Entrance exam score	1478	79.92	21.66
Estimated entrance exam score	1478	79.07	22.44
Mock exam score	1478	73.55	24.02
Recalled mock exam score	1478	75.16	23.71

Table A2: Survey Selection

VARIABLES	(1) Survey	(2) Survey
Official estimated score	-0.000	-0.001
	(0.001)	(0.001)
Zhongkao total score	0.001	0.001
	(0.001)	(0.001)
Male	0.029	0.039
	(0.070)	(0.066)
Class Fixed Effect	No	Yes
Observations	202	202
R-squared	0.028	0.174

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A3: External influence and confidence

	(1) Estimated score	(2) Estimated score	(3) Estimated score
Parent help	0.971	0.134	0.587
	(0.713)	(0.730)	(0.771)
Teacher help	0.290	0.590	-0.145
	(0.772)	(0.730)	(0.759)
Peer help	0.473	0.857	1.213*
	(0.596)	(0.661)	(0.640)
Actual score	0.963***	0.914***	0.909***
	(0.013)	(0.023)	(0.022)
Class fixed effect	No	Yes	Yes
Subject fixed effect	No	Yes	Yes
Demographics	No	No	Yes
Observations	1812	1812	1722

Note: This table presents the results of regression analyses exploring the influence of external help on students' self-estimated scores. The dependent variable in all columns is "Estimated score," representing the students' estimated scores in a particular subject. "Parent help," "Teacher help," and "Peer help" are binary variables denoting whether a certain type of help was received. A student could receive help from multiple sources. "Actual score" reflects the students' actual scores in the entrance exam. The "Demographics" control includes gender, father's education, mother's education, and risk preference. The regressions in columns (2) and (3) control for class and subject fixed effects, as indicated. The number of observations drops in column (3) due to missing data on parents' education. Standard errors are reported in parentheses below the coefficient estimates. Significance levels: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table A4: The supply side of overconfident beliefs by gender

Dependent variable	(1)	(2)	(3)	(4)	(5)	(6)
	Estimated score					
			Male		Female	
	Male	Female	STEM	non-STEM	STEM	non-STEM
Entrance exam score	0.620***	0.597***	0.709***	0.555***	0.659***	0.545***
	(0.044)	(0.057)	(0.054)	(0.055)	(0.053)	(0.098)
Recalled mock score	0.296***	0.366***	0.239***	0.348***	0.320***	0.412***
	(0.033)	(0.053)	(0.043)	(0.041)	(0.042)	(0.100)
Class fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Subject fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Observations	794	684	397	397	343	341
R-squared	0.896	0.917	0.913	0.833	0.938	0.834

Note: This table reports how male and female students combine their recalled mock exam scores and the perceived entrance exam performance when forming their estimated scores. The dependent variable is the estimated entrance exam score. The observation unit is student-subject. All specifications include class and subject fixed effects to control for class-level and subject-specific factors that might affect estimation accuracy. Column (1) and (2) presents results for the male and female students respectively, columns (3) and (4) present results separately for STEM and non-STEM subjects for male students, while columns (5) and (6) present STEm and non-STEM subject results separately for female students. 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 A5: Good news, bad news and the supply side: 5-rank margin

	(1)	(2)	(3)
Dependent	Estimated score		
Variable	Good news	Bad news	Neutral news
Actual entrance exam score	0.613***	0.594***	0.537***
	(0.098)	(0.098)	(0.131)
Recalled mock exam score	0.274***	0.204***	0.331**
	(0.077)	(0.068)	(0.132)
Individual fixed effect	Yes	Yes	Yes
Subject fixed effect	Yes	Yes	Yes
Observations	558	604	316
R-squared	0.954	0.956	0.972

Note: This table reports how students' score estimation respond to good news, bad news and neutral news when good and bad news are defined as performing 5 ranks better in the entrance exam than in the mock exam. The dependent variable is the estimated entrance exam score. 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 accuracy. Column (1) presents results for good news, which is defined as scored 5 ranks higher in the entrance exam than in the mock exam in a subject. Column (2) presents results for bad news, which is defined as scored 5 ranks lower in the entrance exam than in the mock exam in a subject, and columns (3) present results for neutral news, which is defined as scored within 5 ranks in the two exams. 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 A6: Good news, bad news and the supply side: 15-rank margin

	(1)	(2)	(3)
Dependent	Estimated score		
Variable	Good news	Bad news	Neutral news
Actual entrance exam score	0.623***	0.580***	0.556***
	(0.155)	(0.161)	(0.063)
Recalled mock exam score	0.222**	0.154*	0.325***
	(0.087)	(0.085)	(0.055)
Individual fixed effect	Yes	Yes	Yes
Subject fixed effect	Yes	Yes	Yes
Observations	375	399	704
R-squared	0.968	0.968	0.956

Note: This table reports how students' score estimation respond to good news, bad news and neutral news when good and bad news are defined as performing 15 ranks better in the entrance exam than in the mock exam. The dependent variable is the estimated entrance exam score. 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 accuracy. Column (1) presents results for good news, which is defined as scored 15 ranks higher in the entrance exam than in the mock exam in a subject. Column (2) presents results for bad news, which is defined as scored 15 ranks lower in the entrance exam than in the mock exam in a subject, and columns (3) present results for neutral news, which is defined as scored within 15 ranks in the two exams. Standard errors, reported in parentheses, are clustered at the student level to accommodate multiple subjects per student. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.