Updating and Misspecification: Evidence from a Field Experiment*

Marc-Antoine Châtelain† Paul Han‡ En Hua Hu§ Xiner Xu†

August, 2025

Abstract

Misspecification is theoretically linked to failures in belief updating, but empirical evidence remains scarce. Using a field experiment in a university course, we show that misspecified beliefs are a major barrier to accurate updating. Students remain overconfident despite receiving informative test scores, and they substantially overestimate the noisiness in test scores. A randomized controlled trial providing impersonal information about test score noisiness significantly improves students' predictions, closing up to one-third of the gap with a Bayesian benchmark. These results show that misspecification is an important constraint on belief updating but can be mitigated through information interventions.

keywords: belief updating, misspecification, student learning

JEL codes: C93, D83, D84, D91, I23

*We are grateful to all the instructors and students who participated in this study, with special thanks to Xiaoyue Cui and Bernardo Galvão-Souza for their exceptional contributions. We also thank Michael Baker, Guangbin Hong, Robert McMillan, Philip Oreopoulos, Xincheng Qiu, and Basit Zafar for their help and comments. The authors gratefully acknowledge the Department of Economics at the University of Toronto for funding. This research was approved by the research ethics board of the University of Toronto, protocol #00042772.

†University of Toronto

‡University of Toronto; Competition Bureau

§University of Oxford

1 Introduction

Belief formation plays a central role in decision-making, and individuals often rely on performance signals to guide their choices. Students use test results to select academic paths, workers interpret supervisor feedback when deciding whether to seek promotions or change jobs, and managers assess progress reports to reallocate resources or revise timelines. The ability to interpret performance signals accurately is therefore critical, yet individuals often struggle to update their beliefs appropriately in response to new information. A large literature in economics and psychology examines whether individuals update beliefs in a Bayesian manner (Benjamin, 2019), with growing evidence of systematic deviations due to behavioral biases such as overconfidence and motivated reasoning (Ertac, 2011; Eil and Rao, 2011; Buser et al., 2018; Coutts, 2019; Drobner, 2022; Möbius et al., 2022), or to the computational complexity of Bayesian updating (Grether, 1980; Amelio, 2022; Guan, 2023; Gonçalves et al., 2024).

A distinct and comparatively underexplored source of updating error arises when individuals hold incorrect beliefs about the structure of the environment, a mechanism referred to as *misspecification* in the theoretical literature (Berk, 1966; Heidhues et al., 2018; Frick et al., 2020, 2023; Fudenberg et al., 2021; Bohren and Hauser, 2025). In such cases, agents may apply Bayes' Rule correctly but do so using an incorrect model of how signals are generated. Distinguishing between biased updating and misspecification is not only conceptually important but also practically consequential: whereas behavioral biases are often resistant to change, structural misperceptions, such as overestimating the noisiness of performance signals, can plausibly be corrected through information interventions.

This paper provides the first field-experimental evidence that misspecified beliefs about the noisiness of performance signals constitute a first-order barrier to effective belief updating. We study belief formation in the context of a large first-year calculus course at the University of Toronto, and show via a randomized controlled trial that correcting misperceptions about test score noise significantly improves students' ability to form accurate expectations about their future performance.

To conduct our analysis, we collect rich survey data from students over the academic year and link it to administrative records of their performance on all major course assessments. Our panel consists of five surveys, administered shortly before each test, which elicit students' point and probabilistic grade predictions, and their beliefs about the role of randomness in shaping outcomes and prediction errors.[1] Participation is incentivized through small grade

[1] Collecting belief regarding grades is not sufficient to distinguish between misspecification and other updating failures, see Bohren and Hauser (2025).

rewards, and truthful reporting is encouraged by offering students the chance to win monetary prizes based on the accuracy of their predictions, resulting in high response rates (82 to 88% of test-takers complete each survey) and a high-quality panel dataset.

Four central patterns emerge. First, students display substantial overconfidence: across all five tests, only 29% of grade predictions fall below realized outcomes. Second, they markedly overestimate the role of randomness in determining test scores, perceiving the variance of test score noise to be at least three times greater than its actual value. Third, beliefs about the randomness of test outcomes remain strikingly stable over time, despite repeated exposure to informative feedback. In fact, test scores are highly predictive of future performance, with an average pairwise correlation of approximately 0.75 across all five assessments. Fourth, students who place greater weight on luck tend to exhibit larger prediction errors, suggesting that misperceptions about testing noise are associated with lower forecast accuracy.

To identify the causal effects of misspecification on belief updating, we conducted a randomized controlled trial embedded in the final survey administered before the final exam. The treatment provided impersonal statistical information, drawn from administrative data from the previous academic year, demonstrating that term test grades are strong predictors of final exam performance, and that luck plays only a limited role in grade variation. Crucially, the information was framed in general terms, without disclosing students' own scores or peer outcomes, ensuring that any observed changes in beliefs stemmed from shifts in students' understanding of the underlying signal structure rather than from personalized feedback.

To address concerns about potential spillovers from students sharing treatment content with peers, we used a novel design combining staggered treatment implementation with repeated belief elicitation. Students were first randomly assigned to treatment or control groups, and then classified as Early (if accessing the survey within three hours of release) or Late (if accessing it later). Within the Early group, only control students could complete the survey immediately, while treated students were instructed to return after the three-hour window. This design ensured that Early control students could not have been exposed to treated peers prior to completing the survey, allowing for a clean comparison between Early control and Early treated respondents. Balance checks confirm that randomization produced comparable groups across demographic and academic characteristics.

Furthermore, our design elicits beliefs from treated students both before and after the intervention within a single survey, enabling two complementary identification strategies and a direct test for spillovers. A between-group comparison estimates average treatment effects by contrasting treated and control students, while a within-student strategy compares beliefs and predictions immediately before and after treatment. To assess peer spillovers, we compare

pre-treatment beliefs about testing noise between Early treated and Early control students, building on the fact that only Early treated students could have been exposed to spillovers. Finding no significant differences, we conclude that spillovers were minimal or inexistent, supporting the validity of our design.

The information treatment significantly reduced students' perceived role of luck in academic outcomes. On average, the intervention reduced luck-related beliefs by roughly 20 to 25% relative to the control baseline. All effects are statistically significant at the 1% level. Crucially, these belief updates also translated into significant improvements in forecast accuracy. Depending on the identification strategy, absolute prediction errors declined by 0.92 to 2.07 percentage points, corresponding to a 6 to 12% reduction relative to the control baseline.

We leverage the within-student design to classify students based on their belief updating behavior. We define *Aligned Updaters* as those who reduced their perceived testing noise after the intervention, *Null Updaters* as those who reported no change, and *Misaligned Updaters* as those who increased it. Estimating treatment effects by response type, we find that Aligned Updaters exhibit large and significant reductions in absolute prediction errors, while Null and Misaligned Updaters show no statistically significant improvements. By isolating students who decreased their perception of testing noise, these estimates are likely less affected by noise or inattention, and highlight that the treatment's impact is due to a reduction in misspecification.

To interpret the magnitude of treatment effects on prediction accuracy, we compare students' absolute prediction errors to two benchmarks. The first is a Bayesian benchmark, constructed individually for each student using their prior beliefs and a nonparametrically estimated likelihood of observing signals. The second is a conservative benchmark intended to overstate the accuracy attainable by a rational agent. It leverages the full panel dataset and incorporates extensive information, including individual and test fixed effects, lagged test scores, and prior beliefs, yielding highly accurate forecasts with an in-sample R^2 of 0.85.

We estimate the proportion of the gap in prediction accuracy between the control group and each benchmark that is eliminated by the treatment. Relative to the Bayesian benchmark, the treatment closes 39% of the gap in between-group comparisons, 23% in within-student comparisons, and 30% among students classified as *Aligned Updaters*. Even when evaluated against the conservative benchmark the treatment closes 11 to 20% of the gap. These results demonstrate that misspecified beliefs are quantitatively meaningful contributors to prediction errors. By shifting students' beliefs about the role of randomness, the information intervention improved forecast accuracy, eliminating up to one-third of the gap to a rational benchmark.

This paper contributes to several strands of literature. First, it contributes to a growing

body of work documenting systematic deviations from Bayesian updating (Benjamin, 2019; Eil and Rao, 2011; Ertac, 2011; Buser et al., 2018; Coutts, 2019; Drobner, 2022; Möbius et al., 2022; Grether, 1980; Amelio, 2022; Guan, 2023; Gonçalves et al., 2024) by showing that misspecification can be a first-order barrier to accurate updating. Theoretical interest in misspecification dates back to Berk (1966), and recent work has explored the implications of misspecification to settings involving social learning (Heidhues et al., 2018; Frick et al., 2020, 2023) and individual belief updating failures (Fudenberg et al., 2021; Bohren and Hauser, 2025). Despite this theoretical foundation, empirical evidence on the real-world relevance of misspecification remains limited. Related experimental evidence is provided in Chiara and Florian H. (2025). This paper provides the first field-experimental evidence on the causal impact of misspecified beliefs about the noisiness of performance signals, and shows that an information intervention correcting misperceptions leads to significant improvements in forecast accuracy, highlighting the practical relevance of misspecification as a barrier to effective belief updating.

Second, this paper contributes to the literature on belief formation in academic settings. Building on Stinebrickner and Stinebrickner's pioneering work linking administrative records to longitudinal measures of students' expectations (Stinebrickner and Stinebrickner, 2003, 2004, 2006), research has shown that beliefs about ability and expected performance strongly influence educational choices, affecting outcomes such as college dropout (Stinebrickner and Stinebrickner, 2012, 2014a) and major selection (Zafar, 2011; Arcidiacono et al., 2012; Stinebrickner and Stinebrickner, 2014b; Wiswall and Zafar, 2015). Prior work finds that students are often overconfident about their academic ability, and Stinebrickner and Stinebrickner (2012) documents an association between perceptions of luck and how students revise expectations. This paper provides causal evidence from a field experiment that correcting misperceptions about the randomness of test scores substantially improves expectations. The results indicate that misspecified beliefs are a distinct, policy-relevant channel influencing expectations, with potential for scalable interventions to improve belief accuracy.

Third, this paper contributes to the literature combining field experiments with belief elicitation (Haaland et al., 2023; Stantcheva, 2023; D'Acunto and Weber, 2024) by introducing a design that addresses spillover risks when treatment content is easily shared. Building on Wiswall and Zafar (2015), it integrates a within-subject panel with a staggered rollout of a randomized intervention, ensuring a subset of the control group remains unexposed to treated peers and enabling a direct test of spillovers. The within-subject structure also allows to show that gains in prediction accuracy are concentrated among students who lower their perceived test score noise, providing clear evidence of the intervention's mechanism.

These results have practical importance, revealing that in complex environments, misspecification can be a first-order barrier to learning. Correcting such misperceptions does not

require personalized coaching or targeted feedback. Providing impersonal information about the structure of the environment can meaningfully shift beliefs and improve expectations when they are misaligned with the true signal-generating process, underscoring the potential for scalable interventions in settings where misspecification is likely to be widespread.

The remainder of the paper is structured as follows. Section 2 presents the conceptual framework. Section 3 describes the empirical setting and data. Section 4 documents belief formation and updating patterns in the panel. Section 5 presents the experimental design and treatment effects. Section 6 compares students' perceived testing noise with the estimated actual noise. Section 7 benchmarks the magnitude of treatment effects. Section 8 concludes.

2 Conceptual Framework

This section develops a conceptual framework to illustrate how belief updating can fail when students hold incorrect views about the grade-generating process. We consider a stylized setting in which students use past grades to form expectations about future performance but may misperceive the degree of randomness in test outcomes. We distinguish between two sources of deviation from the benchmark of correctly specified Bayesian updating: (1) biased updating, arising from the use of a non-Bayesian updating rule, and (2) misspecification, in which students apply Bayesian updating to a misperceived data-generating process.

2.1 A Stylized Model of Grade Formation and Belief Updating

We begin with a simple example to illustrate how misspecified beliefs about the degree of randomness in test outcomes can lead students to underweight new information and make larger prediction errors about their future performance. Let g_t denote a student's grade on a test taken in period $t \in \{1, 2, 3, 4, 5\}$. Assume a student's grades are determined by an underlying ability θ and a noise component ϵ_t:

$$g_t = \theta + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2).$$

A student does not observe θ, but begins with a prior belief $\theta \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2)$. Upon observing their grade on a first test g_1, a Bayesian student forms a posterior belief about θ and predicts g_2 as:

$$g_2^B = \frac{\frac{\mu_\theta}{\sigma_\theta^2} + \frac{g_1}{\sigma_\epsilon^2}}{\frac{1}{\sigma_\theta^2} + \frac{1}{\sigma_\epsilon^2}}.$$

However, students may hold misspecified beliefs about the variance of the noise component. Suppose a student believes that $\epsilon_t \sim \mathcal{N}(0, \tilde{\sigma}_\epsilon^2)$ with $\tilde{\sigma}_\epsilon^2 > \sigma_\epsilon^2$. Their prediction becomes:

$$g_2^M = \frac{\frac{\mu_\theta}{\sigma_\theta^2} + \frac{g_1}{\tilde{\sigma}_\epsilon^2}}{\frac{1}{\sigma_\theta^2} + \frac{1}{\tilde{\sigma}_\epsilon^2}}.$$

In this case, the student places too little weight on g_1 and relies too heavily on their prior belief μ_θ, resulting in an under-reaction to new information relative to the Bayesian benchmark and, on average, to a lower accuracy when predicting future performance. For overconfident students, this under-reaction implies that mistaken beliefs about ability can persist even in the presence of repeated performance signals.

In addition to this form of misspecification, students may also update in a non-Bayesian way due to cognitive biases or inattention. Therefore, their actual grade prediction, $\hat{g}_2$, may differ from both g_2^B and g_2^M.

2.2 Eliciting Beliefs About Testing Noise

We measure students' beliefs about the noisiness of test scores, $\tilde{\sigma}_\epsilon^2$, using two complementary elicitation strategies. Because $\tilde{\sigma}_\epsilon$ is not a quantity that can be easily reported, we rely on interpretable parameters that map onto this latent parameter under standard assumptions.

First, we elicit the *perceived effect of good luck* on grades. Students are asked to quantify how much a fortunate outcome could improve their grade. Formally, this corresponds to $\mathbb{E}[\epsilon \mid \epsilon \geq 0]$. Under the assumption of normality, there is a one-to-one correspondence between $\mathbb{E}[\epsilon \mid \epsilon \geq 0]$ and $\tilde{\sigma}_\epsilon$, allowing us to directly infer the implied level of perceived noise:

$$\tilde{\sigma}_\epsilon = \sqrt{\frac{\pi}{2}} \cdot \mathbb{E}[\epsilon \mid \epsilon \geq 0]$$

Second, we ask students to report the proportion of their prediction error they believe is due to random luck, as opposed to uncertainty about their own ability. Let r_t denote this reported proportion, which we refer to as the *perceived role of luck* in generating prediction errors. Given an estimate of the variance of the student's subjective distribution over their expected grade, we can isolate the student's perceived testing noise without relying on distributional assumptions:[2]

[2] As shown in Section 6, both methods yield closely aligned estimates of perceived testing noise, with average values of $\tilde{\sigma}_\epsilon = 11.32$ and $\tilde{\sigma}_\epsilon = 12.55$, respectively.

$$\tilde{\sigma}_\epsilon^2 = r_t \cdot \mathbb{E}[(g_t - \hat{g}_t)^2]$$

2.3 Quantifying the Role of Misspecification

To assess the contribution of misspecification to prediction errors, consider the average absolute prediction error that would arise under three scenarios: if students formed expectations as Bayesian agents, as misspecified Bayesian agents, or according to their actual observed forecasts. Let:

$$\begin{aligned}
\Gamma_t^B &= \frac{1}{N}\sum_i |g_{it} - g_{it}^B| && \text{(Bayesian error)},\\
\Gamma_t^M &= \frac{1}{N}\sum_i |g_{it} - g_{it}^M| && \text{(misspecified error)},\\
\Gamma_t &= \frac{1}{N}\sum_i |g_{it} - \hat{g}_{it}| && \text{(observed prediction error)}.
\end{aligned}$$

We define the contribution of misspecification as:

$$\Lambda_t = \frac{\Gamma_t^M - \Gamma_t^B}{\Gamma_t - \Gamma_t^B},$$

which captures the share of excess errors beyond the Bayesian benchmark that can be attributed to misspecified beliefs. The denominator, $\Gamma_t - \Gamma_t^B$, measures the total failure to update optimally, while the numerator, $\Gamma_t^M - \Gamma_t^B$, isolates the portion of this failure that arises specifically from misspecification. It is reasonable to expect that $\Gamma_t^B \leq \Gamma_t^M \leq \Gamma_t$, implying $\Lambda_t \in [0, 1]$.

2.4 Interpreting Experimental Evidence on the Role of Misspecification

To provide causal evidence that misspecification contributes to prediction errors, we conduct a randomized controlled trial. Treated students receive information clarifying that test score noise, σ_ϵ, is low. The intervention serves only to correct beliefs about $\tilde{\sigma}_\epsilon$ implying that any improvement in forecast accuracy can be attributed to reduced misspecification.

To quantify the impact of the treatment, let $\Gamma_5^{\text{control}}$ and Γ_5^{treat} denote the average absolute prediction errors in the control and treatment groups for the final exam taken in period 5. Let Γ_5^B denote the benchmark error under correct Bayesian updating. We define the share of excess prediction error that has been reduced by the treatment as:

$$\Lambda^{\text{ATE}} = \frac{\Gamma_5^{\text{control}} - \Gamma_5^{\text{treat}}}{\Gamma_5^{\text{control}} - \Gamma_5^B}.$$

This measure captures the proportion of prediction error relative to the Bayesian benchmark that is reduced through the information treatment. A value of $\Lambda^{\text{ATE}} = 1$ implies that the treatment fully eliminates the gap between observed prediction errors and the Bayesian benchmark, suggesting that misspecification is the sole source of error. A value of $\Lambda^{\text{ATE}} = 0$ implies that the treatment has no effect, either because students disregard the information or because misspecification does not contribute to prediction errors. We expect $\Lambda^{\text{ATE}} > 0$, reflecting a reduction in absolute prediction errors.

3 Empirical Setting and Data

This section describes the empirical setting and the data sources used in our analysis. We begin by outlining the institutional context of a large first-year calculus course at the University of Toronto, which provides a unique environment for studying belief formation due to its grading policies and challenging assessments. We then detail the administrative records and rich panel survey data that allow us to link students' beliefs to their realized academic outcomes over time. Finally, we summarize key variables and present descriptive statistics to characterize the sample.

3.1 Context

We study a large first-year calculus course at the University of Toronto. This required course for most STEM majors ran from September 2022 to April 2023 and had 1,508 students participate in the first test with 1,155 students participating in the final exam. The final grade is based primarily on four midterms and a final exam, which together account for 70% of the course grade. The remaining 30% come from smaller components, including problem sets and an evaluation of student participation and attendance. Students need to achieve a grade above 50% to pass the class, and many require higher grades to qualify for specific majors.[3]

An important institutional feature of this course is the absence of curving and public grade statistics. Instructors do not release the average, median, or any other distributional information on the grades of students. Furthermore, students are explicitly informed by instructors that all grades received during the course are final and will not be adjusted or curved. As a result, success in the course depends entirely on a student's individual performance on the assessments, and students are likely to form beliefs about their grades with little or no influence from peer outcomes or relative standing.

Our setting is particularly well-suited to studying belief updating and misspecification for

[3]For example, the data science specialist program requires a grade higher than 70% in this class.

several reasons. First, identifying misspecification requires a large sample and multiple test signals in order to estimate the objective noise distribution and compare it to students' subjective beliefs about noise. As the University of Toronto is the largest post-secondary institution in North America, we have an unusually large sample; moreover, whereas most empirical studies rely on a single signal, we observe belief evolution over four distinct signals. Second, the grading policy allows us to rule out most alternative information channels and limits subjective heterogeneity; it also enables our RCT to shift beliefs about test noise independently of beliefs about one's own ability.

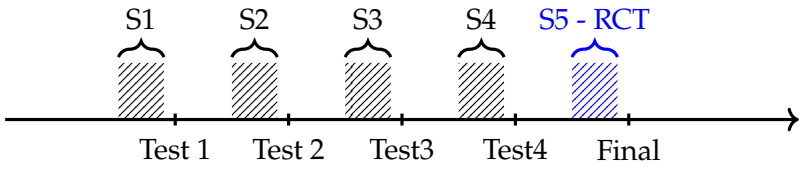
These features are especially useful for detecting failures of belief updating arising from misspecification. However, they limit what we can say about the prevalence of misspecification or the external validity of our results. On the one hand, our context involves relatively informative signals with low ambiguity and a mathematically inclined student population, so misspecification may be less common than in domains such as political beliefs. On the other hand, because our students are highly motivated to hold accurate beliefs and are comfortable with formal reasoning, the effectiveness of our RCT may be stronger than what one would observe in other environments.

3.2 Administrative Data

We obtained access to course-level administrative data on students' grades on all four midterms and the final exam. These grades were recorded independently by the course instructors and are used to calculate each student's final course grade. We link this administrative data to our survey responses at the student level. This linkage allows us to compare students' reported expectations with their realized outcomes across tests. It also enables us to construct measures of prediction error and assess how beliefs respond to test scores.

3.3 Survey Data and Administration

We administered a survey before each test to elicit students' beliefs about their expected performance in the upcoming test and to collect additional variables. In total, five surveys were conducted yielding a rich panel dataset tracking belief formation throughout the year. Each survey took place in a one week window before the corresponding test, after students had received their previous test grade.



To encourage participation, students received a small grade incentive of 0.4% per completed survey, up to a maximum of 2% for completing all five surveys. In addition, we incentivized truthful grade predictions. For each test, the eight most accurate predictions received a $20 reward, and eight additional students were also randomly selected to receive a $20 prize. Students were informed that reporting their best guess would maximize their expected earnings.[4]

The first survey gathered demographic information, including age, gender, international status, year and faculty of study, current or intended program, high school GPA, first-generation college status, and whether the course was required for the student's program. Each of the five surveys included a common set of repeated questions eliciting beliefs about expected grades and perceived testing noise. These repeated measures form a rich longitudinal dataset that tracks how students update their beliefs. To help students revise their beliefs more precisely, we reminded them of their previous responses to recurring survey questions.

The final survey, administered before the final exam, incorporated a randomized controlled trial designed to exogenously shift students' beliefs about testing noise. Details of the RCT are provided in Section 5.

3.4 Key Variables

Grades and Expected Grades. In this paper, we focus on a subset of variables collected on students' beliefs.[5] We observe each student i's grades across the five tests, denoted by g_{it} for $t \in \{1, 2, \ldots, 5\}$. For each test, students reported a prediction of their upcoming grade $\hat{g}_{it}$, and we define the absolute prediction error as $\Gamma_{it} = |g_{it} - \hat{g}_{it}|$. In addition to point predictions, we elicit students' subjective probability distributions over their potential grades. In particular, students provide the probability that their grade g_{it} will be greater than or equal to specific thresholds $X \in \{50, 60, 70, 85\}$.[6] We use these probabilities to construct the cumulative probabilities $G_{it}(X) = \Pr(g_{it} \leq X)$. Because the average class grade is not publicly disclosed, we also elicit students' beliefs about the class average, denoted by $\hat{\bar{g}}_{it}$.

Beliefs About Testing Noise. As discussed in Section 2.2, we also elicit students' beliefs about testing noise by asking them about the role of *luck* in influencing grades and prediction errors. We emphasize to students that luck can exogenously affect grades and provide examples such as getting good or bad sleep, encountering a harsh or lenient grader, or studying the

[4] We use a binarized scoring rule, while it is theoretically incentive-compatible, it can be complicated to explain and bias subject's responses Danz et al. (2022). Therefore, we follow the current best practice and simply inform students that it is in their best interest to report truthfully.

[5] For a complete list of collected variables, see Appendix Section A.6 for the full survey.

[6] We chose 85 instead of 90 because the University of Toronto assigns the maximum GPA value (4.00) to all grades above 85.

right or wrong material. Students are asked how much higher they expect their grade to be if positively affected by luck, which we denote by e_{it} and define as the *perceived effect of good luck*. In addition, we ask them to report the proportion of their prediction errors that they attribute to luck rather than to uncertainty about their own ability, which we denote by r_{it} and define as the *perceived role of luck*.

Additional Variables. Besides the above variables, we also collect information on study effort, beliefs about gender performance, teacher reviews, etc. These variables will be used in a follow-up paper.

3.5 Descriptive Statistics

Demographics. A total of 1,486 students completed at least one survey. As shown in Table 1, among these students, 59% are male, 88% are in their first year of university, 51% are international students, and 92% report that the course is required for their intended program of study. The average high school GPA is 93%, indicating that students have a strong record of academic achievement in high school. Students have an average grade goal of 81%, which corresponds to an A- grade.

Survey Participation. The participation rate in the surveys was high. Table 2 reports the participation rates for each survey among students who completed the corresponding test. The average participation rate across all five surveys is 85%. Among students who participated in the final exam, the average number of surveys completed is 4.3, and 83% completed at least four surveys. The panel dataset comprises 5,501 observations in which students completed the survey corresponding to a test they also took.

Student Achievement. This class is known to be challenging. Among students who participated in at least one test, only 64% obtained a passing final grade,[7] while 19% chose to drop the course, and 17% received a failing grade. The difficulty of the class and the substantial variation in outcomes make this class an ideal setting for studying how students update their beliefs about their academic ability.

Test Grades. As reported in Table 3, average test scores remain fairly stable over time, except for the first test, which recorded a higher mean score than subsequent assessments.[8] Correlations between tests are high, indicating that earlier performance provides significant information about later outcomes. Specifically, the correlations between consecutive tests, $corr(g_{it-1}, g_{it})$, are $(0.78, 0.79, 0.73, 0.81)$. The average pairwise correlation among all five tests is approximately 0.75, reflecting a strong consistency in student performance over time.

[7] This final grade includes assignments, attendance, surveys, and other components in addition to test grades.

[8] We do not disclose the average grade to comply with the course policy prohibiting the release of class averages.

Appendix Table A.1 provides detailed pairwise correlations between all tests.

4 Reduced-Form Analysis

This section leverages our rich panel data to examine how students adjust their beliefs about academic performance and the role of luck in test outcomes. We first describe students' grade expectations and tendencies toward overconfidence, and explore how beliefs about testing noise relate to prediction errors. We then analyze how students update their expectations in response to new performance signals and whether their perceptions of testing noise shift after experiencing unexpected test results.

4.1 Grade Predictions and Beliefs About Testing Noise

Student Grade Predictions. Students' expectations about their grades reveal signs of overconfidence. As shown in Table 4, students tend to overestimate their grades. Across the five tests, 69% of predictions exceeded the actual grade, while only 29% were lower. On average, absolute prediction errors are 9.4 percentage points larger when students overestimate their grades than when they underestimate them. This tendency persists and remains pronounced throughout the course, despite students receiving informative signals about their performance. However, as shown in Table 3, the average expected grade exhibits a slight downward trend over time, suggesting that students gradually adjust their expectations in response to accumulating evidence.

Student Beliefs About Testing Noise. On average, students believe that luck accounts for 37% of their prediction errors, and that the effect of good luck raises their grade by 9.03 percentage points on a test. Table 4 shows that these beliefs remain relatively stable over time. This stability suggests that students are not inferring from the strong correlations in their grades that testing noise is likely to be low. Moreover, the persistent role that students attribute to luck in explaining their prediction errors indicates that they are not gaining significant confidence in their ability to accurately assess their expected performance.

Beliefs About Testing Noise and Prediction Errors. Students who believe that luck plays a larger role in determining their grades tend to have larger absolute prediction errors. Figure 1 shows the relationship between absolute prediction errors and quintiles of the perceived effect of good luck on grades. On average, students in the top two quintiles have absolute prediction errors that are 22% higher than those in the bottom two quintiles, suggesting a potential link between attributing outcomes to luck and reduced accuracy in grade predictions.

4.2 Belief Updating in Response to Test Scores

For each test, the prediction error, $g_{it} - \hat{g}_{it}$, captures the discrepancy between a student's actual performance and their expected grade. As first noted by Zafar (2011), this prediction error can serve as a proxy for the informational content of the test outcome, and provides a basis for analyzing how students adjust their expectations in response to new performance signals. Our survey design is well-suited for this analysis, as it elicits students' beliefs about an upcoming test in a one week window before the exam, and measures updated beliefs at relatively short intervals given the high frequency of our surveys.

Response of Grade Predictions to Test Score Signals. Figure 2 shows the relationship between the change in prediction, $\hat{g}_{it} - \hat{g}_{it-1}$, and the previous test's prediction error, $g_{it-1} - \hat{g}_{it-1}$, by estimating a binned scatterplot regression (Cattaneo et al., 2024).[9] The figure shows that, on average, students whose grade closely matched their prior prediction tend to make little or no adjustment in their subsequent forecast. By contrast, those who received grades lower than predicted tend to revise their next prediction downward, while students who outperformed their prior prediction adjust upward. Nevertheless, the magnitude of these adjustments appears modest, with average changes of about 5 percentage points for students whose prior error exceeded 20 percentage points.

Students who experience large absolute prediction errors may exhibit different updating patterns than those with smaller errors.[10] Our rich panel data enable us to account for this pattern by including student fixed effects in our analysis. Figure 3 shows how the binned scatterplot regression of the prediction changes on the prior prediction errors differs when we control for student fixed effects. The relationship between the change in prediction and the prior prediction error remains similar, but the magnitude of the adjustments tend to be larger.

Table 5 presents the results of a linear regression of the change in prediction on the prior prediction error, estimated both with and without student fixed effects. When student fixed effects are included, the estimated coefficient on the prior prediction error rises from 0.19 to 0.34, implying that receiving a 1 percentage point larger prior prediction error leads to an average increase of 0.34 percentage points in the subsequent prediction. Both models yield statistically significant results at the 1% level. To assess the appropriateness of modeling the relationship between students' prediction changes and prior prediction errors as linear, we implement a nonparametric specification test (Cattaneo et al., 2024).[11] We find no evidence

[9] The figure is constructed using the `binsreg` package (Cattaneo et al., 2025). We fit a piecewise linear function relating changes in grade predictions to prior prediction errors, and display pointwise confidence intervals and a uniform confidence band. The standard errors are clustered at the student level.

[10] In particular, such students may be less responsive to new information if those with more accurate predictions are also better at updating their beliefs or more diligent in completing the surveys.

[11] This test is implemented using the `binstest` package (Cattaneo et al., 2025), and compares a global linear

against the linear model with a p-value of 0.527, suggesting no significant asymmetry in how students adjust their expectations following positive versus negative news.

Response of Beliefs About Testing Noise to Test Score Signals. In contrast to the adjustments students make to their grade predictions, we find that their beliefs about the noisiness of test scores do not respond significantly to prior prediction errors. Figure 4 illustrates the relationship between changes in the perceived effect of good luck, $e_{it} - e_{it-1}$, and the prior prediction error, $g_{it-1} - \hat{g}_{it-1}$, while Figure 5 shows the relationship between changes in the perceived role of luck, $r_{it} - r_{it-1}$, and the prior prediction error. Both models include student fixed effects with standard errors clustered at the student level. In both cases, we find no statistically significant relationship between prior prediction errors and subsequent changes in beliefs about testing noise.

The results presented in this section reveal a notable divergence in how students process performance feedback. While they revise their grade expectations in response to test outcomes, their beliefs about the noisiness of the testing process remain largely unchanged. This rigidity may limit students' ability to learn effectively from feedback and could contribute to persistent biases in their beliefs. To explore whether students' beliefs about testing noise can be influenced through targeted information, and to examine how such shifts might improve the accuracy of their grade predictions, we designed a randomized controlled trial that exogenously alters these beliefs without disclosing any personal performance data. The next section details this experimental intervention.

5 Randomized Controlled Trial

We conducted a randomized controlled trial (RCT) during the final survey administered before the final exam. The RCT was designed to exogenously influence students' beliefs about testing noise and aimed to assess the impact of this intervention on both their beliefs and the accuracy of their grade predictions. If model misspecification is a primary factor contributing to students' failures in belief updating, then providing information that clarifies the nature of testing noise should help students form more accurate grade expectations.

5.1 Treatment Description

The treatment was designed to emphasize that past grades are reliable predictors of performance on the final exam, without revealing any personal information about students' own

polynomial fit to the binscatter estimates. The model includes student fixed effects and clusters standard errors at the student level.

results. This approach enables us to isolate the effect of the treatment on students' beliefs about testing noise, free from the confounding influence of personal feedback. To ensure this separation, the treatment relied on statistics from the same course in the previous academic year, guaranteeing that students received no information specific to their own grades.

Box 1 displays the information provided to students in the treatment group. Treated students were instructed to review this information carefully and were required to remain on the page for two minutes before proceeding with the rest of the survey. The treatment informed students that, based on data from the previous year, performance on the term tests was a strong predictor of outcomes on the final exam. Specifically, we explained that students who scored below the average on the term tests were unlikely to score above the average on the final exam, while those who scored above the average on the term tests were unlikely to fall below the average on the final exam. These statistics were accompanied by aggregate performance data to highlight the large differences in likelihoods across groups.

Box 1: Treatment

If your previous grades were determined largely by luck, then they may not be very helpful in predicting your future grades. However, if luck played only a small role, then your previous grades may be very helpful.

Using **statistics from last year**, we can see that **luck plays only a small role for most students.**

- Amongst students scoring **between 10 to 15 points <u>below</u> the average** across term tests – only **9%** scored **higher than 5 points <u>above</u> the average** on the final.
    - In comparison, ∼**45%** of all students scored **better than 5 points above the average** on the final exam.
- Amongst students scoring **between 10 to 15 points <u>above</u> the average** across term tests – only **9%** scored **worse than 5 points <u>below</u> the average** on the final.
    - In comparison, ∼**40%** of all students scored **worse than 5 points below the average** on the final exam.

Since term test grades predict the final exam grades fairly well, for most students, luck did not seem to play a big role in their grades.

Notes: *This box presents the information shown to students in the treatment group of the RCT. The treatment was designed to convey that past grades are reliable predictors of performance on the final exam and that, on average, luck plays a small role in determining students' grades.*

Information is given as test scores relative to the class average for two reasons. First, it is important that students cannot infer whether the information is explicitly relevant to themselves. This is in order to prevent the treatment from being informative about one's own ability and **only** informative about the testing noise. For example, stating that "70% of students who scored about 70% on the midterms will score above 70% on the exam" will confound the two channels. The institutional policy of withholding class averages further enhanced the impersonal nature of the treatment information, as students had no knowledge of their precise standing relative to the class average. As a result, the intervention provided no clear individualized guidance for predicting their own future performance but instead conveyed the broader message that past grades are generally informative for forecasting future outcomes.

5.2 Treatment Assignment

The impersonal nature of the treatment raises concerns about potential spillover effects. Specifically, if students share the treatment information with their peers, members of the control group might be inadvertently exposed, leading to an underestimation of the true treatment effect.

To account for potential spillovers, we leveraged variation in the timing of survey take-up to construct a control group that would remain unaffected by potential information sharing. A straightforward approach would have been to assign students who completed the survey earlier to the control group and those who completed it later to the treatment group. However, such an approach risks introducing confounding factors, as early survey respondents may systematically differ from those responding later. To address this concern, we implemented a staggered design that establishes an early control group protected from spillover effects while also providing a treated group suitable for valid comparison with this early control group.

We began by randomly assigning half of the potential participants to the treatment group and the other half to the control group. Next, we divided students into two groups based on the timing of their initial attempt to access the survey: those who did so within the first three hours of its release were classified as the *Early* group, while those attempting afterward formed the *Late* group. Within the Early group, students assigned to the control group were permitted to complete the survey immediately, whereas those assigned to the treatment group were instructed to return later to complete it. This design ensures that the early control group remains unaffected by potential spillover effects and enables a valid comparison with the early treatment group to account for selection into early survey participation.

5.3 Measurement of Beliefs Before and After Treatment

To better isolate the causal effects of the information intervention and to test for spillover effects, we measure each treated student's beliefs both before and after exposure to the treatment. Specifically, prior to receiving the treatment, students in the treatment group report their pre-treatment beliefs about their own grade ($\hat{g}_{i5}^{\text{pre}}$), the expected class average ($\hat{\bar{g}}_{i5}^{\text{pre}}$), their distribution of potential grades (G_{i5}^{pre}), the perceived contribution of luck to prediction errors (r_{i5}^{pre}), and the perceived effect of good luck on grades (e_{i5}^{pre}). Following this initial belief elicitation, students are shown the treatment information, and are then asked to report their beliefs again, yielding a set of treated belief measures: $\hat{g}_{i5}^{\text{post}}$, $\hat{\bar{g}}_{i5}^{\text{post}}$, G_{i5}^{post}, r_{i5}^{post}, and e_{i5}^{post}.

This panel structure allows for two complementary empirical strategies to estimate treatment effects. A first approach is to use *between-group* comparisons, which leverages the randomized controlled trial design to compare post-treatment beliefs of treated students with the beliefs of control students. For beliefs measured both before and after the treatment, the repeated measurements also enable a *within-student* estimation strategy that compares each treated student's beliefs before and after receiving the treatment. This design provides unusually rich information, allowing us to observe individual-level responses to the treatment.

Unlike the between-group approach, the identification strategy using within-student comparisons does not rely on random assignment, but instead requires that the treatment does not affect students' pre-treatment beliefs and that there are no time-varying confounders in the short time window where students report their pre- and post-treatment beliefs. These assumptions are plausible given that beliefs were elicited immediately before and after treatment exposure, and that the survey content was identical to the control version up to the point of treatment. We discuss the assumptions underlying this identification strategy more formally in Section A.1.1.

Throughout the analysis, we report results from both identification strategies to provide a comprehensive assessment of the impact of the intervention on students' beliefs.

5.4 Descriptive Statistics of Treatment and Control Groups

In this subsection, we present descriptive statistics for the treatment and control groups to assess the balance achieved through randomization. Table 6 reports the average characteristics of students in each group, along with the differences between them. The table indicates that the groups are well balanced with respect to demographics and prior academic performance, showing no statistically significant differences across any of the variables.

Table 7 presents the average characteristics of students in the Early and Late groups by

treatment status. The results show that randomization successfully balanced the treatment and control groups within each timing group, with no significant differences in demographics or prior academic performance, except for the share of male students and first-year students in the Early group, where statistically significant differences are observed. Consistent with the motivation for our staggered design, the table also reveals systematic differences between the Early and Late groups, with students in the Late group scoring lower on average in prior tests. Overall, the balance checks support the validity of our design and randomization.

5.5 Testing for Spillover Effects

To assess potential spillover effects, we compare pre-treatment beliefs about testing noise between students in the Early treatment and Early control groups. If spillovers were present, the Early treatment students who respond to the survey later may have been exposed to the treatment already. Hence, we would expect their pre-treatment beliefs to be different from those of the Early control. In this case, we would expect lower pre-treatment beliefs in the treatment group, as some treated students might have received information from peers suggesting that luck plays a minimal role in determining grades before completing the survey.

Table 8 shows the average pre-treatment belief about the effect of good luck is 9.55 percentage points in the treatment group and 9.14 in the control group. Similarly, the average belief about the contribution of luck to prediction errors is 38.53% in the treatment group versus 37.03% in the control group. Both differences are not statistically significant. The absence of meaningful differences in these measures suggests that spillover effects were minimal or absent, supporting the integrity of the research design.

5.6 Treatment Effects on Beliefs About Testing Noise

We first assess whether the information intervention shifted students' beliefs about test score noise. Table 9 reports average treatment effects on the perceived effect of good luck on grades and the perceived role of luck in explaining prediction errors. We estimate average treatment effects using both between-group and within-student identification strategies.

The between-group results indicate that treated students report significantly lower beliefs about the role of luck after the intervention. On average, the perceived effect of good luck on grades is 1.75 percentage points lower in the treatment group compared to the control group, while the perceived contribution of luck to prediction errors is 8.60 percentage points lower. The within-student analysis, which compares pre- and post-treatment beliefs for each treated student, yields similar effects. On average, the perceived effect of good luck declines by 1.83 percentage points, while the perceived contribution of luck to prediction errors drops by 9.57 percentage points. All effects are statistically significant at the 1% level.

These results show that the treatment had a substantial and statistically significant impact with reductions in luck-related beliefs by roughly 20 to 25% relative to the control baseline and both identification strategies yielding consistent evidence.

5.7 Treatment Effects on Students' Grade Predictions

We next assess whether the treatment improved the accuracy of students' grade predictions. Table 10 presents the average treatment effects on students' absolute prediction errors. The between-group estimate shows that, on average, treated students' absolute prediction errors are 2.07 percentage points lower than those of students in the control group. This effect is statistically significant at the 5% level and represents a reduction of approximately 12% relative to the control group's average baseline error. The within-student estimate shows that, on average, treated students' absolute prediction errors are 0.92 percentage points lower than their pre-treatment absolute prediction errors. This effect is statistically significant at the 1% level and corresponds to a 5.7% improvement relative to their average pre-treatment baseline.

These results indicate that the information treatment not only shifted students' beliefs about testing noise, but also improved the accuracy of their grade predictions.

5.8 Heterogeneity by Change in Beliefs About Testing Noise

The within-student identification strategy enables an analysis of how changes in beliefs about testing noise mediate the treatment's effect on grade prediction accuracy. This feature is uncommon in standard experimental settings, where outcomes are often observed only post-treatment and comparisons are limited to between-group averages. By contrast, our pre- and post-treatment measurements make it possible to estimate heterogeneous treatment effects as a function of observed belief updating. This added granularity offers unique insights into the mechanisms through which the treatment operates.

We define *response types* based on the sign of observed belief updating following the information treatment. Let $\Delta e_i = e_{i,5}^{\text{post}} - e_{i,5}^{\text{pre}}$ and $\Delta r_i = r_{i,5}^{\text{post}} - r_{i,5}^{\text{pre}}$ denote the within-student changes in the perceived effect of good luck and the perceived role of luck, respectively. We classify students into one of three response types for each of these changes:

- **Aligned Updaters** ($\Delta e_i < 0$), ($\Delta r_i < 0$): Students who reduced their beliefs about testing noise in response to the information treatment. These students updated in the hypothesized direction, consistent with learning that test scores are more informative than previously believed.
- **Null Updaters** ($\Delta e_i = 0$), ($\Delta r_i = 0$): Students whose beliefs about testing noise remained unchanged following the treatment. This may reflect already accurate prior beliefs or

limited engagement with the informational content of the intervention.

- **Misaligned Updaters** ($\Delta e_i > 0$), ($\Delta r_i > 0$): Students who increased their beliefs about testing noise after receiving the information. This response may indicate that these students had previously underestimated the role of noise, or it could reflect misinterpretation of the information or inattention during the survey.

We estimate both the distribution of response types and the average treatment effects of each type. However, identifying the distribution of response types requires stronger assumptions than those needed to estimate average treatment effects. Specifically, we assume idiosyncratic shocks are unlikely to reverse the sign of the change in beliefs. This assumption is plausible given that students are likely to form a clear sense of whether the treatment reinforces or challenges their prior beliefs, even if the magnitude of the adjustment may vary due to attention or the survey context.[12]

Table 11 reports the joint distribution of response types. A majority of students (64.31%) are classified as *Aligned Updaters* on at least one dimension, with 35.69% aligning on both. At the same time, a sizable share of students (53.53%) chose to leave at least one of their reported beliefs unchanged, and 23.14% are classified as *Null Updaters* on both measures. In contrast, only 17.65% of students increased their beliefs about testing noise on at least one measure, and just 5.69% did so on both. These results indicate that most students either reduced their beliefs about testing noise or maintained their prior views.

Table 12 reports average treatment effects (ATEs) by students' response types. For each group, we present both the pre-treatment average and the ATE for three outcomes: absolute prediction error, the perceived effect of good luck on grades, and the perceived contribution of luck to prediction errors.

Panel A classifies students based on their updating in the perceived effect of good luck, while Panel B uses changes in the perceived contribution of luck to prediction errors. Across both classifications, students identified as *Aligned Updaters* exhibit sizable and statistically significant improvements in prediction accuracy. These students also begin with higher baseline beliefs about the effect and role of luck than their peers, suggesting that they were more prone to overestimating the noisiness of test scores prior to the intervention.

In contrast, *Misaligned Updaters* show no statistically significant improvements in prediction accuracy. On average, these students reported lower pre-treatment beliefs about the noisiness of test scores in the dimension used to define their response type. This pattern suggests that

[12]Nevertheless, inattention or distraction may occasionally alter the sign of reported updating. In such cases, the observed distribution of response types reflects not only belief updating but also variation in response quality as described in the definition of the response types.

many of them may have initially underestimated the role of luck. However, this group also exhibits the highest pre-treatment prediction errors, raising the possibility that some students in this category were less engaged with the survey overall.

Null Updaters show some heterogeneity across Panels A and B: when classified by the perceived effect of good luck, they exhibit significant gains in prediction accuracy, but no significant change when classified by the perceived role of luck. However, a large share of these students are aligned on at least one belief dimension. To account for this overlap, we next consider categories that combine both dimensions of belief updating.

Panel C aggregates students into three mutually exclusive groups: those who aligned on at least one belief dimension, those who remained null on both dimensions, and those who misaligned on at least one dimension and were never aligned. The results reinforce the finding that improvements in prediction accuracy are concentrated among students who updated in the intended direction. Students who were aligned on either measure exhibit a statistically significant reduction of 1.37 percentage points in their absolute prediction errors. By contrast, students who were misaligned and never aligned or who remained null on both dimensions show no statistically significant improvements in prediction accuracy.[13]

The randomized controlled trial shows that misspecification is a major obstacle to accurate updating from performance signals. We now turn to imposing additional structure to compare students' beliefs about testing noise to actual estimates, and to interpret the magnitude of treatment effects as a share of the observed deviation from a Bayesian benchmark.

6 Perceived vs. Actual Testing Noise

In this section, we estimate the actual test score noise and compare it to students' beliefs about testing noise. We provide evidence that students' beliefs about testing noise are indeed misaligned with the actual noisiness of the grade-generating process.

Unlike laboratory experiments, we do not control the signal-generating process. Grades are produced in a high-stakes environment where manipulation is neither feasible nor ethical. As a result, any model of grade production is necessarily misspecified to some extent. Since our estimates rely on a parsimonious model, students may possess private information, such as study time or health conditions, not captured by our specification. Consequently, our

[13]Notably, students classified as null updaters on both dimensions exhibit the lowest average pre-treatment prediction errors across all groups. They also report the lowest pre-treatment beliefs about the effect and role of luck. This pattern suggests that many of these students may have already held relatively accurate beliefs prior to the intervention.

estimates of testing noise likely overstate the true level of randomness in grades so that the estimated gap between students' perceived and actual noise provides a conservative measure of misalignment: the true discrepancy is likely to be even larger.

We proceed by assuming the following data-generating process for student grades:

$$g_{it} = \theta_i + \eta_{it} + \bar{g}_t + \epsilon_{it}, \quad \epsilon_{it} \sim \mathcal{N}(0, \sigma_\epsilon^2),$$

where θ_i is a student fixed effect representing the time-invariant ability in this course, η_{it} is a transitory skill shock, $\bar{g}_t$ is a test fixed effect capturing the test difficulty, and ϵ_{it} represents exogenous testing noise. Our goal is to estimate σ_ϵ^2, the variance of this noise term.

Note that the residual from a model including student fixed effects and period fixed effects is composed of both the transitory skill component η_{it} and the idiosyncratic noise ϵ_{it}, so the residual variance yields only an upper bound on σ_ϵ^2. The informativeness of test scores cannot be directly recovered without additional structure. To solve this problem, we impose a simple autoregressive model on the transitory component. Specifically, we assume that η_{it} follows a stationary AR(2) process:

$$\eta_{it} = \beta_1 \eta_{it-1} + \beta_2 \eta_{it-2} + \nu_{it}.$$

This assumption allows us to separately identify the variance components associated with η_{it} and ϵ_{it}. The AR(2) assumption captures the idea that transitory factors affecting student performance are likely to persist across a few exams, allowing us to separate these structured fluctuations from truly random testing noise. It provides a flexible representation of short-term persistence in performance, while remaining feasible to estimate given our short panel structure with 5 time periods.

Appendix Section A.2 details the identification and estimation of the variance of testing noise. We estimate $\sigma_\epsilon = 3.75$. As shown in Section 2.2, we can use the elicited beliefs about the perceived effect of good luck and the perceived role of luck to compare students' implied perceived testing noise to this estimate. First, on average, students believe that the effect of good luck on their grades is approximately 9.03 percentage points. This implies a perceived testing noise of $\tilde{\sigma}_\epsilon = 11.32$. Second, on average, students believe that the contribution of luck to prediction errors is approximately 37%. Using students' elicited subjective probability distributions to estimate the variance of their expected grades, we find a perceived testing noise of $\tilde{\sigma}_\epsilon = 12.55$.[14]

These estimates indicate that students perceive test score noise to be roughly three times as large as its actual variance. Even under our conservative estimate that likely overstates

[14] Appendix Section A.3 details our approach to estimating the subjective variance of expected grades.

true randomness, students substantially overestimate the role of luck, highlighting the value of interventions that correct beliefs about performance signals in this setting.

7 Interpreting the Magnitude of Treatment Effects

The randomized controlled trial provides causal evidence that correcting students' beliefs about test score noise improves the accuracy of their grade predictions. However, since some prediction error is unavoidable due to inherent noise, it is not reasonable to expect the treatment to eliminate all errors. Consequently, we require a benchmark for what constitutes an accurate forecast in order to interpret the magnitude of treatment effects. In this section, we construct two benchmarks and estimate the share of prediction error eliminated by the intervention relative to these benchmarks.

7.1 Constructing the Bayesian Benchmark

We first construct a benchmark representing optimal Bayesian updating. This benchmark reflects the forecast that would be made by a Bayesian agent who shares the student's prior beliefs and updates rationally based on a relevant signal.

We follow the procedure described in Appendix Section A.4. For each student, we construct a posterior predictive distribution over final exam grade bins using (i) prior beliefs elicited in the fourth survey and (ii) the fourth term test grade as the signal. The likelihood function is estimated nonparametrically from the joint distribution of signals and final exam outcomes. We compute the posterior expected grade as:

$$\hat{G}_i^{\text{Bayes}} = \sum_{k=1}^{5} \tilde{\pi}_{ik} \cdot \mu_k,$$

where $\tilde{\pi}_{ik}$ is student i's posterior probability of grade bin k and μ_k is its midpoint.

7.2 Constructing a Conservative Benchmark

As discussed in Section 6, studying belief formation in natural settings requires acknowledging that models of endogenous signals are inevitably misspecified. To ensure a conservative comparison, we construct a flexible benchmark that overstates the accuracy achievable under Bayesian updating, and thus understates the share of prediction error eliminated by the treatment.

This benchmark is constructed from a linear regression model that predicts test scores

using our rich panel data. Formally, we estimate the following model:

$$g_{it} = \alpha_i + \delta_t + \beta_1 g_{i(t-1)} + \sum_{k=1}^{4} \gamma_k \cdot \pi_{itk} + \varepsilon_{it},$$

where g_{it} is the test score of student i at time t, α_i are student fixed effects, δ_t are test fixed effects, $g_{i(t-1)}$ is the lagged test score, and π_{itk} denotes the prior probability for grade bin k which was elicited at time $t-1$. The model is estimated using all available periods, and we use the estimated coefficients to construct fitted predictions of final exam grades in period $t = 5$, which we denote by $\hat{G}_i^{\text{Flex}}$. The high explanatory power of the flexible benchmark with an $R^2 = 0.85$, confirms that final exam outcomes are largely predictable using lagged performance, prior beliefs, and individual and test fixed effects. We note that the goal of the model is to present an upper bound on the accuracy achievable under Bayesian updating; hence, the potential issue of overfitting is not problematic. If the actual accuracy achievable is lower, then misspecification will explain a larger share of mistakes; hence, overfitting only downwardly biases the effect of misspecification.

7.3 Interpreting Treatment Effects

We now use our estimated benchmarks to interpret the effect of the information treatment. Let $\Gamma_5^{\text{control}}$ and Γ_5^{treat} denote the average absolute prediction error on the final exam for the control and treatment groups, respectively. Let Γ_5^B denote the benchmark average absolute prediction error. As described in Section 2.4, we define the proportion of excess prediction error eliminated by the treatment as:

$$\Lambda^{\text{ATE}} = \frac{\Gamma_5^{\text{control}} - \Gamma_5^{\text{treat}}}{\Gamma_5^{\text{control}} - \Gamma_5^{B}}.$$

We estimate Λ^{ATE} using both the flexible and Bayesian benchmarks, reporting results based on between-group and within-student identification strategies. We also present within-student estimates for the subgroup of students whose change in beliefs were aligned with the treatment along at least one dimension as defined in Section 5.8.

Table 13 reports the estimates. Panel A shows that, in between-group comparisons, the treatment closes 39% of the gap relative to the Bayesian benchmark and 20% relative to the flexible benchmark. Both effects are statistically significant at the 5% level.

Panels B and C present within-student estimates. Among all students (Panel B), the treatment closes 23% of the gap relative to the Bayesian benchmark and 11% relative to the flexible benchmark. Among students whose belief changes aligned with the treatment (Panel

C), the treatment closes 30% and 16% of the gap, respectively. All estimates are statistically significant at the 1% level.

Together, these results indicate that the treatment led to meaningful improvements in forecasting accuracy. The estimated effects remain substantial even when compared to the flexible benchmark, which overstates the accuracy achievable by a rational agent. This conservative benchmark yields a lower-bound estimate of 11%, while the Bayesian benchmark suggests that the treatment closes 23% of the gap overall and 30% among students whose belief changes aligned with the treatment. These effects can be attributed entirely to our information treatment, underscoring the central role of misspecification in limiting the predictive accuracy of students in this setting.

8 Conclusion

Leveraging rich panel data on students' grade expectations and a randomized information intervention, this paper provides the first causal evidence on the effects of correcting misperceptions about the noisiness of performance signals in a natural, high-stakes setting with endogenous signals. We find that students systematically overestimate the randomness in their academic performance, leading to substantial forecasting errors. These errors stem from a misspecified mental model of the grade-generating process where students fail to recognize the high signal value of prior grades and consequently underreact to informative feedback.

The experimental intervention provided impersonal statistical information on the predictive power of prior test scores. It significantly reduced students' perceptions of the role of luck and led to substantial improvements in their ability to predict their academic performance. Among students who reduced their perceived testing noise, the intervention eliminated 30% of the gap in prediction accuracy relative to a Bayesian benchmark. These improvements were concentrated among students whose prior beliefs were most misaligned with the data, underscoring the central role of misspecification in limiting students' ability to form accurate expectations.

Our findings have broader implications for models of learning. They highlight that persistent prediction errors need not only arise from irrational updating, but may instead reflect incorrect beliefs about the structure of the environment. This distinction is consequential. While behavioral biases are often difficult to remediate, structural misperceptions, such as beliefs about the noisiness of signals, can be shifted through improved information. Theoretically, our results affirm the relevance of misspecification as a distinct impediment to learning. Methodologically, our experimental design combines repeated belief elicitation and a stag-

gered implementation of a randomized treatment to test for spillover effects, and allows a rich study of heterogeneous responses. Practically, the findings show that information interventions targeting misperceptions of the signal-generating process can meaningfully improve expectations, with important implications for environments where accurate forecasting and effective responses to performance feedback are essential for decision-making.

References

Amelio, A. (2022). Cognitive Uncertainty and Overconfidence. ECONtribute Discussion Papers Series 173, University of Bonn and University of Cologne, Germany.

Arcidiacono, P., V. J. Hotz, and S. Kang (2012). Modeling college major choices using elicited measures of expectations and counterfactuals. *Journal of Econometrics 166*(1), 3–16.

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

Berk, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics 37*(1), 51–58.

Bohren, A. and D. N. Hauser (2025). The behavioral foundations of model misspecification. Working Paper.

Buser, T., L. Gerhards, and J. Van Der Weele (2018). Responsiveness to feedback as a personal trait. *Journal of Risk and Uncertainty 56*, 165–192.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng (2024, May). On binscatter. *American Economic Review 114*(5), 1488–1514.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng (2025). Binscatter regressions. *The Stata Journal 25*(1), 3–50.

Chiara, A. and S. Florian H. (2025). Weighting competing models. Working Paper.

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

D'Acunto, F. and M. Weber (2024). Why survey-based subjective expectations are meaningful and important. *Annual Review of Economics 16*(Volume 16, 2024), 329–357.

Danz, D., L. Vesterlund, and A. J. Wilson (2022). Belief elicitation and behavioral incentive compatibility. *American Economic Review 112*(9), 2851–2883.

Drobner, C. (2022). Motivated beliefs and anticipation of uncertainty resolution. *American Economic Review: Insights 4*(1), 89–105.

Eil, D. and J. M. Rao (2011). The good news-bad news effect: Asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics 3*(2), 114–138.

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

Frick, M., R. Iijima, and Y. Ishii (2020). Misinterpreting others and the fragility of social learning. *Econometrica 88*(6), 2281–2328.

Frick, M., R. Iijima, and Y. Ishii (2023). Belief convergence under misspecified learning: A martingale approach. *Review of Economic Studies 90*(2), 781–814.

Fudenberg, D., G. Lanzani, and P. Strack (2021). Limit points of endogenous misspecified learning. *Econometrica 89*(3), 1065–1098.

Gonçalves, D., J. Libgober, and J. Willis (2024). Retractions: Updating from complex information. Working Paper.

Grether, D. M. (1980). Bayes rule as a descriptive model: The representativeness heuristic. *Quarterly Journal of Economics 95*(3), 537–557.

Guan, M. (2023). Choosing between information bundles. Working Paper.

Haaland, I., C. Roth, and J. Wohlfart (2023, March). Designing information provision experiments. *Journal of Economic Literature 61*(1), 3–40.

Heidhues, P., B. Kőszegi, and P. Strack (2018). Unrealistic expectations and misguided learning. *Econometrica 86*(4), 1159–1214.

Jensen, R. (2010). The (perceived) returns to education and the demand for schooling. *The Quarterly Journal of Economics 125*(2), 515–548.

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

Oreopoulos, P. and R. Dunn (2013). Information and college access: Evidence from a randomized field experiment. *The Scandinavian Journal of Economics 115*(1), 3–26.

Oreopoulos, P. and U. Petronijevic (2019). The remarkable unresponsiveness of college students to nudging and what we can learn from it. Technical report, National Bureau of Economic Research.

Oreopoulos, P. and U. Petronijevic (2023). The promises and pitfalls of using (mostly) low-touch coaching interventions to improve college student outcomes. *Economic Journal 133*(656), 3034–3070.

Stantcheva, S. (2023). How to run surveys: A guide to creating your own identifying variation and revealing the invisible. *Annual Review of Economics 15*(Volume 15, 2023), 205–234.

Stinebrickner, R. and T. R. Stinebrickner (2003). Understanding educational outcomes of students from low-income families. *Journal of Human Resources 38*(3), 591–617.

Stinebrickner, R. and T. R. Stinebrickner (2004). Time-use and college outcomes. *Journal of Econometrics 121*(1), 243–269. Higher education (Annals issue).

Stinebrickner, R. and T. R. Stinebrickner (2006). What can be learned about peer effects using college roommates? Evidence from new survey data and students from disadvantaged backgrounds. *Journal of Public Economics 90*(8), 1435–1454.

Stinebrickner, R. and T. R. Stinebrickner (2012). Learning about academic ability and the college dropout decision. *Journal of Labor Economics 30*(4), 707–748.

Stinebrickner, R. and T. R. Stinebrickner (2014a). Academic performance and college dropout: Using longitudinal expectations data to estimate a learning model. *Journal of Labor Economics 32*(3), 601–644.

Stinebrickner, R. and T. R. Stinebrickner (2014b). A major in science? Initial beliefs and final outcomes for college major and dropout. *Review of Economic Studies 81*(1), 426–472.

Wiswall, M. and B. Zafar (2015). Determinants of college major choice: Identification using an information experiment. *Review of Economic Studies 82*(2), 791–824.

Zafar, B. (2011). How do college students form expectations? *Journal of Labor Economics 29*(2), 301–348.

9 Tables

Table 1: Demographics

Male	0.59
Age	18.51
First Year	0.88
International	0.51
Course is Required	0.92
Grade Goal	81.01
High School GPA	93.50
Observations	1,488

Notes: This table reports summary statistics for the 1,488 students who completed at least one survey in the study sample. The variable *Male* is an indicator equal to 1 for male students. *Age* is reported in years. *First Year* is an indicator for students in their first year of university. *International* is an indicator for students who are classified as international students. *Course is Required* is an indicator if the student reports that the course is required for their intended program of study. *Grade Goal* refers to the student's self-reported target final grade for the course (on a 0-100 scale). *High School GPA* is the self-reported high school average, also on a 0-100 scale. All values are means across students.

Table 2: Survey Completion by Test

	Time Period				
Variable	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$
Completed Survey	0.85	0.82	0.86	0.85	0.88
Number of Completed Surveys	3.75	3.94	4.15	4.32	4.33
Completed 5 Surveys	0.51	0.54	0.59	0.64	0.65
Completed at Least 4 Surveys	0.67	0.71	0.77	0.83	0.83
Observations	1,508	1,399	1,278	1,129	1,155

Notes: This table reports survey participation rates across the five main tests in the study. Each column corresponds to a specific time period $t \in \{1, \ldots, 5\}$, and the sample includes only students who completed the corresponding test in that period. *Completed Survey* is the share of students who completed the corresponding survey. *Number of Completed Surveys* is the average number of surveys completed by students who took the test at time t. *Completed 5 Surveys* is the share of students who completed all five surveys, and *Completed at Least 4 Surveys* is the share who completed four or more. Observations report the number of students who completed the test in each period. Participation rates are high throughout, with 82% to 88% of test-takers completing each survey on average. The panel dataset includes a total of 5,501 student-test observations where both test and survey responses are observed.

Table 3: Student Grades and Expectations by Test

	Time Period				
Variable	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$
Grade g_{it}					
Mean Relative to Test 1	0.00	−9.23	−8.04	−11.79	−5.25
Std. Dev.	20.54	22.10	23.04	20.56	24.35
N	1,508	1,399	1,278	1,129	1,155
Expected Grade $\hat{g}_{it}$					
Mean Relative to Test 1	0.00	0.61	−1.49	−3.04	−4.84
Std. Dev.	14.36	14.34	14.96	15.38	15.80
N	1,278	1,153	1,095	962	1,013
Absolute Prediction Error Γ_{it}					
Mean	15.62	19.47	18.28	18.32	16.80
Std. Dev.	12.83	14.71	14.99	14.21	13.65
N	1,278	1,153	1,095	962	1,013

Notes: This table reports summary statistics for realized grades g_{it}, expected grades $\hat{g}_{it}$, and absolute prediction errors $\Gamma_{it} = |g_{it} - \hat{g}_{it}|$ across five tests ($t \in \{1,\ldots,5\}$). For each variable, we report the mean and standard deviation, as well as the number of observations with non-missing values. To comply with course policy, grades and expected grades are reported relative to the mean of test 1. The number of observations varies across tests due to course withdrawals and survey non-response.

Table 4: Student Beliefs by Test

	Time Period				
Variable	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$
Prediction > Grade	0.57	0.79	0.72	0.80	0.60
Perceived Effect of Good Luck	9.39	9.04	8.89	8.77	8.96
Perceived Role of Luck	35.64	36.01	38.10	37.34	37.60
Change in Expected Grade	–	0.27	-3.02	-2.19	-1.45
Observations	1,278	1,153	1,095	962	1,013

Notes: This table reports summary statistics on students' belief measures across five tests ($t \in \{1,\ldots,5\}$). *Prediction > Grade* is the share of students whose grade prediction exceeded their realized grade. *Perceived Effect of Good Luck* is the average expected gain in grade points that students believe they would receive if they were lucky on the test. *Perceived Role of Luck* denotes the average student belief about the percentage of their prediction error attributable to random luck. *Change in Expected Grade* is the average change in students' grade expectations relative to the previous test, this variable is undefined for $t = 1$. Observations correspond to the number of students who completed the survey and the test in each period.

Table 5: Response of Grade Predictions to Test Score Signals

	Δ Prediction	Δ Prediction
Past Prediction Error	0.19***	0.34***
	(0.01)	(0.02)
Student Fixed Effects	No	Yes
Observations	3,555	3,350
Number of Students	1,219	1,014

Standard errors in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table reports estimates from linear regressions of the change in students' expected grade ($\hat{g}_{it} - \hat{g}_{it-1}$) on the prior test's prediction error ($g_{it-1} - \hat{g}_{it-1}$). The dependent variable in both columns is the change in point prediction between consecutive tests. Column 1 presents estimates without student fixed effects, while Column 2 includes them to control for time-invariant individual heterogeneity. Standard errors are clustered at the student level. The coefficient increases from 0.19 to 0.34 when fixed effects are included, indicating that students adjust their subsequent prediction upward by 0.34 percentage points for every 1 percentage point underprediction on the previous test. We trim the prior prediction error at the 5th and 95th percentiles to reduce the influence of extreme outliers. We test the linearity of the relationship using a nonparametric specification test (Cattaneo et al., 2024) and find no evidence against linearity ($p = 0.527$) for the model including student fixed effects.

Table 6: Descriptive Statistics of Experimental Groups

	Treatment	Control	Difference
Test 1 grade	0.00	-1.34	1.34
	(0.77)	(0.77)	(1.09)
Test 2 grade	0.00	-0.69	0.69
	(0.92)	(0.90)	(1.28)
Test 3 grade	0.00	-0.73	0.73
	(0.96)	(0.98)	(1.38)
Test 4 grade	0.00	-0.14	0.14
	(0.91)	(0.91)	(1.28)
Male	0.58	0.62	-0.04
	(0.02)	(0.02)	(0.03)
International	0.54	0.52	0.02
	(0.02)	(0.02)	(0.03)
First Year	0.94	0.91	0.02
	(0.01)	(0.01)	(0.02)
Age	18.25	18.33	-0.08
	(0.05)	(0.05)	(0.07)
Goal Mark	81.34	81.49	-0.15
	(0.47)	(0.46)	(0.66)
First Generation	0.19	0.19	0.00
	(0.02)	(0.02)	(0.02)
High School GPA	94.38	93.99	0.39
	(0.18)	(0.19)	(0.26)
Observations	510	538	1048

Standard errors in parentheses.

Stars are only displayed for the Difference column.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table reports mean characteristics for students in the treatment and control groups, along with the difference between the two. To comply with course policy, academic performance measures are expressed relative to the average performance of the treatment group. Demographic variables are indicators for male, international, and first-year student status. *Goal Mark* is the student's self-reported target grade (on a 0-100 scale), and *High School GPA* is the self-reported high school grade point average (on a 0-100 scale). *First Generation* is an indicator equal to 1 if the student is the first in their family to attend college. Standard errors are shown in parentheses. No statistically significant differences are found between groups, indicating successful randomization. Stars indicate significance levels for the difference column only.

Table 7: Descriptive Statistics of Experimental Groups by Early Status

	Early			Late		
	Treatment	Control	Difference	Treatment	Control	Difference
Test 1 grade	0.00	-1.96	1.96	-4.47	-5.55	1.08
	(1.33)	(1.35)	(1.90)	(0.93)	(0.92)	(1.31)
Test 2 grade	0.00	-0.01	0.01	-3.06	-4.01	0.95
	(1.53)	(1.65)	(2.25)	(1.13)	(1.07)	(1.55)
Test 3 grade	0.00	0.69	-0.69	-3.74	-5.00	1.26
	(1.66)	(1.74)	(2.40)	(1.18)	(1.18)	(1.67)
Test 4 grade	0.00	0.15	-0.15	-6.52	-6.77	0.25
	(1.57)	(1.76)	(2.35)	(1.10)	(1.04)	(1.51)
Male	0.60	0.71	-0.10*	0.57	0.59	-0.02
	(0.04)	(0.04)	(0.05)	(0.03)	(0.03)	(0.04)
International	0.53	0.48	0.05	0.54	0.53	0.01
	(0.04)	(0.04)	(0.06)	(0.03)	(0.03)	(0.04)
First Year	0.96	0.89	0.07**	0.93	0.92	0.00
	(0.02)	(0.03)	(0.03)	(0.01)	(0.01)	(0.02)
Age	18.18	18.31	-0.14	18.28	18.33	-0.05
	(0.07)	(0.09)	(0.11)	(0.06)	(0.06)	(0.09)
Goal Mark	82.30	82.69	-0.39	80.95	81.01	-0.06
	(0.87)	(0.83)	(1.20)	(0.56)	(0.55)	(0.79)
First Generation	0.14	0.21	-0.07	0.21	0.18	0.03
	(0.03)	(0.03)	(0.04)	(0.02)	(0.02)	(0.03)
High School GPA	94.73	94.27	0.46	94.22	93.87	0.35
	(0.33)	(0.33)	(0.47)	(0.22)	(0.23)	(0.31)
Observations	146	153	299	364	385	749

Standard errors in parentheses.

Stars are only displayed for the Difference columns.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table reports mean characteristics for students in the Early and Late groups, disaggregated by treatment assignment. To comply with course policy, academic performance measures are expressed relative to the average performance of the Early treatment group. Demographic variables are indicators for male, international, and first-year student status. *Goal Mark* is the student's self-reported target grade (on a 0–100 scale), and *High School GPA* is the self-reported high school grade point average (on a 0–100 scale). *First Generation* is an indicator equal to 1 if the student is the first in their family to attend college. Standard errors are shown in parentheses. Stars indicate significance levels for the treatment–control difference within each timing group. Differences between Early and Late groups in average test performance reflect selective take-up timing and support the motivation for the staggered design.

Table 8: Testing for Spillover Effects Using Pre-Treatment Beliefs of Early Students

	Early		
	Treatment	Control	Difference
Baseline Effect of Good Luck	9.55	9.14	0.41
	(0.44)	(0.40)	(0.59)
Baseline Role of Luck	38.53	37.03	1.50
	(2.26)	(2.16)	(3.13)
Observations	146	153	299

Standard errors in parentheses.

Stars are only displayed for the Difference columns.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table reports pre-treatment beliefs about testing noise for students in the Early treatment and Early control groups, who attempted to access the final survey within the first three hours of its release. *Baseline Effect of Good Luck* is the student's pre-treatment reported estimate of the number of grade points they expect to gain if lucky on the test. *Baseline Role of Luck* is the student's pre-treatment reported belief about the percentage of their prediction error attributable to luck. Standard errors are shown in parentheses. The comparison tests whether Early treated students may have been exposed to treatment content prior to completing the survey. Because Early control students could not have been exposed to spillover effects, and no significant differences are observed between groups, the results suggest that spillover effects from peer information sharing were minimal or absent. Stars indicate significance levels for the difference column only.

Table 9: Average Treatment Effects on Beliefs About Testing Noise

	Treatment	Control	Difference
Between-Group Identification			
Perceived Effect of Good Luck	7.21	8.96	-1.75***
	(0.21)	(0.20)	(0.29)
Perceived Role of Luck	28.53	37.13	-8.60***
	(1.05)	(1.08)	(1.51)
Observations	510	538	1048
Within-Student Identification			
Perceived Effect of Good Luck	7.21	9.04	-1.83***
	(0.21)	(0.21)	(0.16)
Perceived Role of Luck	28.53	38.10	-9.57***
	(1.05)	(1.12)	(0.83)
Observations	510	510	510

Standard errors in parentheses.

Stars are only displayed for the Difference column.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table reports average treatment effects on students' beliefs about testing noise, using both between-group and within-student identification strategies. *Perceived Effect of Good Luck* is the student's reported estimate of the number of grade points they expect to gain if lucky on the test. *Perceived Role of Luck* is the student's reported belief about the percentage of their prediction error attributable to luck. Between-group comparisons are based on randomly assigned treatment and control groups, while within-student estimates compare pre- and post-treatment responses among treated students. All models are estimated using OLS with standard errors clustered at the student level. Stars denote significance levels for the difference column only.

Table 10: Average Treatment Effects on Students' Grade Predictions

	Treatment	Control	Difference
Between-Group Identification			
Absolute Prediction Error	15.30	17.36	-2.07**
	(0.55)	(0.62)	(0.83)
Observations	495	518	1013
Within-Student Identification			
Absolute Prediction Error	15.30	16.22	-0.92***
	(0.55)	(0.59)	(0.33)
Observations	495	495	495

Standard errors in parentheses.

Stars are only displayed for the Difference column.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table reports average treatment effects on students' absolute prediction errors using both between-group and within-student identification strategies. Absolute prediction error is defined as the absolute difference between a student's predicted and realized grade on the final exam. Between-group estimates compare treated and control students using randomized assignment. Within-student estimates compare pre- and post-treatment prediction errors for treated students, using repeated measures within the same survey. Standard errors are shown in parentheses and clustered at the student level. Stars denote significance levels for the difference column only.

Table 11: Joint Distribution of Response Types

Perceived Effect of Good Luck	**Perceived Role of Luck**			
	Aligned	Null	Misaligned	**Total**
Aligned	35.69	10.20	2.35	48.24
Null	13.33	23.14	2.94	39.41
Misaligned	2.75	3.92	5.69	12.36
Total	51.77	37.26	10.98	100

Notes: This table reports the joint distribution of response types based on within-student changes in two belief measures: the perceived effect of good luck on grades (Δe_i) and the perceived contribution of luck to prediction errors (Δr_i). *Aligned Updaters* are students who reported a decrease in the relevant belief after the information treatment, *Null Updaters* reported no change, and *Misaligned Updaters* reported an increase. Rows classify students based on changes in the perceived effect of good luck, while columns classify them based on changes in the perceived role of luck. Cell values represent the percentage of treated students in each combination. The majority of students updated beliefs in the intended direction along at least one dimension, supporting the effectiveness of the intervention.

Table 12: Average Treatment Effects by Response Types

	Absolute Prediction Error		Perceived Effect of Good Luck		Perceived Role of Luck	
Response Types	Pre-Treatment	ATE	Pre-Treatment	ATE	Pre-Treatment	ATE
Panel A: Perceived Effect of Good Luck						
Aligned	15.91	-0.98**	10.03	-4.47***	41.11	-16.53***
	(0.84)	(0.43)	(0.28)	(0.20)	(1.52)	(1.23)
Null	16.08	-1.35***	8.23	0.00	34.62	-5.30***
	(0.91)	(0.37)	(0.36)	(.)	(1.93)	(0.88)
Misaligned	17.90	0.77	7.47	2.85***	36.53	3.90
	(1.92)	(1.81)	(0.52)	(0.36)	(3.45)	(2.87)
Panel B: Perceived Role of Luck						
Aligned	16.43	-1.47***	9.73	-3.19***	44.67	-21.54***
	(0.85)	(0.44)	(0.29)	(0.24)	(1.46)	(1.05)
Null	15.56	-0.35	7.93	-0.57***	30.59	0.00
	(0.89)	(0.36)	(0.34)	(0.15)	(1.96)	(.)
Misaligned	17.44	-0.25	9.26	0.46	31.48	14.07***
	(1.93)	(1.83)	(0.67)	(0.57)	(3.06)	(2.18)
Panel C: Combined Response Types						
Aligned on Effect or Role	16.43	-1.37***	9.71	-3.22***	41.81	-16.87***
	(0.74)	(0.39)	(0.26)	(0.20)	(1.33)	(1.02)
Null on Both Effect and Role	15.37	-0.64	7.58	0.00	30.82	0.00
	(1.18)	(0.39)	(0.45)	(.)	(2.59)	(.)
Misaligned on Effect or Role and Never Aligned	16.69	0.89	8.05	2.11***	31.56	10.00***
	(1.74)	(1.63)	(0.55)	(0.32)	(3.24)	(2.01)

Standard errors in parentheses.

Stars are only displayed for the ATE column.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table reports average treatment effects (ATEs) by students' response types. Response types are defined based on the sign of within-student changes in two belief measures: the perceived effect of good luck and the perceived role of luck. Panels A and B classify students separately based on each measure, while Panel C presents mutually exclusive groups based on whether students aligned on at least one, remained null on both, or misaligned on at least one and never aligned. For each group, we report the pre-treatment average and the ATE in absolute prediction error and both belief measures. Standard errors are shown in parentheses and clustered at the student level. Stars denote significance levels for the ATE columns only.

Table 13: Share of Predictive Gap Closed by the Treatment
Relative to Bayesian and Flexible Benchmarks

	Benchmarks	
	Bayesian	Flexible
Panel A: Between-Group Identification		
Λ^{ATE}	0.39***	0.20**
SE	0.15	0.08
N	862	862
Panel B: Within-Student Identification		
Λ^{ATE}	0.23***	0.11***
SE	0.08	0.04
N	430	430
Panel C: Aligned Students		
Λ^{ATE}	0.30***	0.16***
SE	0.09	0.04
N	275	275

$^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$

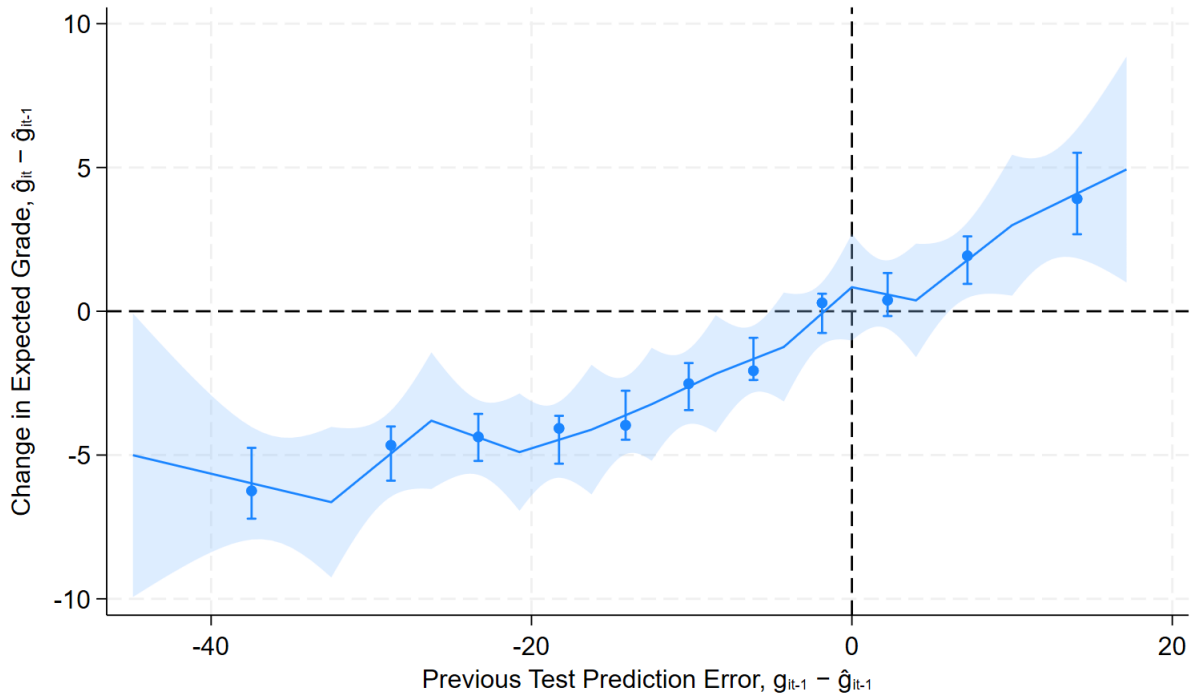
Notes: This table reports estimates of Λ^{ATE}, the proportion of excess prediction error eliminated by the information treatment, relative to two counterfactual benchmarks. For each specification, $\Lambda^{\text{ATE}} = (\Gamma_5^{\text{control}} - \Gamma_5^{\text{treat}}) / (\Gamma_5^{\text{control}} - \Gamma_5^{B})$, where $\Gamma_5^{\text{control}}$ and Γ_5^{treat} denote the average absolute prediction error on the final exam for control and treatment groups, and Γ_5^{B} is the benchmark prediction error. Two benchmarks are used: the *Bayesian benchmark*, which simulates optimal predictions by an agent who shares the student's prior and updates rationally given their Test 4 grade, and the *Flexible benchmark*, which uses a high-dimensional fixed-effects regression model to predict outcomes. The flexible model has an R^2 of 0.85 and overstates predictive accuracy, yielding conservative estimates of treatment gains. Panel A presents estimates based on between-group comparisons. Panel B reports within-student estimates comparing each treated student's post-treatment error to their pre-treatment baseline. Panel C restricts the within-student analysis to students whose belief changes were aligned with the treatment on at least one belief dimension, as defined in Section 5.8. In all panels, standard errors are reported beneath the estimates, and sample sizes reflect the number of observations with non-missing values for both the benchmark and observed prediction error. Stars denote significance levels for Λ^{ATE}.

10 Figures

Figure 1: Association Between Absolute Prediction Errors and Quintiles of Perceived Effect of Good Luck

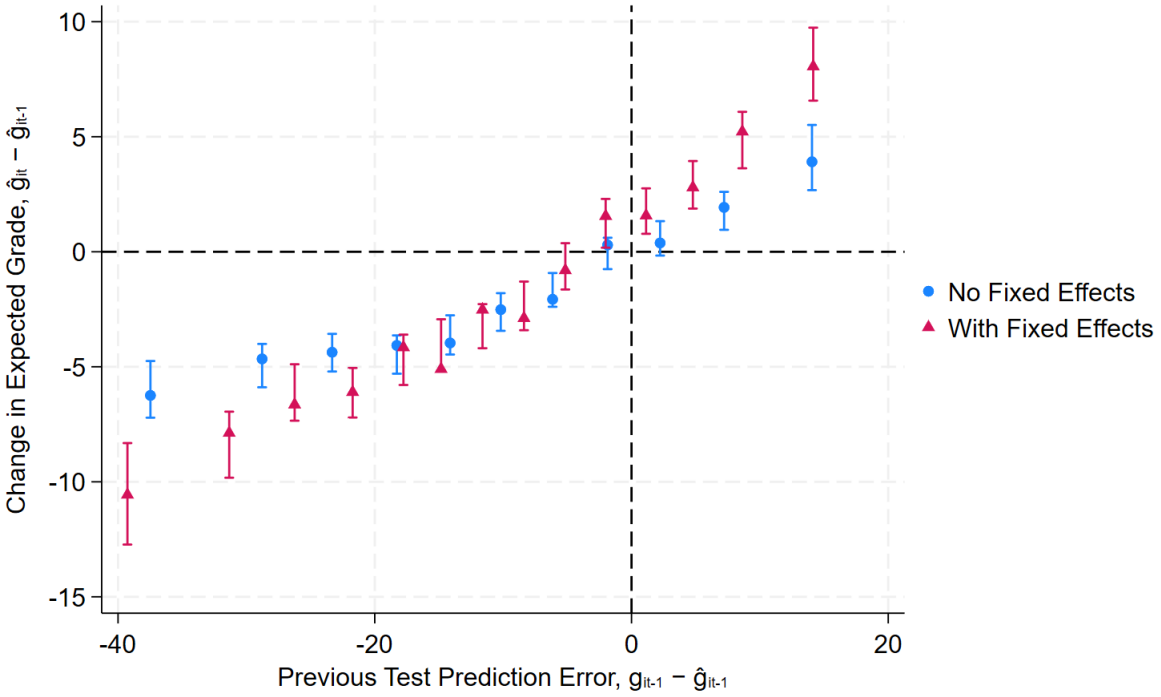
Notes: This figure plots average absolute prediction errors across quintiles of students' perceived effect of good luck on test grades. The perceived effect of good luck is elicited as the number of grade points a student expects to gain if lucky on the test. Error bars indicate 95% confidence intervals based on standard errors clustered at the student level. Students in the top two quintiles of perceived luck effect exhibit significantly larger prediction errors, suggesting that stronger beliefs in the effects of luck are associated with reduced predictive accuracy.

Figure 2: Binscatter Plot of Change in Grade Prediction
by the Previous Test's Prediction Error



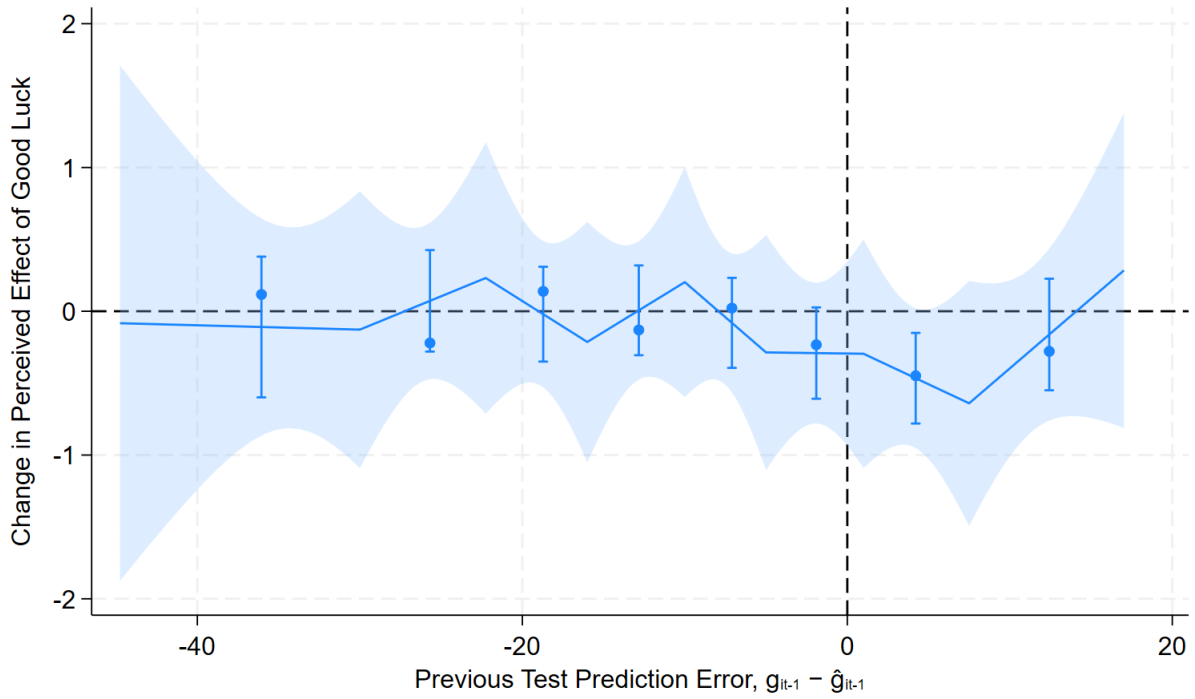
Notes: This figure plots a binned scatterplot of the relationship between the change in expected grade ($\hat{g}_{it} - \hat{g}_{it-1}$) and the prediction error from the previous test ($g_{it-1} - \hat{g}_{it-1}$). Each point reflects the average change in expected grade within a bin of prior prediction error. The blue shaded area represents the uniform 95% confidence band, and vertical lines denote pointwise confidence intervals. Estimates are based on a piecewise linear fit using the `binsreg` package (Cattaneo et al., 2025), with standard errors clustered at the student level. The figure shows that students partially adjust their expectations in response to surprising test outcomes, though the magnitude of revision is modest. We trim the prior prediction error at the 5th and 95th percentiles to reduce the influence of extreme outliers.

Figure 3: Comparing Binscatter Plots of Change in Grade Prediction
by the Previous Test's Prediction Error With and Without Student Fixed Effects



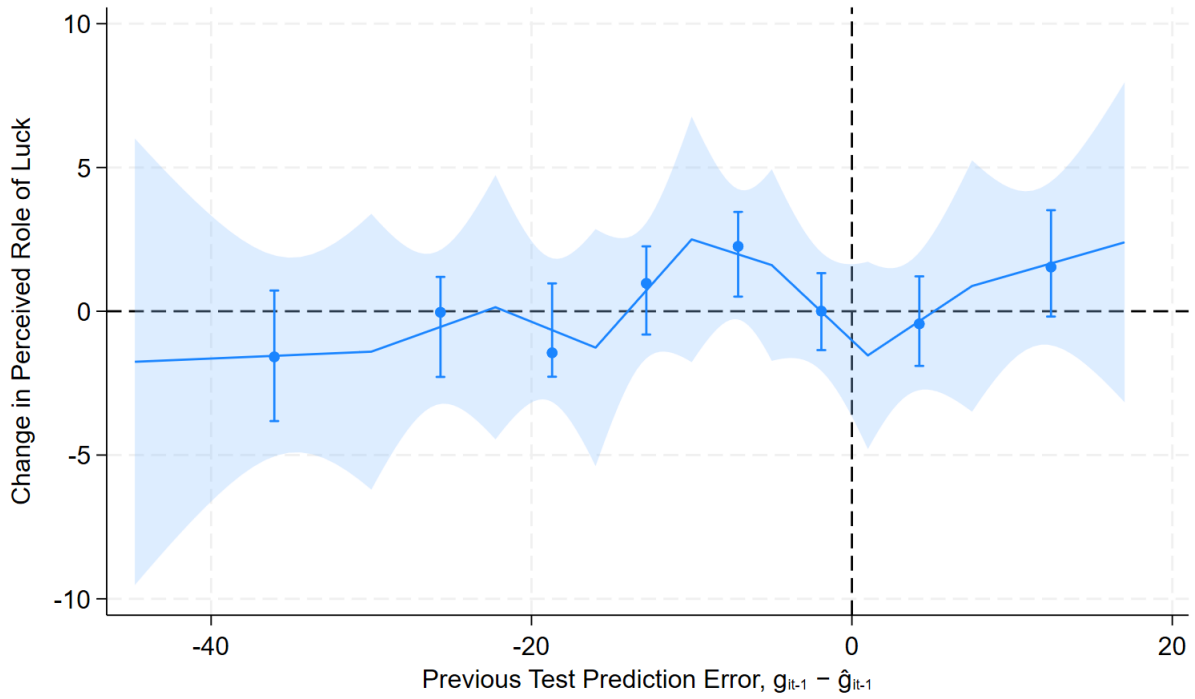
Notes: This figure compares binned scatterplots of the relationship between the change in expected grade ($\hat{g}_{it} - \hat{g}_{it-1}$) and the prior test prediction error ($g_{it-1} - \hat{g}_{it-1}$), estimated with and without student fixed effects. The blue circles show estimates from a model without fixed effects, while the red triangles show estimates from a model that includes student fixed effects. Vertical bars represent 95% confidence intervals with standard errors clustered at the student level. Including student fixed effects leads to steeper estimated responses, particularly among students with large prediction errors, suggesting that controlling for unobserved heterogeneity increases the estimated responsiveness of grade expectations to performance signals. We trim the prior prediction error at the 5th and 95th percentiles to reduce the influence of extreme outliers.

Figure 4: Binscatter Plot of Change in Perceived Effect of Good Luck by the Previous Test's Prediction Error



Notes: This figure shows the relationship between the change in the perceived effect of good luck ($e_{it} - e_{it-1}$) and the prior test prediction error ($g_{it-1} - \hat{g}_{it-1}$), estimated using binned scatterplots with student fixed effects. The sample is trimmed at the 5th and 95th percentiles of the prior prediction error distribution to reduce the influence of extreme outliers. Points represent average changes in beliefs within each bin. Vertical bars indicate 95% confidence intervals, and the shaded region denotes the uniform confidence band. The flat slope suggests that students' beliefs about the effect of luck remain stable regardless of prior surprises in performance.

Figure 5: Binscatter Plot of Change in Contribution of Luck to a Student's Prediction Error by the Previous Test's Prediction Error



Notes: This figure plots the relationship between the change in the perceived contribution of luck to a student's prediction error ($r_{it} - r_{it-1}$) and the prior test prediction error ($g_{it-1} - \hat{g}_{it-1}$), using a binned scatterplot with student fixed effects. The sample is trimmed at the 5th and 95th percentiles of the prior prediction error distribution to reduce the influence of extreme outliers. Points show average changes in beliefs within each bin. Vertical lines represent 95% confidence intervals, and the shaded area denotes the uniform confidence band. The estimates show no significant pattern, indicating that students' beliefs about the role of luck in generating prediction errors are largely unresponsive to prior surprises in performance.

A Appendix

A.1 Randomized Controlled Trial

A.1.1 Identification Assumptions for Within-Student Strategy

Our experimental design elicits each treated student's beliefs immediately before and immediately after exposure to the treatment. Let i index students and $t \in \{\text{pre}, \text{post}\}$ denote the time period relative to treatment exposure. Let $Y_{it}(d)$ denote the potential outcome for student i at time t under treatment status $d \in \{0, 1\}$, where $d = 1$ denotes exposure to the treatment. The treatment is administered between the two measurements $Y_{i,\text{pre}}(1)$ and $Y_{i,\text{post}}(1)$ within a single survey session.

We define the individual treatment effect of interest as:

$$\tau_i = Y_{i,\text{post}}(1) - Y_{i,\text{post}}(0),$$

which corresponds to the causal effect of the treatment in the post-treatment period. The potential outcome $Y_{i,\text{post}}(1)$ is observed for treated students, while $Y_{i,\text{post}}(0)$ is an unobserved counterfactual representing what the student's outcome would have been in the post-treatment period had they not received the treatment.

For treated students, we observe the within-student change in outcome:

$$\Delta Y_i = Y_{i,\text{post}}(1) - Y_{i,\text{pre}}(1),$$

which captures the difference in the outcome immediately before and after treatment exposure. The gap between the individual treatment effect τ_i and the observed change ΔY_i is given by:

$$\tau_i - \Delta Y_i = Y_{i,\text{pre}}(1) - Y_{i,\text{post}}(0) \equiv \nu_i,$$

where ν_i denotes the difference between the pre-treatment outcome under the eventual exposure to treatment, and the post-treatment counterfactual outcome had the student not been treated. In the discussion that follows, we first outline the assumptions required to identify τ_i directly. While these assumptions are informative, they are more restrictive than those necessary for our main analysis. We then turn to the assumptions under which the estimands of interest can be identified from the observed ΔY_i.

A.1.1.1 Building Intuition: Identification of Individual Treatment Effects τ_i

Illustrative Assumption 1 (IA1: Stability of Untreated Potential Outcomes). In the absence of treatment, the outcome would remain constant across the two measurement periods:

$$Y_{i,\text{post}}(0) = Y_{i,\text{pre}}(0).$$

This assumption rules out time-varying shocks that could influence outcomes independently of the treatment between the pre- and post-measurement periods. In our experimental setting, where the two measurements are taken only a few minutes apart, this assumption is likely to hold. However, it is not strictly required for identification. As we discuss below, the analysis can accommodate deviations from this assumption by allowing for idiosyncratic noise in either the pre- or post-treatment periods.

Assumption 2 (A2: No Anticipation). Pre-treatment outcomes are unaffected by treatment status:

$$Y_{i,\text{pre}}(1) = Y_{i,\text{pre}}(0).$$

This assumption rules out anticipation effects, in which students might alter their pre-treatment responses in expectation of receiving the treatment. It also precludes spillover effects from treated peers. In our setting, anticipation is unlikely to be a concern, as the treatment is administered immediately after the pre-treatment measurement and students are not informed of the intervention prior to completing the pre-treatment portion of the survey. Moreover, our experimental design allows for a direct test of spillover effects, as discussed in Section 5.5, and we find no evidence of such effects.

Under IA1 and A2, the treatment effect τ_i can be rewritten as:

$$\begin{aligned}
\tau_i &= Y_{i,\text{post}}(1) - Y_{i,\text{post}}(0) \\
&= Y_{i,\text{post}}(1) - Y_{i,\text{pre}}(0) \quad \text{(by IA1)} \\
&= Y_{i,\text{post}}(1) - Y_{i,\text{pre}}(1) \quad \text{(by A2)} \\
&= \Delta Y_i.
\end{aligned}$$

IA1 and A2 jointly allow for the identification of individual-level treatment effects using within-student changes in outcomes. However, assumption IA1 is more stringent than required for our main analysis, which focuses on average treatment effects. In the next subsection, we relax this assumption and outline an identification strategy for the average treatment effect without relying on the point identification of individual treatment effects.

A.1.1.2 Identification of Average Treatment Effects

We now relax Assumption IA1 by allowing for idiosyncratic variation in untreated potential outcomes across the two measurement periods, and turn to identifying average treatment effects.

Assumption 1 (A1: Mean Stability of Untreated Potential Outcomes).

$$\mathbb{E}\big[Y_{i,\text{post}}(0) \mid i\big] = \mathbb{E}\big[Y_{i,\text{pre}}(0) \mid i\big].$$

This assumption allows for idiosyncratic variation in untreated outcomes across the two measurement periods, but rules out time-varying confounders.

$$\begin{aligned}
\mathbb{E}\big[Y_{i,\text{post}}(0) \mid i\big] &= \mathbb{E}\big[Y_{i,\text{pre}}(0) \mid i\big] && \text{(by A1)} \\
&= \mathbb{E}\big[Y_{i,\text{pre}}(1) \mid i\big] && \text{(by A2)} \\
&= \mathbb{E}\big[Y_{i,\text{post}}(0) + \nu_i \mid i\big] && \text{(by definition of } \nu_i\text{)}
\end{aligned}$$

Hence, Assumptions A1 and A2 together imply that:

$$\mathbb{E}\big[\nu_i \mid i\big] = 0.$$

It follows that:

$$\mathbb{E}[\tau_i \mid i] = \mathbb{E}[\Delta Y_i \mid i].$$

The average treatment effect among a group of treated individuals is therefore identified by the average within-student change in outcome among this group.

A.2 Estimating the Variance of Testing Noise

To identify the variance of the exogenous noise component ϵ_{it} in the grade-generating process, we must separate it from the transitory skill shock η_{it}. Assume that η_{it} follows a stationary AR(p) process:

$$\eta_{it} = \sum_{k=1}^{p} \beta_k \eta_{it-k} + \nu_{it}, \tag{1}$$

where ν_{it} is a shock that is independent of past realizations. We set $p = 2$ to balance flexibility and tractability given that we observe 5 time periods. This yields five unknown parameters: $(\sigma, \sigma_\eta, \sigma_\nu, \beta_1, \beta_2)$.

The variance of η_{it} under the AR(2) process in Equation 1 is:

$$\sigma_\eta^2 = \frac{(1-\beta_2)\sigma_\nu^2}{(1+\beta_2)(1-\beta_1-\beta_2)(1+\beta_1-\beta_2)} \tag{2}$$

which serves as our first identifying equation.

We estimate a fixed effects model of the form:

$$g_{it} - \bar{g}_t = \theta_i + \eta_{it} + \epsilon_{it},$$

where $\bar{g}_t$ is a test fixed effect and θ_i is a student fixed effect. The residual variance from this regression, $\text{Var}(\eta_{it} + \epsilon_{it})$, is estimated as 9.5^2, providing a second equation:

$$\sigma_\eta^2 + \sigma^2 = 9.5^2.$$

Next, we use the variance of first-differenced residuals. Let Δ_{it} denote the first difference of the residuals:

$$\Delta_{it} = (\eta_{it} + \epsilon_{it}) - (\eta_{it-1} + \epsilon_{it-1}).$$

Its variance is:

$$\text{Var}(\Delta_{it}) = \left(\frac{2 - 2\beta_1 - 2\beta_2}{1 - \beta_2}\right)\sigma_\eta^2 + 2\sigma^2,$$

which we estimate as 14.6^2. This provides a third identifying equation. This equality holds only under stationarity, which imposes $|\beta_2| < 1$, $\beta_1 + \beta_2 < 1$, and $\beta_2 - \beta_1 < 1$. We show that these conditions are satisfied by our estimates.

Since η_{it} is unobserved, we cannot directly estimate the AR(2) process. However, we can regress the predicted residuals on their lags:

$$\eta_{it} + \epsilon_{it} = \hat{\beta}_1(\eta_{it-1} + \epsilon_{it-1}) + \hat{\beta}_2(\eta_{it-2} + \epsilon_{it-2}) + \xi_{it},$$

where ξ_{it} represents a measurement error component. These estimates are biased, but under standard assumptions:

$$\text{p-lim}(\hat{\beta}_k) = \beta_k \cdot \frac{\sigma_\eta^2}{\sigma_\eta^2 + \sigma^2}.$$

We estimate $(\hat{\beta}_1, \hat{\beta}_2) = (-0.24, -0.32)$. Both estimates are statistically significant at the 1% level. These deliver the final two identifying equations.

Solving the system yields:

$$(\sigma, \sigma_\eta, \sigma_\nu, \beta_1, \beta_2) = (3.75, 8.72, 7.9, -0.28, -0.37).$$

The implied process is therefore stationary.

Residual Normality

Figure A.1 displays the distribution of residuals, $\eta_{it} + \epsilon_{it}$, from a student fixed effects regression using the demeaned grades on the left-hand side. The residuals appear approximately normal, consistent with the Gaussian assumption on ϵ_{it}.

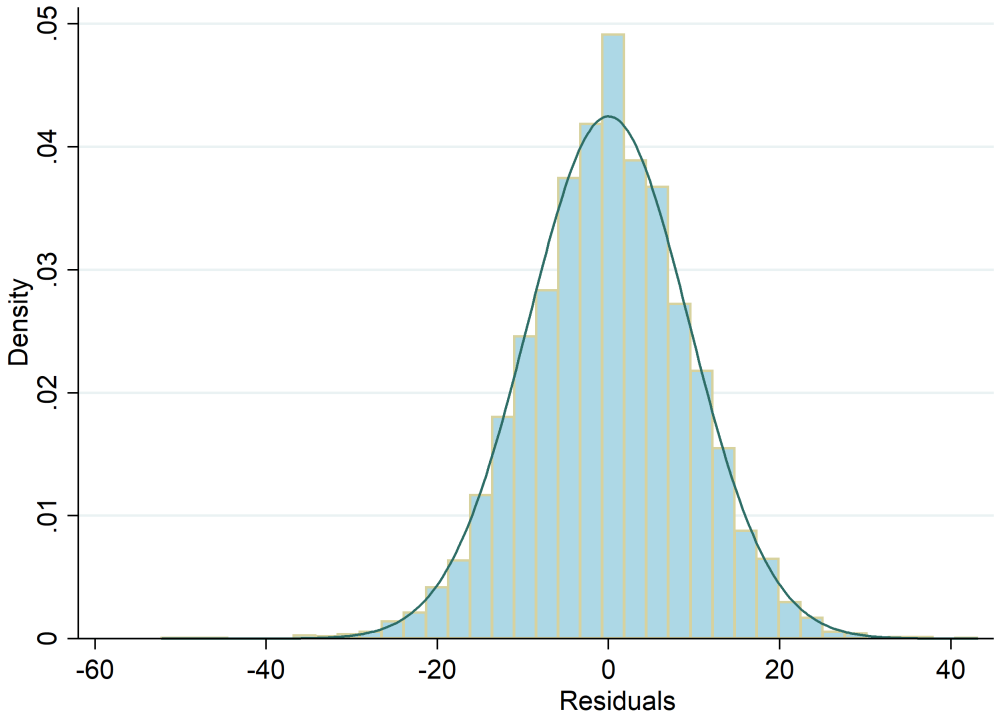


Figure A.1: Residuals from grade model with student and test fixed effects

A.3 Estimating the Subjective Variance of Grade Expectations

To quantify the uncertainty students associate with their own academic performance, we estimate the variance implied by their reported beliefs over grade outcomes. Let g_{it} denote the student's grade in period t, and let $\hat{g}_{it}$ represent the student's subjective expectation of g_{it}, computed as the mean of their reported beliefs.

Each student reports a probability distribution $\{\pi_{ik}\}_{k=1}^{K}$ over a set of discrete grade bins $\{g_1, \ldots, g_K\}$, where $\pi_{ik} = \Pr(g_{it} = g_k)$ and $\sum_{k=1}^{K} \pi_{ik} = 1$. For each grade bin, we use the midpoint value μ_k to construct a student's expected grade:

$$\hat{g}_{it} = \sum_{k=1}^{K} \pi_{ik} \mu_k. \tag{3}$$

The subjective variance is then defined as the expected squared deviation from the student's own prediction:

$$\mathbb{E}[(g_{it} - \hat{g}_{it})^2] = \sum_{k=1}^{K} \pi_{ik}(\mu_k - \hat{g}_{it})^2. \tag{4}$$

This measure captures the extent of uncertainty each student expresses about their future performance. We compute this quantity for each student using their reported probabilities for each test.

A.4 Constructing a Bayesian Benchmark

To assess how closely students' forecasts align with rational updating, we construct a Bayesian benchmark that combines each student's subjective prior with the observed informativeness of grade signals. This benchmark reflects the final exam grade prediction that would be made by a Bayesian agent who shares the student's prior beliefs and updates rationally using empirically estimated signal distributions.

A.4.1 Setup

Let $G_i \in \{g_1, \ldots, g_5\}$ denote the final exam grade bin of student i, and let S_i denote an observed signal prior to the final exam. We use the student's grade on the fourth term test that took place before the final as this signal. Before observing S_i, student i reports a subjective prior $\{\pi_{ik}\}_{k=1}^{5}$ over the final grade bins, where $\pi_{ik} = \Pr(G_i = g_k)$ and $\sum_k \pi_{ik} = 1$. Because students only report posterior beliefs over the final exam grades, we proxy π_{ik} using their reported probabilities of scoring in each grade bin on the fourth term test. The signal S_i is observed prior to the final exam and is informative about G_i.

A.4.2 Empirical Likelihood Estimation

We estimate the conditional distribution of signal values given observed final exam outcomes. For each grade bin g_k, we construct the likelihood function $\ell_k(s) = \Pr(S_i = s \mid G_i = g_k)$ using a nonparametric approach. Specifically, we discretize the signal S_i into Q quantile bins and estimate the conditional probabilities by computing the relative frequency of each signal bin within each grade bin.

To avoid assigning zero probability to unobserved signal-grade combinations, we use

additive smoothing. Specifically, we define the smoothed empirical likelihood as:

$$\ell_k(s) = \frac{n_{ks} + \lambda}{n_k + Q \cdot \lambda}$$

where n_{ks} is the number of students with grade bin g_k and signal bin s, n_k is the total number of students with grade bin g_k, Q is the total number of signal bins, and $\lambda > 0$ is a smoothing parameter. In our implementation, we set $\lambda = 0.001$ and $Q = 200$.

A.4.3 Bayesian Updating

Given the prior $\{\pi_{ik}\}$ and the empirically estimated likelihoods $\{\ell_k(S_i)\}$, we compute the posterior probability that student i belongs to grade bin g_k using Bayes' rule:

$$\tilde{\pi}_{ik} = \Pr(G_i = g_k \mid S_i) = \frac{\pi_{ik} \cdot \ell_k(S_i)}{\sum_{j=1}^{K} \pi_{ij} \cdot \ell_j(S_i)}.$$

This posterior represents how a Bayesian agent with the student's prior beliefs would rationally update upon seeing the signal S_i.

A.4.4 Posterior Prediction

The Bayesian benchmark prediction of the student's final grade is given by the posterior expected grade:

$$\hat{G}_i^{\text{Bayes}} = \sum_{k=1}^{K} \tilde{\pi}_{ik} \cdot \mu_k,$$

where μ_k is the midpoint of grade bin g_k. This benchmark captures the prediction that would be made by a well-specified Bayesian forecaster who incorporates prior beliefs and signal information. Comparing student predictions to this benchmark allows us to quantify deviations from Bayesian updating.

A.5 Appendix Tables

Table A.1: Pairwise Correlation Matrix of Grades

	Test 1	Test 2	Test 3	Test 4	Test 5
Test 1	1				
Test 2	0.78	1			
Test 3	0.73	0.79	1		
Test 4	0.67	0.71	0.73	1	
Test 5	0.69	0.76	0.80	0.81	1

Notes: This table reports pairwise Pearson correlation coefficients between test grades across the five major assessments in the course. Each entry shows the correlation between grades on the corresponding pair of tests, based on the sample of students who took both tests. All coefficients are statistically significant at the 1% level. The high correlations indicate that prior performance is highly predictive of subsequent outcomes.

A.6 Survey Instructions

A.6.1 Introduction

Welcome to the research study!

This study is conducted by the Economics Department at the University of Toronto. **Your responses are strictly anonymous and stored in a secure server. They will not be shared with course instructors or teaching staff.**

This survey should take about 7 minutes to complete. **You can only the submit your response once**. If this is your first survey, it might take slightly longer.

You will receive 0.4 MAT137 course mark, for completing this survey today. Your participation in this research is voluntary. You have the right to withdraw at any point during the study.

Additionally, we request your consent to access your responses for research purposes. Your identity will be kept confidential throughout the research and publication stages of the project. If you consent to be part of the study, **you have the chance to win a cash reward up to $20 in each round of survey.**

Frequently Asked Questions

Who can I contact if I have any questions regarding the survey?

If you have any questions, concerns or need additional information about this study, you can reach the research team using the contact form here.

Who can I contact if I have complaints or concerns regarding the survey?

If you have any concerns or complaints about your rights as a research participant and/or your experiences while participating in this study, contact the Office of Research Ethics at ethics.review@utoronto.ca or 416-946-3273.

Could participating in the study be bad for me?

We do not think there is anything in this study which could harm you. On the contrary, we expect that this study may help you better understand your own study habits and improve your study effectiveness. In addition, you could win a cash reward.

What will you do with the study results?

We will use the results to improve future student course satisfaction and we expect the results to be published in an academic journal.

By clicking the button below, you acknowledge:

- Your participation in the study is voluntary.
- You are aware that you may choose to terminate your participation at any time for any reason.

A.6.2 Effort Elicitation

Think back to last week, that is **the week starting on Monday, April 3rd and ending on Sunday, April 9th**.

How many hours did you spend on studying for MAT137 during that week, outside lectures and tutorials?

Your answer last survey: Fewer than 5 hours (0-1 hour per weekday)

- ○ Fewer than 5 hours (0-1 hour per weekday)
- ○ 5 to 10 hours (1-2 hours per weekday)
- ○ 10 to 15 hours (2-3 hours per weekday)
- ○ 15 hours or more (more than 3 hours per weekday)

To the best of your memory, how many hours did you spend on studying for MAT137 during that week, outside lectures and tutorials?

Now think back to the past 24 hours before this survey. How many hours did you spend on studying for MAT137 **in the past 24 hours, outside** lectures and tutorials?

Your answer last survey: 6 hour(s)

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

Hours

Done

How many hours do you plan to study for MAT137 **on the 3 days between Monday, April 17th and the final exam on Wednesday, April 19th**, outside lectures and tutorials?

Your answer last survey: 15 hour(s)

A.6.3 Belief Elicitation: Grades

The following questions may be selected for a cash reward.

After each round of survey, 16 students will receive a cash reward ranging $5 to $20. The 8 students who produced the best predictions will each receive $20. The other 8 winners are **randomly** drawn. The amount of cash they receive is based on the quality of their prediction in a randomly selected question.

In short, the reward is determined by the accuracy of your predictions. **In order to maximize your payment, you should answer truthfully. Individual responses are strictly anonymous and will never be shared with instructors.**

- ○ Yes, I understand.
- ○ No, I do not understand.

The following questions are about test 4 on March 24, 2023. The test was out of 40.

These questions may be selected for payment. The reward is determined by the accuracy of your predictions. In order to maximize your payment, you should answer truthfully. Individual responses are strictly anonymous and will never be shared with instructors.

In the previous survey, you predicted the class average for test 4 would be 24.4 out of 40.

What do you think the class average actually was for test 4 on March 24?

0 4 8 12 16 20 24 28 32 36 40

Test Average out of 40

In the previous survey, you predicted that you would outperform **34%** of the students in test 4.

What percent of MAT137 students do you think that <u>you actually outperformed</u> in test 4 on March 24?

Ex: if your answer is 80, it means you believe that you scored <u>higher</u> than 80% of your classmates. In other words, you believe that you were be in the top 20% of the class.

You were in the **bottom** of the class — **Median** — You were at the **top** of the class

0 10 20 30 40 50 60 70 80 90 100

Percent

In the previous survey, you predicted male students would score 2.8 point(s) (out of 40) higher in test 4 compared to female students, on average.

How much <u>higher</u> do you think male students <u>actually</u> scored in test 4 on March 24, compared to female students, on average?

Women did better — Men did better

-10 -8 -6 -4 -2 0 2 4 6 8 10

Points out of 40

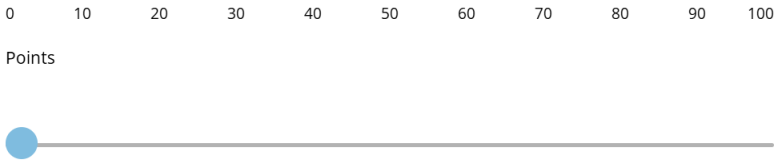
-2

The following questions are about the final exam on Wednesday, April 19. In all of the questions, you can assume that the final exam is graded out of 100 points.

These questions may be selected for payment. The reward is determined by the accuracy of your predictions. In order to maximize your payment, you should answer truthfully. Individual responses are strictly anonymous and will never be shared with instructors.

What do you think will be the **class average** in the **final exam** (out of 100)?

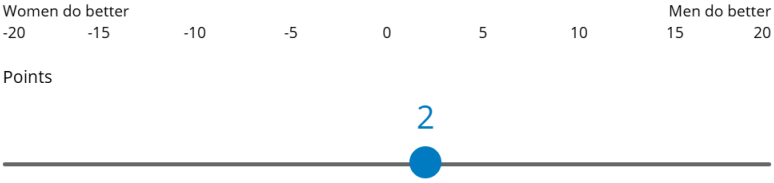
Your answer for **test 4**:60



Assume the upcoming final exam is out of 100 points, **how much higher do you think male students will score compared to female students, on average?**

For example, if your answer is **20**, it means you believe that male students will score 20 points **higher** on average. If your answer is **-20**, it means you believe that male students will score 20 points **lower** on average.

Your answer last survey: 7



Assume the final exam is out of 100 points, **how likely are the following events (as a percent)?**

If your answer is 0, it means the event will never happen. If your answer is 100 , it means the event will happen with absolute certainty. The five numbers should add up to 100.

Men perform **a lot better** (e.g. more than 5 points higher) than women on average	0
Men perform **slightly better** (e.g. 2 to 5 points higher) than women on average	0
Men perform **about the same** (e.g. between 2 points lower and 2 points highter) as women on average	0
Men perform **slightly worse** (e.g. 2 to 5 points lower) than women on average	0
Men perform **a lot worse** (e.g. more than 5 points lower) than women on average	0
Total	0

Done

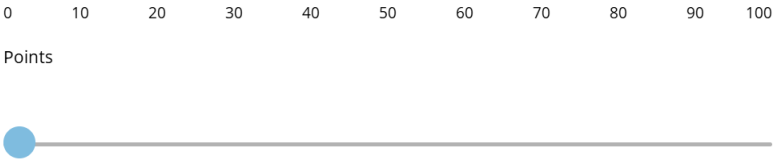
On average, which group of students do you think will have a better outcome?

For example, if your answer is "small male advantage", it means that you think male-identifying students will do better than female-identifying students, but the difference is not very large.

	Strong female advantage	Small female advantage	About the same	Small male advantage	Strong male advantage
Performance in **MAT137** (the course, not just the final exam)	○	○	○	○	○
Effort and work ethics in **MAT137**	○	○	○	○	○
Performance in **MAT137, if both groups worked equally hard**	○	○	○	○	○
Natural ability in **math** (the discipline, not just MAT137).	○	○	○	○	○

You estimated that the average is going to be **68**. **What grade do you think you will get in the upcoming final exam?**

Your answer for **test 4**: 59

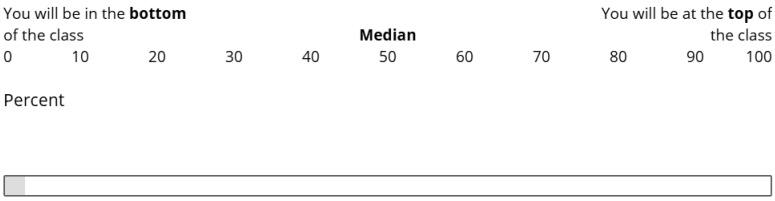


Done

What percent of MAT137 students do you think that you will outperform, in the upcoming exam?

Ex: if your answer is 80, it means you believe that you will score higher than 80% of your classmates. In other words, you believe that you will be in the top 20% of the class.

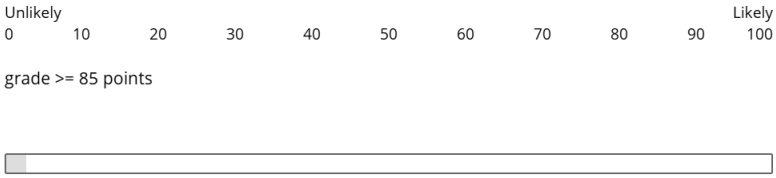
Your answer for **test 4**: 34



How likely (as a percent) do you think that <u>your</u> final exam grade is going to be higher than the following cutoffs?

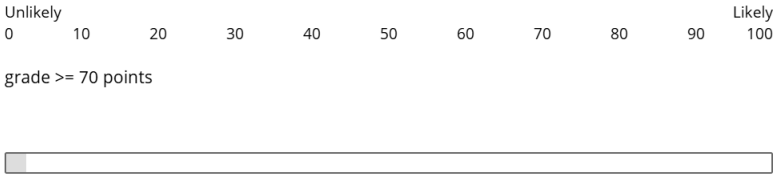
0 means it is impossible. 100 means it will happen with absolute certainty.

Your answer for **test 4**: 11%



Done

Your answer for **test 4**: 40%



A.6.4 Belief Elicitation: Testing Noise

You estimated that you would score **53** out of 100 in the final exam.

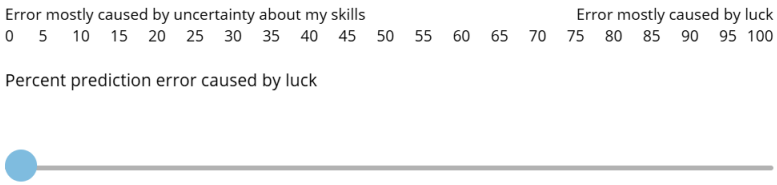
Making an accurate prediction is difficult because grades are determined by both **your MAT137 skills** and **luck.**

Let's call the difference between your actual grade and your predicted grade **"prediction error"**.

Prediction errors exist because either you were not very sure about your MAT137 skills when you made the prediction, or because you could not have possibly foreseen your luck (e.g. generous grading, poor sleep the night before,...).

How much (as a percent) do you think luck contributes to your prediction error?

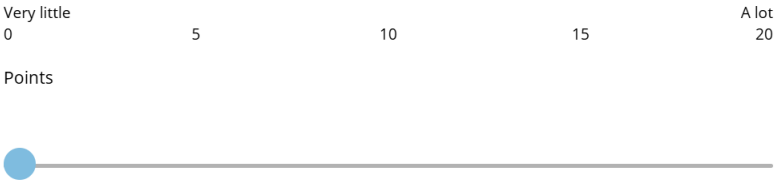
Your answer last survey: 60



Luck can have a positive or negative impact on one's test scores.

How much higher (in points, out of 100) **do you think you would be able to score in the upcoming final exam, if you were struck by good luck?**

Your answer last survey: 5



Some students can experience fortunate events that benefit their test outcome, e.g. coming across similar questions during the review process.

Consider all of these "lucky" students, and the effect of good luck. How likely are the following events (as a percent)?

For example, 0 means it will never happen; 100 means it will happen with absolute certainty. The three numbers should add up to 100.

Your answer last survey: 80\15\5

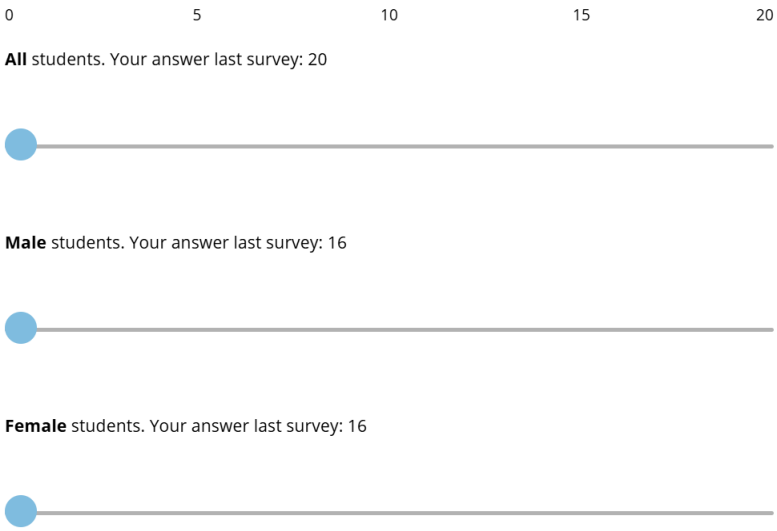
On average, luck increases their test score **slightly** : 0 to 3 points	0
On average, luck increases their test score **moderately** : 3 to 10 points	0
On average, luck increases their test score **dramatically** : more than 10 points	0
Total	0

A.6.5 Other Questions

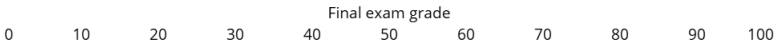
The following questions may be selected for payment. The reward is determined by the accuracy of your predictions. In order to maximize your payment, you should answer truthfully. Individual responses are strictly anonymous and will never be shared with instructors.

You reported that you studied **10 to 15 hours (2-3 hours per weekday)** last week, that is **the week starting on Monday April 3, and ending on Sunday April 9**

How many hours do you believe **other students in MAT137** spent on studying for this course during that week, **outside** lectures and tutorials?



What grade do you think you would get in the upcoming final exam, if you consistently studied the following numbers of hours every week?



0 hours per week. Your answer last survey: 17

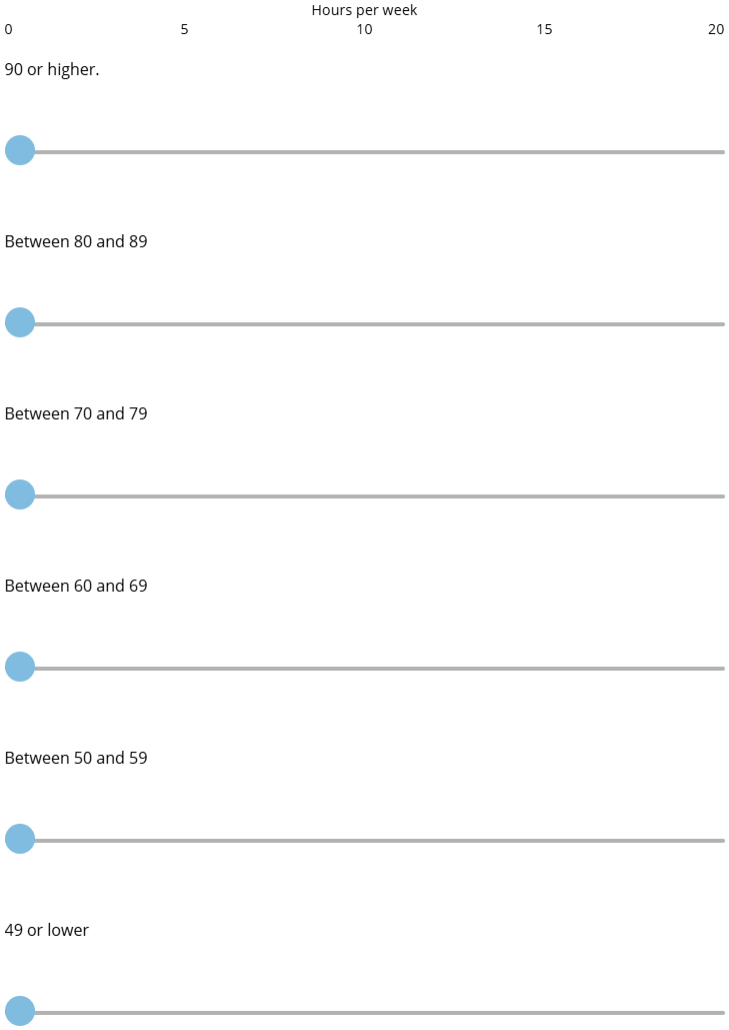
5 hours per week. Your answer last survey: 28

10 hours per week. Your answer last survey: 71

15 hours per week. Your answer last survey: 81

more than 15 hours per week. Your answer last survey: 88

What is the **maximum** hours you are willing to study **weekly** to **guarantee** the following grades in the **upcoming exam**?



At University of Toronto, letter grades and numerical marks follow the conversion scale below:

A+: 90-100
A : 85-89
A- : 80-84

B+: 77-79
B : 73-76
B- : 70-72

C+: 67-69
C : 63-66
C- : 60-62

D+: 57-59
D : 53-56
D- : 50-52

F : 0-49

The following question may be selected for payment. The reward is determined by the accuracy of your predictions. In order to maximize your payment, you should answer truthfully. Individual responses are strictly anonymous and will never be shared with instructors.

What do you think **your course mark in MAT137** will be?
Your answer last survey: 60

0 10 20 30 40 50 60 70 80 90 100

Mark

Your responses are **strictly confidential** and will not be shared with your teachers.

The results are used for research purposes only and will not have an impact on your instructors.

	Completely untrue	Mostly untrue	Somewhat	Mostly true	Completely true
My classmates behave the way my instructor wants them to.	○	○	○	○	○
My instructor explains difficult things clearly.	○	○	○	○	○
In this class, we learn a lot in each lecture.	○	○	○	○	○
My instructor makes learning enjoyable.	○	○	○	○	○
My instructor makes me want to learn more math.	○	○	○	○	○