Updating and Misspecification: Evidence from the Classroom*

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

August 27, 2024

Abstract

Misspecification is theoretically linked with updating failures, but empirical evidence has been lacking. We document the empirical relevance and estimate the impact of misspecification on updating. We collect a novel high-frequency dataset on students' beliefs about grades in a freshman course. Students are overconfident, their beliefs do not improve over time, and they overestimate the testing noise by a factor of 3. Our RCT exogenously shocks and improves students' belief in the testing noise. Treated students reduce their prediction errors by 32%. We estimate the impact of misspecification structurally and find that a lower bound of 25% of prediction errors can be attributed to misspecification. Our finding suggests that misspecification is a major obstacle to processing information correctly, but it can be alleviated via simple interventions.

keywords: belief updating, misspecification, student learning

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

*We thank the instructors of MAT135/137 for their involvement in the project, with particular thanks to Xiaoyue Cui and Bernado Galvao-Souza. We also thank Michael Baker, Robert McMillan, Philip Oreopoulos, 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 Toronto; University of Oxford

1 Introduction

Why do individuals fail to learn from highly informative signals? What can help them learn better?

The bulk of the literature on belief updating is experimental and has focused on behavioral and cognitive explanations to explain the failure of Bayesian updating.[1] These explanations range from overconfidence and ego-driven motivated beliefs[2] to highlighting the computational difficulty of Bayesian updating.[3] It is salient from introspection that people are often over-confident and averse to updating negatively about themselves. Similarly, Bayesian updating was discovered contemporaneously with calculus, highlighting its non-triviality. If these channels drive a large share of belief updating mistakes, then it may be challenging to improve how people update. Learning probability theory is hard, and learning to be less egocentric is perhaps impossible.

We posit that a perceptual channel also drives updating failures in the real world. In empirical settings, distinct from the lab, agents may face uncertainty and be *misspecified* regarding how to interpret information. Theoretically, such agents will update suboptimally even when they apply Bayes rule correctly. Therefore, this channel differs from *updating biases* of the earlier channels. Misspecification has been studied theoretically, initially by Berk (1966), which shows that a misspecified Bayesian belief need not converge to the truth asymptotically. More recently, the implication of misspecification has been explored in social learning (Heidhues et al. (2018); Frick et al. (2020, 2023)) as well as for individual belief updating failures (Fudenberg et al. (2021); Bohren and Hauser (2023)). However, we are unaware of works that empirically document the relevance in the real world.[4] In this paper, we empirically document the phenomenon's existence, quantify the extent to which it impacts belief updating, and study whether one can alleviate its adverse effects.

Our empirical setting is a large first-year class in calculus. We study students who receive past test grades and predict their future grades. The correlation coefficient between tests is 0.8; hence, past test grades are highly informative signals. Previous studies (Zafar (2011); Stinebrickner and Stinebrickner (2014); Wiswall and Zafar (2015); Oreopoulos and Petronijevic (2023)) have shown that students have difficulty in learning, which can lead to costly mistakes in the allocation of study hours, dropping and field majoring decisions. In our setting, students begin the term with an absolute prediction mistake of 15.62

[1] See Benjamin (2019) for a fairly recent survey.

[2] Ertac (2011); Eil and Rao (2011); Buser et al. (2018); Coutts (2019); Drobner (2022); Möbius et al. (2022).

[3] Grether (1980); Amelio (2022); Guan (2023); Gonçalves et al. (2024).

[4] Castillo and Youn (2023) and Chiara and Florian H. (2024) explore this issue in the lab.

percentage points (pp) and are generally overconfident. We observed no improvement in beliefs after four test grades, as their prediction mistake on the fifth test was 16.8pp.

We hypothesize that, along with updating biases, students are misspecified and underestimate the correlation between tests. They may overattribute the realized grade to noise rather than their underlying ability. This would result in underreacting to the information received. We then would expect that informing students about the tests' informativeness, without the need to teach them how to update or give them additional information, would improve their predictions.

We conduct a randomized control trial (RCT) and shock exogenously students' beliefs regarding the testing noise. Importantly, our treatment also must *not* change their belief in their own ability. Our treatment informs students about the correlation between test grades in a salient manner without specific reference to any particular grade. Therefore, there is no information relevant to themselves that the student can infer from the treatment. Yet, we find that treated students make 31% less absolute prediction mistakes. We take this to be strong causal evidence in favor of misspecification.

Such a treatment is challenging to perform in most empirical settings for three reasons. First, we aim to give subjects truthful information. This requires the researchers to know the testing informativeness. We overcome this by accessing past year's test grades. This allows us to credibly inform students about testing informativeness and reduces the information's relevance for inferring their ability. Second, spillover effects pose a real threat. If the information we give is helpful and impactful, then we must acknowledge that students might share it with others. We overcome this with a staggered rollout of the treatment. We also checked and confirmed that there were no spillover effects (which would only bias our results towards the null). Third, we want to recover the treatment effect on updating *only*. However, if treated students also studied harder, leading to better predictions, we would recover two confounding treatment effects. Put differently, we are in a non-standard situation where we do not want the treatment to induce behavioral changes. Therefore, we administered the treatment only three days before the final exam to limit behavioral adaptation, and we also observed no difference in the performance between the control and treatment groups.

While our RCT provides causal evidence on the existence of the channel, it does not measure the extent to which it impacts belief updating. To do so, one needs to recover both the objective and subjective, potentially misspecified information structures. Recovering the objective information structure requires a large sample of repeated measurement, which we have in our setting with a large class and multiple tests. Estimating the subjective information structure empirically is challenging, as this requires recovering the agent's

belief regarding the variance of the noise. To our knowledge, we are the first paper to do so empirically. We estimate the objective testing noise's standard deviation to be 3.75, while students overestimate it to be 11, more than twice as much. Students who believe the testing noise is higher also make larger prediction mistakes. Using a structural model, we estimate the absolute prediction mistakes of a correctly specified and misspecified Bayesian agent. Comparing the two along with the empirical absolute prediction mistake offers us a measure of the effect of misspecification without committing to a particular model of non-Bayesian updating. We find that 25% of prediction mistakes (lower bound) are due to misspecification, and the effect of misspecification is particularly strong for the first test and the final exam.

Taken together, our results suggest that misspecification is a relevant channel for the failure of updating, but unlike other channels, it can be alleviated via simple interventions.

The paper is organized as follows. In section 2, we provide an exposition of our channel and our results via a conceptual framework. Section 3 goes over our empirical setting and data. Sections 4 and 5 provide descriptives and results from our RCT. Section 6 uses a structural model to estimate the impact of misspecification, and section 7 concludes the paper.

2 Conceptual Framework

Consider a simple model to illustrate misspecification's role in the failure of belief updating. Denote periods by $t \in \{0, 1\}$. Suppose for simplicity that grades are generated by

$$g_t = \theta + \epsilon_t.$$

Where g_t is the grade of a student at time t, and it is a function of the student's ability θ and a testing noise $\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$. The student believes instead that $\epsilon_t \sim \mathcal{N}(0, \tilde{\sigma}_\epsilon^2)$ and faces uncertainty regarding her ability as $\theta \sim \mathcal{N}(\mu_\theta, \tilde{\sigma}_\theta^2)$. In our data, the vast majority (90%) of students overestimate the testing noise, $\tilde{\sigma}_\epsilon^2 > \sigma_\epsilon^2$. We ask students to predict their next test grade, g_2, after observing g_1. After observing g_1, one should predict their expected belief of θ. For a Bayesian and misspecified student, the expected grades g_2^B and g_2^M are

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

Note that, for a Bayesian, the larger σ_ϵ^2 is, the less they update towards the signal. Since the larger σ_ϵ is perceived to be, the less weight the student puts on the signal g_1 as opposed

to her prior μ_θ. Therefore, an overconfident student with an incorrect belief $\mu_\theta > \theta$ will update slower if they believe the testing noise to be high. Additionally, the student's actual expected grade $\hat{g}_2$ may differ from both g_2^B and g_2^M if she displayed biased non-Bayesian updating.

Previous studies in education economics[5] as well as behavioral economics[6] have both documented a failure of learning. It is standard to collect the expected grade, $\hat{g}_2$, and prediction error, $|g_2 - \hat{g}_2|$. However, it is necessary to compute the Bayesian posterior *distribution* to quantify the failure of Bayesian updating. This requires knowledge of the prior μ_θ and σ_θ. Additionally, in an empirical setting, we also need to collect sufficient data to estimate the testing noise σ_ϵ and minimize other potential unobserved signals. To overcome these challenges, we repeatedly collect students' expected grades and their belief distribution right before and after each test in an incentive-compatible manner. Collecting their belief distribution allows us to recover the prior, and having repeated measures enables us to estimate σ_ϵ. We find that test grades are highly correlated between periods, hence extremely good signals.[7] The testing noise, σ_ϵ, is estimated to be 3.75 on average. Given this and the student's biased prior beliefs, we find the Bayesian absolute prediction error is 11.4pp on average. Yet, the average prediction error of students is 17.6pp, which does not get smaller over time.

Collecting these distributional variables still would not allow a researcher to disentangle whether updating failure is due to biases or misspecification. To do so, the researcher must recover the objective testing noise σ_ϵ and the subjective noise $\tilde{\sigma}_\epsilon$. We, therefore, additionally collect belief data regarding the testing noise. As σ_ϵ is a complex object, one cannot elicit it directly by asking. We elicit it indirectly in two ways. First, we elicit the *effect of good luck* on grades, formally, it is $\mathbb{E}[\epsilon \mid \epsilon \geq 0]$. Given the normal distribution, this maps bijectively to σ_ϵ. Simultaneously, we also elicit the proportion of prediction mistakes due to testing noise as opposed to uncertainty about their ability. As the prediction error squared, $\mathbb{E}[(g_t - \hat{g}_t)^2]$, is equal to $\sigma_\theta^2 + \tilde{\sigma}_\epsilon^2$ (without the need of normality), this also allows us to recover σ_ϵ^2. These independent methods yield close and consistent measures of $\tilde{\sigma}_\epsilon = 11.32$ and $\tilde{\sigma}_\epsilon = 11.00$ on average. Therefore, students overestimate the testing noise by almost three times.

We disentangle the effect of biased updating and misspecification by examining how changes in students' prediction errors over time can be attributed to these factors. We focus on the changes in prediction error as primitive and do not restrict our analysis to

[5]See Zafar (2011); Stinebrickner and Stinebrickner (2014); Wiswall and Zafar (2015); Oreopoulos and Petronijevic (2023).

[6]See Grether (1980); Coutts (2019); Barron (2021); Möbius et al. (2022).

[7]The correlation coefficient between tests is around 0.8.

any particular model of non-Bayesian updating. Let $\Gamma_t = \sum_i |\hat{g}_{it} - g_{it}|$ denote the observed prediction error, we denote by $\Gamma_t^B = \sum_i |g_{it}^B - g_{it}|$ and $\Gamma_t^M = \sum_i |g_{it}^M - g_{it}|$ the Bayesian and misspecified agents' prediction error. Then consider $\Lambda_t = \frac{\Gamma_t^M - \Gamma_t^B}{\Gamma_t - \Gamma_t^B}$, the denominator is a measure of failure of updating as Γ_t^B is the theoretical minimal amount of mistakes one can make. The numerator measures a misspecified Bayesian's mistakes vis-a-vis a correctly specified Bayesian. In general, we expect $\Gamma_t^B \leq \Gamma_t^M \leq \Gamma_t$, therefore $\Lambda_t \in [0, 1]$. If $\Lambda_t = 1$ for all time periods, then all reductions in mistakes are due to misspecification, as incorporating Bayesian updating alone does not reduce aggregate mistakes. Similarly, if $\Lambda_t = 0$, we see that misspecification does not introduce any prediction mistakes. Therefore, Λ_t captures the proportion of mistakes explained by misspecification.

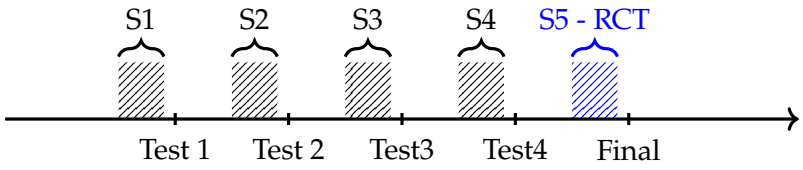
We find that misspecification plays a bigger role in the first test and on the final exam. In particular, Λ_t is 0.71 and 0.55 for the first test and final exam, respectively. It is lower, at 0.26 and 0.29 for the second and third tests, respectively. We note that these should be considered lower bounds for several reasons. First, our estimate of Γ_t^B is an upper bound because our Bayesian model is itself misspecified and students may observe more information than we have. Therefore, an actual Bayesian should be making even less mistakes than our estimate. Second, Γ_t^M is likely a lower bound as we cannot cover all potential misspecifications in our misspecified model. For instance, we assume students' misspecified model has normal testing noise. As testing noise is normal in our data, this assumption removes some potential prediction mistakes due to misspecification.

To offer further evidence that misspecification plays a key role in belief updating, we conduct an RCT before the final exam. Theory and our data show that students who believe the testing noise to be higher also make worse predictions. Therefore, we conduct an RCT which informs students that σ_ϵ, the testing noise, is very low - the treated group became more responsive to information and improved by lowering their prediction mistakes by 32%. Students are given an information treatment that provides exogenous shocks to $\tilde{\sigma}_\epsilon$. Importantly, this treatment does not offer them any information regarding their ability; hence, it *only* impacts their belief regarding testing noise. We found that the treatment successfully informed students that the testing noise was lower than they thought, and reduced their prediction error. To quantify the magnitude of reduction, we consider $\Lambda^{treat} = 1 - \frac{\Gamma^{treat} - \Gamma_5^B}{\Gamma^{control} - \Gamma_5^B}$. Here Γ^{treat} and $\Gamma^{control}$ are the prediction error of the treatment and control groups. If the treatment, which reduces only misspecification, is fully effective and there is no updating bias, then the numerator of the second term is 0 and $\Lambda^{treat} = 1$. Alternatively, if the treatment is ineffective or misspecification does not contribute to prediction mistakes, then $\Gamma^{treat} = \Gamma^{control}$ and $\Lambda^{treat} = 0$. We find $\Lambda^{treat} = 0.32$ and interpret this as our treatment reduced prediction mistakes by around 32%.

3 Empirical Setting and Data

Background. We collected data from a large first-year calculus course that is the prerequisite for most STEM majors at the University of Toronto. The course was run from September 2022 until April 2023. A total of 1,508 students start in the class, and a total of 1,155 finish the class.

Survey. The class grade is determined mainly by 4 midterms and a final exam, which make up 70% of their grades, along with some other minor components such as problem sets and attendance. We ran five surveys, one for each test, and asked students to predict their grades and elicited other variables. These surveys were run three days before their next test, and students always had their last test grade returned before the survey.[8] To ensure a high takeup rate, we incentivized survey completion with a participation grade in the class, totaling up to 2% for completing all five surveys (0.4% for each survey). This led to a high takeup rate of 88% and 90.7% on the first and last surveys, respectively. Additionally, grade predictions are incentivized based on accuracy.[9] One important institutional feature of the class is that the instructors do not release any average, median, or other statistical information regarding the grade distributions - and there is no curving. The instructors clearly state that the grade you see is the grade you get. Therefore, we take students to believe their grades are not influenced by their peers nor accounting for curves when making inferences.



Variables. In this paper, we focus on a subset of the collected variables regarding student beliefs.[10] First, we observe each student i's grades across the five tests, which we denote by g_{it}, for $t \in \{1, 2, ..., 5\}$. We ask each student for their expected grade, $\hat{g}_{it}$, and we denote by $\Gamma_{it} = |g_{it} - \hat{g}_{it}|$ the absolute prediction error. Along with the expected grade, we also elicit the probability this grade falls in the different ranges. In particular, we ask for the probability that g_{it} is less or equal to X, for $X \in \{50, 60, 70, 85\}$.[11] Formally, we

[8] Additionally, we remind them of their grades.

[9] We pay $20 to the best 20 predictions, and an additional 20 random students are paid using an incentive-compatible mechanism. Students are told that truthful reporting of their best guesses maximizes their expected earnings.

[10] For the set of collected variables, please see the appendix for the full survey.

[11] We chose 85 instead of 90 as the University of Toronto counts all grades above 85 as a 4.00 (perfect for GPA).

denote this by $G_{it}(X) = p(g_{it} \leq X)$. As the average grade is not publicized, we also elicit students' beliefs regarding the class average and denote it by $\hat{\bar{g}}_{it}$.

Finally, we elicit their belief regarding the informativeness of testing. As the variance of the testing noise is a complex object, we do not elicit it by asking directly. First, we highlighted to them that "luck" can impact their grade exogenously and gave them examples such as good or bad sleep, a harsh or generous grader, and studying for the right or wrong questions. They are then asked how much higher they expect their grades to be if they were positively impacted by luck, which we denote by e_{it}, and formally, $e_{it} = \mathbb{E}[\epsilon_{it} \mid \epsilon \geq 0]$. Additionally, we ask them for the proportion of prediction errors r_{it} due to luck as opposed to uncertainty regarding their ability.

4 Descriptive Summary

Test Grades. Students have five grades from the tests, and the averages are very consistent over time, with the first-term test being on the easier side. We note that most of these students need to pass this class (score above 50), and often, many need a higher grade to qualify for different majors.[12] At the instructor's request, we do not share class averages; nevertheless, in our dataset, only 64% of the initial student body has a passing grade.[13] Therefore, this is an extremely challenging class, and learning about one's ability in this class is beneficial. The correlation between tests is also very high, suggesting that test grades are highly predictive of each other and that past grades are very informative signals of future performances. The correlation, $corr(g_{it-1}, g_{it})$ is $(0.78, 0.79, 0.73, 0.81)$ between the 5 time periods.

Student Beliefs Regarding Grades. As in the figure below, students begin with optimistic beliefs, and this optimism has a slight downward trend over time. On average, the prediction errors are 20.72pp and 11.32pp when the student overestimates and underestimates her grades, respectively. On average, a student who receives a grade lower and higher than her prediction adjusts her next period's prediction down and up by 3.95pp and 3.01pp respectively.[14] In our sample, 69% and 29% of predictions are greater and lower than the realized grade, respectively. Both of these errors and the percentage are stable throughout the periods. In general, we see that prediction errors persist and remain high throughout the class despite what are statistically informative signals. The descriptives suggest students are overconfident and are underreacting to the information. However,

[12]The data science specialist program requires a grade higher than 70% in this class.

[13]Note: this includes assignments, attendance, surveys, and other grades besides the test grades.

[14]This is not necessarily a sign of asymmetric updating, as students are getting different strength of signals in the two groups.

Table 1: Summary Statistics of Test Grades and Expected Grades

	Time Period				
Variable	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$
Grade g_{it}					
Mean (relative to test 1)	0.00	-9.22	-8.04	-11.78	-5.25
Std. Dev.	20.54	22.10	23.04	20.56	24.34
N	1,508	1,399	1,278	1,129	1,155
Expected Grade $\hat{g}_{it}$					
Mean	68.58	69.09	66.93	65.13	63.84
Std. Dev.	14.51	14.52	15.22	15.82	15.79
N	1,333	1,211	1,145	1,043	1,048
Absolute Prediction Error Γ_{it}					
Mean	15.62	19.46	18.28	18.31	16.80
Std. Dev.	12.83	14.72	14.99	14.21	13.66
N	1,276	1,151	1,093	960	1,011

Note: We do not report the average grade as per the instructor's request.
Grades are reported with test 1 as the baseline.
Sample size N changes due to dropping and survey take-up.

we cannot quantify whether students are updating enough given the information without making specific modeling assumptions.

Student Beliefs Regarding Testing Noise. On average, students believe that 37% of their prediction errors, r_{it}, is due to luck and that good luck, e_{it}, raises their grade by 9.03pp on a test. Both of these remain stable throughout the periods. The fact that e_{it} remains stable suggests that students are not inferring from their grades (which are highly correlated) that the testing noise must be low. And the fact that r_{it} remains stable jointly implies students do not believe overall their predictions are getting better. When a student's prediction last period was higher (and lower) than their actual grade, then the average e_{it} is 9.07pp (and 8.55pp), while the average r_{it} is 38.23% (and 34.93%), significant at any p-value. These behaviors align with theoretical models such as Bénabou and Tirole (2004), which suggest that misspecification of information can be motivated by ego.

Belief Change and Testing Noise. One pattern one may find natural is when students get a grade lower than what they predicted; then they should adjust their next prediction downward. We verify this, and Figure 1 shows the change in prediction, $\hat{g}_{it} - \hat{g}_{it-1}$, given last period's prediction error, $g_{it-1} - \hat{g}_{it-1}$. We note that, in general, students are not willing

to lower their prediction by more than 5pp even when their prediction error is close to 40pp. This is not necessarily a sign of non-Bayesian updating. If signals are viewed as very uninformative, a Bayesian could display this pattern. Furthermore, Figure 2 shows that students who tend to make larger and negative prediction errors are precisely those who believe testing to be noisier.

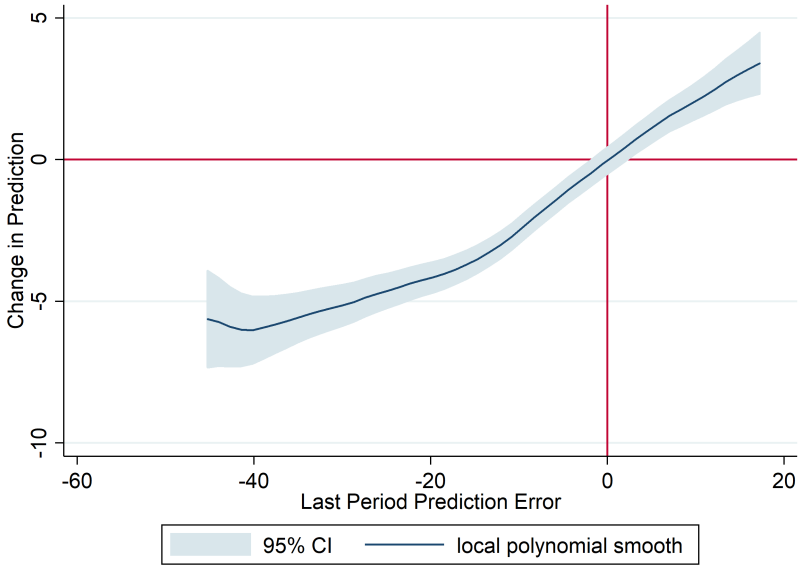


Figure 1: Prediction change and prediction error

Heterogeneity. We find significant heterogeneity in terms of test grades across gender, first-generation status, and international student status. On average, male students score 6pp higher than their female counterparts; domestic students score 2.5pp higher than international students; and non-first-generation students score 5pp higher than first-generation students. All of these differences are statistically significant at any standard p-value. For prediction errors, We find that male students tend to have 0.6pp higher absolute prediction errors (p-the value of 0.104); international and first generations students have absolute prediction errors that are 1.7pp and 2pp higher than their counterparts, respectively (significant at any p-value). Additionally, we find that first-generation and domestic students believe testing to be less noisy, but the differences are not economically significant.[15]

[15]First-generation students believe that 35.7% of their prediction errors stem from noise while non-first-generation student believe it to be 37.1% (statistically significant differences at any p-value). For domestic and international students, we find these numbers to be 35.5% and 38%, respectively (statistically significant differences at any p-value).

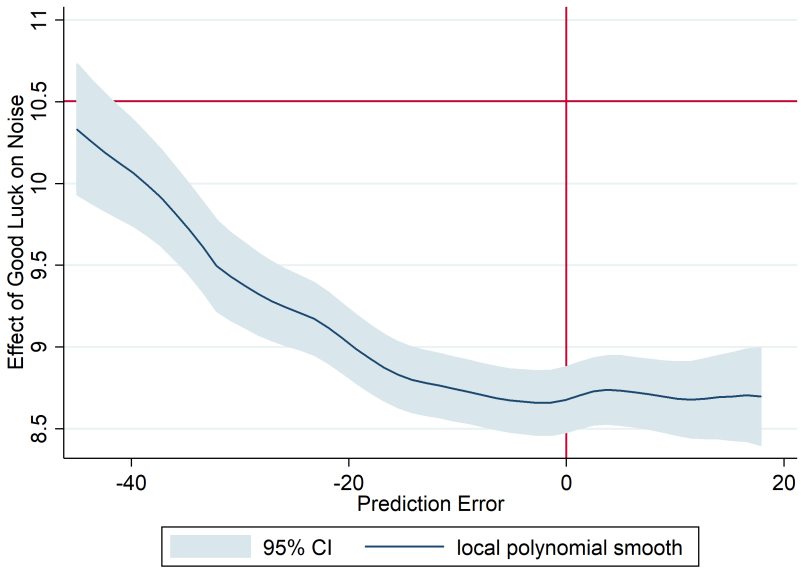


Figure 2: Effect of good luck and prediction error

5 Randomized Control Trial

Treatment. We conduct an RCT before the last test. The goal of the RCT was to exogenously influence students' beliefs regarding the testing noise - without giving them additional personal information regarding their performance. To accomplish this, we leverage the fact that the class did not release any information regarding averages. In particular, treated students first reported their beliefs via $\hat{g}_{i5}, \hat{\tilde{g}}_{i5}, G_{i5}, r_{i5}$ and e_{i5}, then they are shown that the effect of luck is minimal (see Figure 3). In particular, we inform them of the probability, in the previous year, that a student who scored 10pp to 15pp **above** and *below* the average across midterms also scores 5pp **below** and *above* the average on the final exam, respectively. Furthermore, we show the total percentage for all of the classes - to highlight that these grades are truly highly correlated. Because students do not know class averages, it is difficult to know if they lie in these grade bins exactly. Hence, this provides no information to help predict their future grade, only that they should use their previous grades more. Students are asked to consider this information carefully and are locked on this page for two minutes before they can proceed. Additionally, students are shown their previous grades, g_{it}, and their expected class average, $\hat{\tilde{g}}_{it}$. We then collected belief variables again, and we denote by $\hat{g}^T_{i5}, \hat{\tilde{g}}^T_{i5}, G^T_{i5}, r^T_{i5}$ and e^T_{i5}, the treated version of these variables.

Group Assignment. We note that spillover effects are a serious threat to such designs,

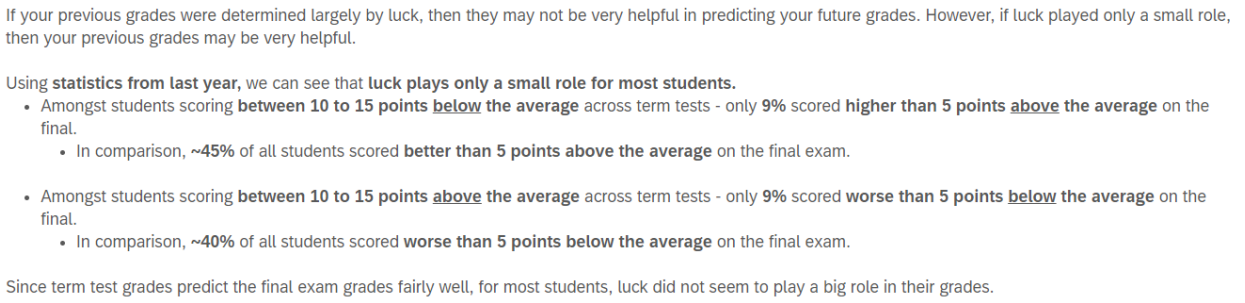


Figure 3: Treatment Example

and in general, their potential existence implies that the treatment will be underestimated. We take many measures to carefully minimize spillover. One approach would be to run the survey without treatment first. This obtains an uncontaminated control group. However, timing matters, and students who complete the survey early differ from those who complete it late. To overcome this, we perform a staggered implementation. First, students either attempted to complete the survey in the first three hours or did not. This classifies them as *Early* or *Late*. Among the early group, half of the students proceeded with the survey without treatment, while the other half were asked to perform the survey later. This allowed us to ensure that the early control group was uncontaminated from potential spillover effects, and comparing them with the early treatment group allowed us to rule out heterogeneity due to time. Additionally, each treated student first reports their beliefs without being treated and reports another belief post-treatment.

Table 2 documents some student characteristics. First, there are no significant differences between the control and treatment groups for observable characteristics such as past grades and gender ratio. Second, between the treatment and control groups, their reported beliefs on their expected grade and belief in the effect of noise do not differ before the treatment. Similarly, the average past prediction error does not differ between the treatment and control, fixing the timing. Finally, there is a noticeable difference between the early and late groups regarding actual grades, highlighting that timing matters.

There are two natural ways to estimate the effect of our treatment. We call these *within* and *between* group effects. The between-group effect consists of comparing the control group with the treatment group. The within-group effect consists of comparing the pre-treatment reports to the post-treatment reports within the treatment group. Throughout the analysis, we showcase both effects.

Table 2: Group Statistics

	Treatment Early		Control Early		Treatment Late		Control Late	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Test 1 grade g_{i1}	0.00	(1.33)	-1.96	(1.36)	-4.47	(0.94)	-5.56	(0.92)
Test 2 grade g_{i2}	0.00	(1.53)	0.04	(1.66)	-3.13	(1.13)	-4.01	(1.07)
Test 3 grade g_{i3}	0.00	(1.66)	0.53	(1.74)	-3.83	(1.18)	-5.00	(1.18)
Test 4 grade g_{i4}	0.00	(1.57)	-0.08	(1.77)	-6.61	(1.10)	-6.85	(1.04)
Male Prop.	0.61	(0.04)	0.72	(0.04)	0.59	(0.03)	0.61	(0.03)
First-gen Prop.	0.14	(0.03)	0.21	(0.03)	0.21	(0.02)	0.18	(0.02)
International Prop.	0.53	(0.04)	0.48	(0.04)	0.54	(0.03)	0.53	(0.03)
Pre-treatment expect grade $\hat{g}_{i5}$	65.35	(1.30)	66.35	(1.38)	63.53	(0.76)	62.56	(0.83)
Pre-treatment effect of luck $\hat{e}_{i5}$	9.55	(0.44)	9.14	(0.39)	8.84	(0.24)	8.88	(0.24)
Average past prediction error	-9.52	(0.79)	-10.19	(0.88)	-10.42	(0.63)	-11.10	(0.63)
Sample Size	146		153		364		385	

Note: Test grades are relative to the Treatment Early group.

Treatment on Belief of Noise. Recall our treatment seeks to show students that testing noise is low. We first show the average treatment effect on the effect of luck, e_{it}, and the proportion of prediction mistakes due to noise, r_{it}. For the control group, e_{it} and r_{it} are, on average, 8.96pp and 37.13%, respectively. For the treatment group's pre-treatment report, we find almost identical values of 9.04pp and 38.1%. For the post-treatment reports, these are lower, with values of 7.21pp and 28.53%, respectively. As they are virtually identical, we pool the control and pre-treatment reports in our figures. Figures 4 and 5 plot the confidence intervals, showing that the treatment successfully reduced the student's perception of testing noise. We report the overall average treatment effect in the figure, but it turns out that both the timing effect and spillover are minimal. The appendix shows the other estimations that are similar. In Appendix A, we show the treatment effect by early/late status, gender, international status, and first-gen status. We find the same effect across these subgroups.

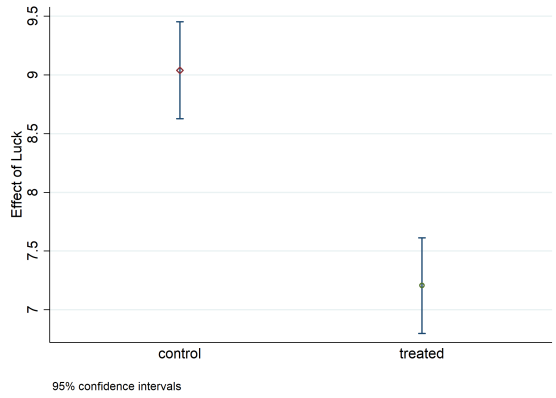


Figure 4: Belief Regarding Effect of Good Luck on Grade

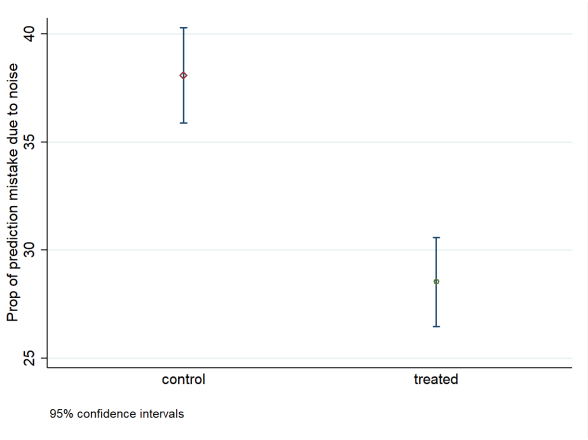


Figure 5: Belief Regarding Effect of Luck on Prediction Mistakes

Placebo Tests. We might worry that predictions differ because students study harder after the treatment. That is why we only ran the experiment three days ahead of the final exam. We also find that the treated students do not score differently from the control students.[16] This suggests that our implementation did not induce behavioral changes that would confound with effects from our updating channel. Additionally, mechanical changes induced by the treatment, such as demand effects, could be in play. We check whether the treatment has impacted other variables. We first check that the treatment did not impact a student's belief regarding the average grade. This is important as we did want students to infer about the class average to make better predictions - only to infer that grades are correlated. We find that treated students do not report a different prediction of

[16]The difference between the groups is 0.05pp with a p-value of 0.73. We cannot reject the null that the groups have the same mean.

the average than control students and that they do not change their predictions of average post-treatment. Therefore, we are confident that the treatment did not induce mechanical changes in their beliefs.

Effect on Responsiveness to Information. We investigate whether treated students are more responsive to information. This prompts the natural question: what is responsiveness to information? This, in turn, prompts us to ask: what are response measures and information measures? The right response measure is the unit the student is adjusting. And the right information measure is what a student is taking into account when adjusting the response. A particular response measure is natural in our empirical setting: expected grade. As our treatment itself asks them to consider the effect of luck on grades, it is natural that students potentially respond via their grade predictions. Two measures of information stand out as plausible candidates: past grades relative to the average and past prediction errors. Our treatment itself divulges information via the class average, and we show students their previous grades and expected class averages. Therefore, we find it likely that students may consider their past grades relative to their expected averages to be informative. Similarly, past prediction errors are natural for students to recall, as we highlight in the treatment that luck may impact their prediction error. A priori, we do not know how and whether students actually think through these measures. Therefore, we empirically examine whether these response and information measures are related and whether our treatment impacts these relationships.

Denote by $i \in \mathcal{T}$ if student i is in the treatment group. We consider the response measure to be how much a student adjusts her predicted grade from previous periods: $\Delta_i = \hat{g}_{i5}^T - \frac{1}{4}\sum_{t<5} \hat{g}_{it}$ if $i \in \mathcal{T}$ and $\Delta_i = \hat{g}_{i5} - \frac{1}{4}\sum_{t<5} \hat{g}_{it}$ otherwise. We denote by $\Phi_i = \frac{1}{4}\sum_{t=1}^{4}(g_{it} - \hat{\hat{g}}_{it})$ their past deviation from the average. And we denote the average past prediction errors by $\Psi_i = \frac{1}{4}\sum_{t=1}^{4}(g_{it} - \hat{\hat{g}}_{it})$.

To measure the effect of the treatment on responsiveness to information, we consider the following regressions

$$\Delta_i = \beta_0 + \beta_1 \Psi_i + \beta_2 \mathbb{1}_{\{i\in\mathcal{T}\}}\Psi_i + \beta_3 \mathbb{1}_{\{i\in\mathcal{T}\}} + \beta_4 \sum_{t<5} \frac{1}{4} g_{it} + \beta_5 g_{i5} + \beta_5 X_i + \xi_i \quad (1),$$

$$\Delta_i = \beta_0 + \beta_1 \Phi_i + \beta_2 \mathbb{1}_{\{i\in\mathcal{T}\}}\Phi_i + \beta_3 \mathbb{1}_{\{i\in\mathcal{T}\}} + \beta_4 \sum_{t<5} \frac{1}{4} g_{it} + \beta_5 g_{i5} + \beta_5 X_i + \xi_i \quad (2).$$

We capture the responsiveness to information of the control via β_1 and the treatment effect on responsiveness via β_2. Additionally, we expect the treatment effect to depend

on actual past grades and not only the prediction effect or deviation from the average. Therefore we control for the average past grades. Additionally, we control for the actual final grade to capture individual variations in the final period unaccounted for from previous periods. Finally, we include controls such as gender, international status, age, and first-generation status in X_i.

Table 3: Responsiveness to Information

	Ψ_i Prediction Error		Φ_i Average Deviation	
	Between	Within	Between	Within
β_1 control responsiveness	0.24^{***}	0.24^{***}	0.04	0.07
	(0.07)	(0.06)	(0.07)	(0.06)
β_2 treatment effect on responsiveness	0.16^{***}	0.16^{**}	0.10^{**}	0.09^{*}
	(0.07)	(0.07)	(0.05)	(0.05)
N	796	401×2	886	446×2

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

Note: Robust standard errors in parentheses

The regression results show that the treatment has a significant effect on responsiveness. On average, the treated student adjusts her prediction at least by an extra 0.15pp upward for every 1pp she scores above her prediction on previous tests. In the case of scoring above the expected average, we see that the control student does not respond to this measure.[17] However, as our treatment makes it salient that this measure contains information, we see treated students react to it and, on average, raise their prediction by an additional 0.10pp upwards for every 1pp she scores above her expected average. These results are not dependent on controls; see Appendix for regression without controls.

Effect on Prediction Quality. We then investigate whether this increased responsiveness to information translates to better beliefs. The average absolute prediction error Γ_{i5} for the control group was 17.35pp, and 15.31pp for the treated group. This difference is statistically significant with a *p*-value of 0.015. Similarly, at a within-student level, the pre-treated absolute prediction error was 16.22pp and dropped to 15.31pp for the treated group; this difference, too, is statistically significant at a *p*-value of 0.06. To quantify the significance, we use the Bayesian prediction mistakes (derived in the next section) and

[17]This makes sense in this institutional setting as they are never told the actual average.

compute $\Lambda^{treat} = 1 - \frac{\frac{1}{|\mathcal{T}|}\sum_{i\in\mathcal{T}}\Gamma_{i5}^{treat}-\Gamma_{i5}^{B}}{\frac{1}{|\mathcal{T}^C|}\sum_{i\notin\mathcal{T}}\Gamma_{i5}^{control}-\Gamma_{i5}^{B}} = 1 - \frac{15.31-10.81}{17.35-10.81} \approx 0.32$ for the between-student effect. Similarly, we obtain 0.18 for the within-student effect. Our findings suggest that our treatment, which exogenously shocks belief in testing noise, reduced around 18% and 32% of prediction mistakes relative to the Bayesian benchmark in the within and between-student cases, respectively.

6 Structural Estimation

In this section, we analyze students' updating patterns. We recover measures of non-Bayesian bias as well as misspecification.

To analyze the updating patterns of a student, it is not sufficient to observe merely how predictions change. Rather, it is necessary to estimate the actual statistical relationship between grades. Only given this statistical informativeness can we estimate how a Bayesian would update. Similarly, to study the effect of misspecification, it is not sufficient to estimate the objective statistical informativeness. Additionally, we need to estimate the *perceived* informativeness. These factors lead us to explicitly model the relationship between grades to recover an objective and a subjective measure of informativeness.

To illustrate the issue at hand, consider a dataset containing the expected belief at a time t, $\hat{g}_{it}$, and an updated expected belief, $\hat{g}_{it+1}$, at a future time. However, this type of dataset does not allow the researcher to discuss the updating process as it 1) does not document belief in terms of distribution and 2) does not document the signal or the signal strength. In experimental settings, the researcher elicits G_{it} and G_{it+1}, which are prior and posterior probability distributions. The researcher knows the signal's informativeness, tells it to the subject, and can estimate how much the subject's belief differs from a Bayesian one. However, in our setting, we do not know the objective informativeness of test scores, and more importantly, the students may have a different perception of this informativeness. Therefore, our first contribution is to collect distributional belief data G_{it}s empirically. Then, to estimate non-Bayesian updating and misspecification, we leverage our rich dataset to build a structural model. This allows us to 1) estimate the objective informativeness of test grades and 2) estimate the students' perceived informativeness of the test grades via estimating their belief in the model's parameters.

Model. We posit that the following general functional form can capture the grade-generating function:

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

In particular, we posit that three sets of variables determine the grade. First, θ_i and η_{it} are measures of the student's skill. These are the objects the student ultimately faces uncertainty about. We only assume that η_{it} may be correlated over time. In particular, θ_i is an underlying fixed effect that remains constant throughout the course. Additionally, η_{it} represents the student's shock to skill in different periods, such as having more time to study or finding this test's material more familiar. Second, there is a class-wide test fixed effect, which we assume is captured by the average test grade $\bar{g}_t$. In all of our models, replacing $\bar{g}_t$ with a test fixed effect makes no difference, but we take this interpretation to calibrate the student's perceived model. Finally, ϵ_{it} denotes exogenous testing noise, which neither the student nor the researcher can explain, and our main interest is in the variance σ^2 of this noise and how student perception of this variance impacts updating. We assume it is normally distributed and give some supporting evidence of this assumption in the appendix.

Estimation of Objective Model. As we wish to understand the true informativeness of grades, given this model, we estimate σ.

A natural first stab at estimating σ is via rewriting (2) as $g_{it} - \bar{g}_t = \theta_i + \eta_{it} + \epsilon_{it}$. This groups all observables on the left-hand side, and we can estimate a simple fixed effects model. The residual of the fixed effects model will be $\eta_{it} + \epsilon_{it}$; therefore, the standard deviation of these residuals will be an upper bound for σ. Doing this exercise yields 9.7 as an upper bound for σ. However, as this includes individual shocks, it may severely overestimate the actual testing noise. It turns out we can do better given that η_{it}s are autocorrelated. We assume that $\eta_{it} = \beta_1\eta_{it-1} + \beta_2\eta_{it-2} + \nu_{it}$ follows an $AR(2)$ process.[18] This assumption gives us a clean identification of σ and disentangles ϵ's variance from η's variance.[19]

Proposition 1. *If η_{it} is a stationary AR(2) process, then σ is identified.*[20]

We go over the result in Appendix B. The intuition is that by observing the variation of the residual $\eta_{it}+\epsilon_{it}$ over time, we can estimate the variance of η (given the $AR(2)$ functional form) and separate it from σ. We recover $\sigma = 3.75$. To give the reader a general sense of the estimate, the effect of testing noise, $\mathbb{E}[\epsilon_{it} \mid \epsilon_{it} \geq 0]$, would be equal to 2.95. Therefore, the expected change to your grade due to luck is around ± 2.95, whereas the average student believes it to be around ± 9.03.

[18] We have 5 time periods, $AR(1)$ and $AR(2)$ offer similar results.

[19] We thank Xincheng Qiu and Guanbing Hong for pointing us in the right direction for this result.

[20] As we only have five periods total, we look at $AR(2)$, but if we had more periods, then this result can be extended for $AR(p)$.

We note that the objective model is itself misspecified, so the recovered testing noise would be an upper bound. However, we are confident that the model nevertheless captures most of the variations and that η is correlated over time. First, running the naive fixed effect regression (2) already returns an R^2 of 0.8. Second, we note in our estimation of β_1 and β_2 that they are statistically significantly different from 0 at any p-value. Third, adding lagged residuals in (2) significantly raises the R^2 to 0.87.

Estimation of Subjective Model. To retrieve a student's subjective mental model to quantify misspecification, we must retrieve, for each period, their uncertainty in their skill, $\hat{\theta}_i + \hat{\eta_{it}}$, their belief in the noisiness of testing $\hat{\sigma}$, and their belief in the class average $\hat{\bar{g}}_t$. Their belief in the class average is retrieved directly from our survey. To recover the first two items, we collected $p_{it}(\hat{g_{it}} \geq X)$, the quantile functions at values of $X \in \{50, 60, 70, 85\}$. Using the cubic spline method of Bellemare et al. (2016), we recover a distribution $F_{it}(\hat{g}_{it})$ which captures the student's belief over her grades at time t. Their belief about the distribution of their grade encapsulates both their uncertainty regarding skill and their belief regarding the noisiness of the testing. From F_{it}, we compute $\mathrm{Var}(\hat{g}_{it})$, the student's perceived variance in her grade. In particular, in our model, $\mathrm{Var}(\hat{g}_{it}) = \mathrm{Var}(\theta_i + \eta_{it}) + \hat{\sigma}^2_{it}$. We ask them directly about the proportion of mistakes in predicting their grades due to exogenous luck, r_{it}. As the expected squared error of the prediction is the variance, we recover $r_{it} = \frac{\hat{\sigma}^2_{it}}{\mathrm{Var}(\hat{g_{it}})}$. This, along with the normality assumption, allows us to recover a belief distribution over $\hat{\theta}_i + \hat{\eta_{it}}$.[21] In particular, we recover the student's belief regarding her skill, with pdf h_{it}, via

$$\begin{aligned} h_{it}(\theta_i + \eta_{it} = x) &= p(\hat{g}_{it} = x + \hat{\bar{g}}_{it} + \epsilon_{it}\ ;\ \hat{\sigma}_{it}), \\ &= \int f_{it}(g) p(\epsilon_{it} = g - x - \hat{\bar{g}}_{it};\ \hat{\sigma}_{it}) dg, \\ &= \int f_{it}(g) \frac{\exp[-\frac{1}{2}(\frac{g - \hat{\bar{g}}_{it} - x}{\hat{\sigma}_{it}})^2]}{\hat{\sigma}_{it}\sqrt{2\pi}} dg. \end{aligned}$$

To obtain h_{it}, we estimated the integral above with a discrete sum of 200 values of g drawn from F_{it} and estimated for every value of $\theta_i + \eta_{it}$ from -100 to 100 with a stepsize of $\frac{1}{2}$. We find that students remain overly optimistic about their performance apart from an initial adjustment.

Hence, the average student continues to expect to score about 7 points higher than the

[21] Note this assumes that the students correctly believe the noise to be normally distributed, and therefore, our model will *underestimate* the effect of misspecification. Our analysis, therefore, always provides a lower bound for the effect of misspecification.

Table 4: Summary Statistics of Estimated Subjective Variables

	Time Period				
Variable	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$
$\mathbb{E}[\theta_i + \eta_{it}]$	10.91	7.99	8.14	7.37	7.19
$\hat{\sigma}_{it}$	10.67	10.99	11.55	11.54	11.33

average. Similarly, we find students believe σ to be around 11, which implies an expected effect of noise, $\mathbb{E}[\epsilon \mid \epsilon \geq 0, \sigma = 11] = 8.78$. We also elicited this measure by asking students about their beliefs regarding the impact of good luck. Our directly elicited measure e_{it} has a similar value of 9.03. This gives us some confidence that these elicited values are internally consistent. We find that only 10% of estimated σ across students and the five tests are below the objectively estimated $\sigma = 3.7$, so the overestimation of the testing noise is widespread.

Updating and Misspecification. We now consider misspecification and non-Bayesian updating and decompose their effects on the prediction error. To illustrate our approach, define first $\Gamma_{it} = |\hat{g}_{it} - g_{it}|$ to be the absolute prediction error. Similarly, we define $\Gamma^B_{it} = |\hat{g}^B_{it} - g_{it}|$ and $\Gamma^M_{it} = |\hat{g}^M_{it} - g_{it}|$ to be the absolute prediction error of a Bayesian with the correct belief $\sigma = 3.75$ and that of a Bayesian with misspecified belief $\sigma = \hat{\sigma}_{it}$. Therefore, $\sum_i \Gamma_{it} - \Gamma^B_{it}$ represents the reduction in the absolute prediction error from Bayesian updating with correct specifications at time t, and $\sum_i \Gamma^M_{it} - \Gamma^B_{it}$ represents the reduction in the absolute prediction error from correcting misspecification for a Bayesian at time t. Therefore, $\Lambda_t = \frac{\sum_i \Gamma^M_{it} - \Gamma^B_{it}}{\sum_i \Gamma_{it} - \Gamma^B_{it}}$ is our measure of the proportion of prediction error due to misspecification in the dataset at time t. If $\Lambda_t = 1$, then this means that all the reduction in mistakes is due to misspecification, as adding Bayesian updating alone does not reduce the aggregate mistake. By focusing on the prediction errors, we can consider a tangible effect of non-Bayesian updating and misspecification without committing to any particular model of non-Bayesian updating.

With the above methodology laid out, we now show how Γ^M_{it} and Γ^B_{it} are recovered. Note first that recovering these is equivalent to recovering the posterior distribution of $\hat{g}^M_{it}$ and $\hat{g}^B_{it}$. We have already identified $\hat{\sigma}_{it}, \hat{\bar{g}}_{it}$, σ, and $\bar{g}_t$ for the misspecified and correctly specified models. So the uncertainty in updating depends on values of $\theta_i + \eta_{it}$. Students observe η_{it}; therefore, learning through past grade is only through learning about the distribution of θ_i. However, we do not observe η_{it} as the researcher. We take a minimalistic and misspecified approach and assume it is constant over time. This implies that our model

is inherently misspecified. This implies the actually correctly specified Bayesian prediction error should be lower than our estimates, and the same is true for our misspecified model's prediction error.

To simplify notations as we proceed, denote by $\kappa_i = \theta_i + \eta_{it}$, note we recover at each period the subject's uncertainty regarding κ_i, which we denote by a pdf h_{it}. This implies that to recover the posterior distribution of $\hat{g}_{it}^M$ and $\hat{g}_{it}^B$, we need to estimate the posterior distribution of κ_i given the last period's grade g_{it-1}. We derive the posteriors of κ_i in the misspecified model as

$$\begin{aligned} q_{it}^M(\kappa_i = x | g_{it-1}, \hat{\sigma}_{it}, \hat{\bar{g}}_{it}) &= \frac{h_{it-1}(x) p(g_{it-1} = \hat{\bar{g}}_{it-1} + x + \epsilon_{it-1} \mid \hat{\sigma}_{it})}{f_{it-1}(g_{it-1})}, \\ &= \frac{h_{it-1}(x)}{f_{it-1}(g_{it-1})} \frac{\exp[-\frac{1}{2}(\frac{g_{it-1} - \hat{\bar{g}}_{it-1} - x}{\hat{\sigma}_{it}})^2]}{\hat{\sigma}_{it}\sqrt{2\pi}}. \end{aligned}$$

And the correctly specified model has the following posteriors:

$$\begin{aligned} q_{it}^B(\kappa_i = x | g_{it-1}, \sigma, \bar{g}_t) &= \frac{h_{it-1}(x) p(g_{it-1} = \bar{g}_{t-1} + x + \epsilon_{it-1} \mid \sigma)}{f_{it-1}(g_{it-1})}, \\ &= \frac{h_{it-1}(x)}{f_{it-1}(g_{it-1})} \frac{\exp[-\frac{1}{2}(\frac{g_{it-1} - \bar{g}_{it-1} - x}{\sigma})^2]}{\sigma\sqrt{2\pi}}. \end{aligned}$$

Given this posterior regarding skill κ_i, we can derive the Bayesian posterior distribution regarding grades of $\hat{g}_{it}^M$ and $\hat{g}_{it}^B$. For the misspecified model, we obtain that

$$\begin{aligned} p_{it}^M(g_{it} | g_{it-1}) &= \int q_{it}^M(\kappa | g_{it-1}, \hat{\sigma}_{it}, \bar{g}_t) p(\epsilon_{it} = g_{it} - \kappa - \hat{\bar{g}}_{it}) d\kappa, \\ &= \int q_{it}^M(\kappa | g_{it-1}, \hat{\sigma}_{it}, \bar{g}_t) \frac{\exp[-\frac{1}{2}(\frac{g_{it} - \hat{\bar{g}}_{it} - \kappa}{\hat{\sigma}_{it}})^2]}{\hat{\sigma}_{it}\sqrt{2\pi}} d\kappa. \end{aligned}$$

For the correctly specified model, we have instead

$$\begin{aligned} p_{it}^B(g_{it} | g_{it-1}) &= \int q_{it}^B(\kappa | g_{it-1}, \sigma, \bar{g}_t) p(\epsilon_{it} = g_{it} - \kappa - \bar{g}_t) d\kappa, \\ &= \int q_{it}^B(\kappa | g_{it-1}, \sigma, \bar{g}_t) \frac{\exp[-\frac{1}{2}(\frac{g_{it} - \bar{g}_t - \kappa}{\sigma})^2]}{\sigma\sqrt{2\pi}} d\kappa. \end{aligned}$$

Once we obtain the posterior distribution over grades, we take the expected grade for each of these two models as the prediction to compute Γ_{it}^M and Γ_{it}^B. As a sanity check, we show in the table below that $\sum_i \Gamma_{it} > \sum_i \Gamma_{it}^M > \sum_i \Gamma_{it}^B$ holds on the aggregate, but of

course, given the testing noise is random, it needs not hold for each individual.[22] We can reject at any statistical significance the equality between any pairs of $\Gamma_{it}, \Gamma_{it}^M$, and Γ_{it}^B for any of these periods.

Table 5: Prediction Errors - Actual, Misspecified and Bayesian

	Test2	Test3	Test4	Test5
Γ_{it}	19.3	17.5	17.9	16.2
Γ_{it}^M	17.0	12.6	14.0	13.9
Γ_{it}^B	11.4	11.0	12.4	11.0
Λ_t	0.71	0.26	0.29	0.55

These gives us values of Λ_t equal to $0.71, 0.26, 0.29,$ and 0.55 respectively. We see that misspecification has a greater impact on using the first test to make predictions and for predicting the final exam. Throughout this paper, our methods provide a lower bound for the effect of misspecification on updating, and here, we conclude that at least 25% of prediction mistakes stem from misspecification concerns.

7 Summary

We showed that misspecification plays a key role in the failure of Bayesian updating in a common and relevant empirical setting. This has important policy consequences as we show that simple interventions to correct misspecification can significantly improve beliefs. While one can never be sure of the external validity of our findings, we are confident the present setting was not special in this phenomenon. Additionally, all of our results suggest only a lower bound on the effect of misspecification in this setting.

Whereas a large literature classifies failure of Bayesian updating as due to errors or mistakes in the updating process, we highlight it may instead be caused by mistakes in the perception of the information. The present work presents evidence that misspecification exists by collecting a unique rich dataset which additionally allows us to estimate its impact. Additionally, we further support this hypothesis with causal evidence provided by our RCT.

[22]The absolute prediction error is slightly different from our summary statistics table as we report here only for subjects for whom we can compute Γ_{it}^M, which requires completion of both surveys at time t and $t-1$ as well as having taken tests at both times.

We provide descriptive evidence that prediction error is increasing in measures of misspecification; implement a randomized experiment that generates exogenous shocks to misspecification and find significant effects on prediction error and beliefs; and estimate a structural model that quantifies the magnitude of this channel vis-a-vis other non-misspecification related channels. Both our reduced form and structural results suggest a conservative lower bound of 25% of prediction errors are due to misspecification.

References

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

Barron, K. (2021). Belief updating: does the 'good-news, bad-news' asymmetry extend to purely financial domains? *Experimental Economics 24*, 31–58.

Bellemare, C., L. Bissonnette, and S. Kröger (2016). Simulating power of economic experiments: the powerbbk package. *Journal of the Economic Science Association 2*, 157–168.

Bénabou, R. and J. Tirole (2004). Willpower and personal rules. *Journal of Political Economy 112*(4), 848–886.

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, J. A. and D. N. Hauser (2023). *The Behavioral Foundations of Model Misspecification: A Decomposition*. 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.

Castillo, M. and S. Youn (2023). When biased beliefs lead to optimal action: An experimental study. Working Paper.

Chiara, A. and S. Florian H. (2024). Weighting competing models.

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

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.

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

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 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.

Stinebrickner, R. and T. R. Stinebrickner (2014). 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.

Appendix

A Randomized Control Trial

A.1 Heterogeneity Treatment Effect of Belief about Effect of Luck

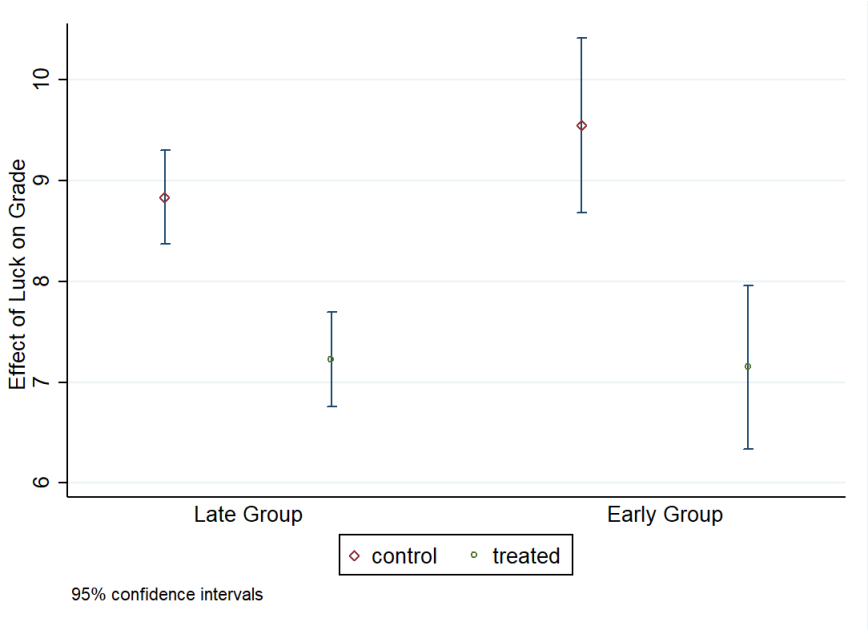


Figure 6: Belief Regarding Effect of Good Luck on Grade by Timing

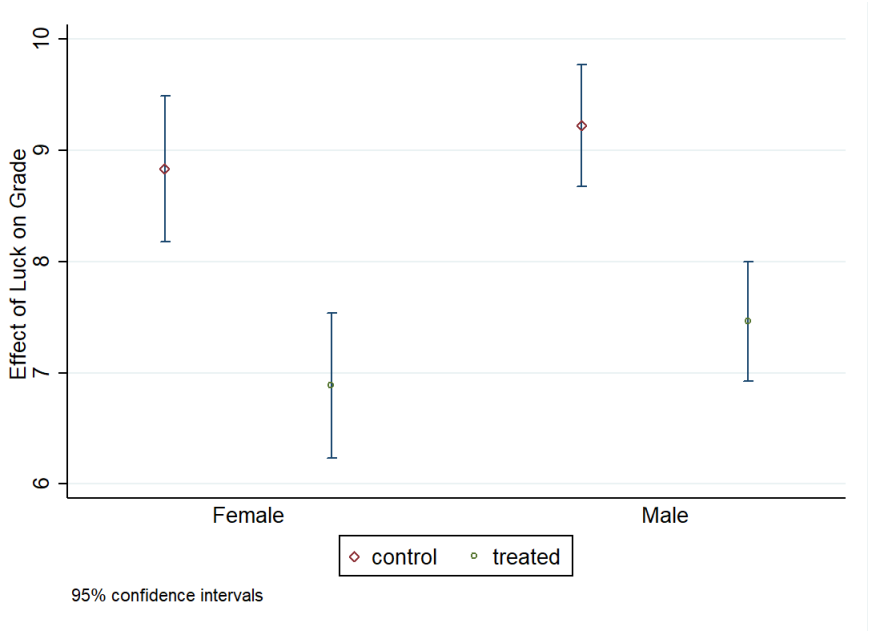


Figure 7: Belief Regarding Effect of Good Luck on Grade by Gender

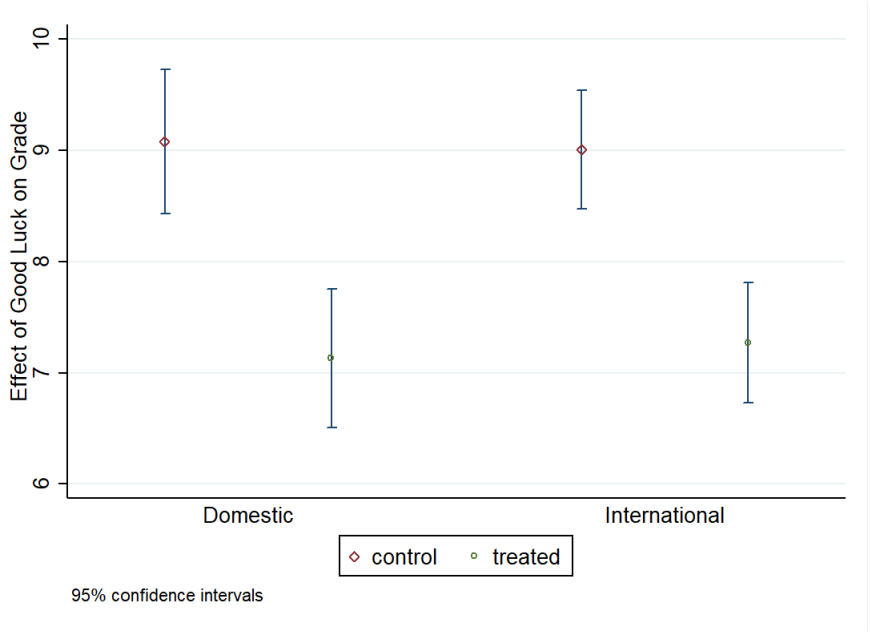


Figure 8: Belief Regarding Effect of Good Luck on Grade by International Status

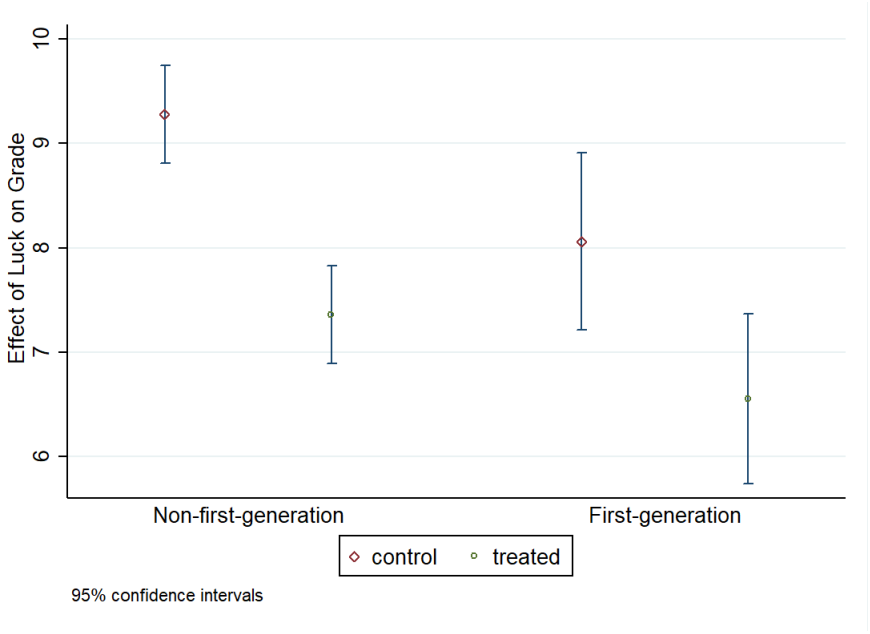


Figure 9: Belief Regarding Effect of Good Luck on Grade by First-Gen Status

A.2 Belief Regarding Effect of Luck on Prediction Mistakes

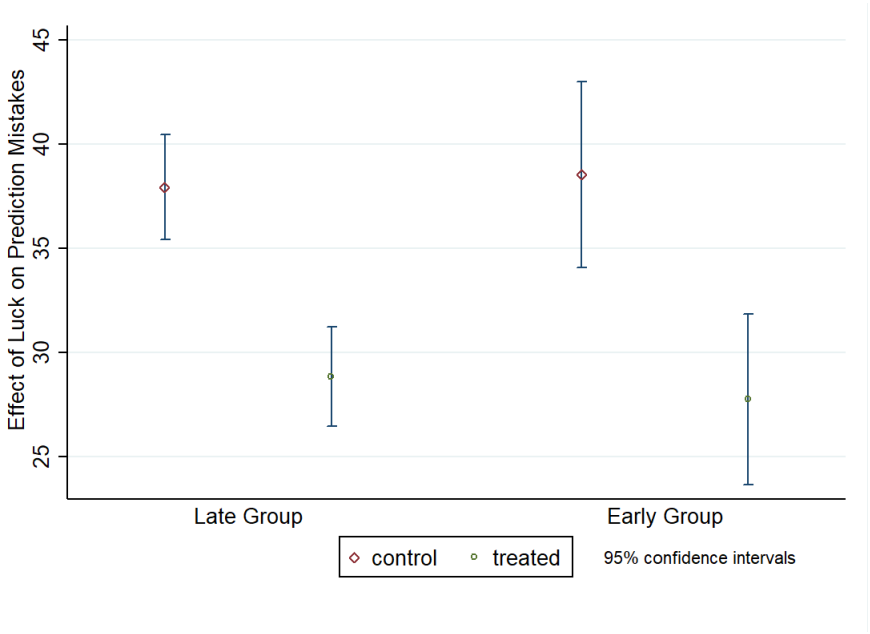


Figure 10: Belief Regarding Effect of Luck on Prediction Mistakes by Early Status

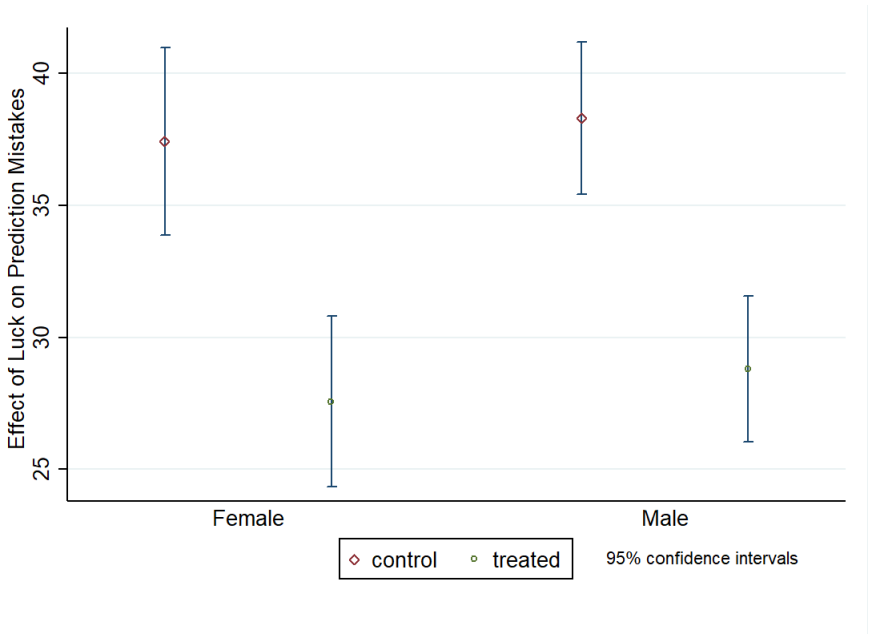


Figure 11: Belief Regarding Effect of Luck on Prediction Mistakes by Gender

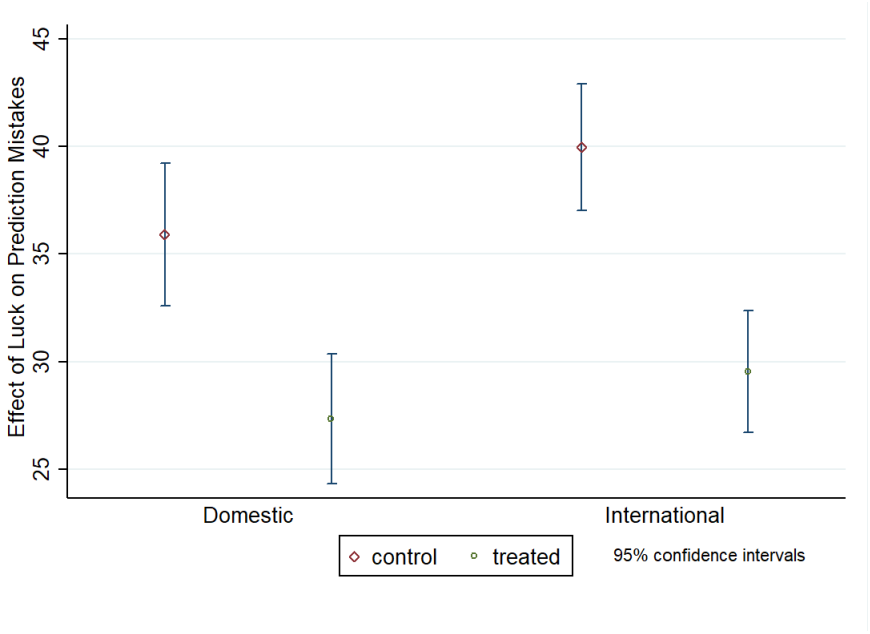


Figure 12: Belief Regarding Effect of Luck on Prediction Mistakes by International Status

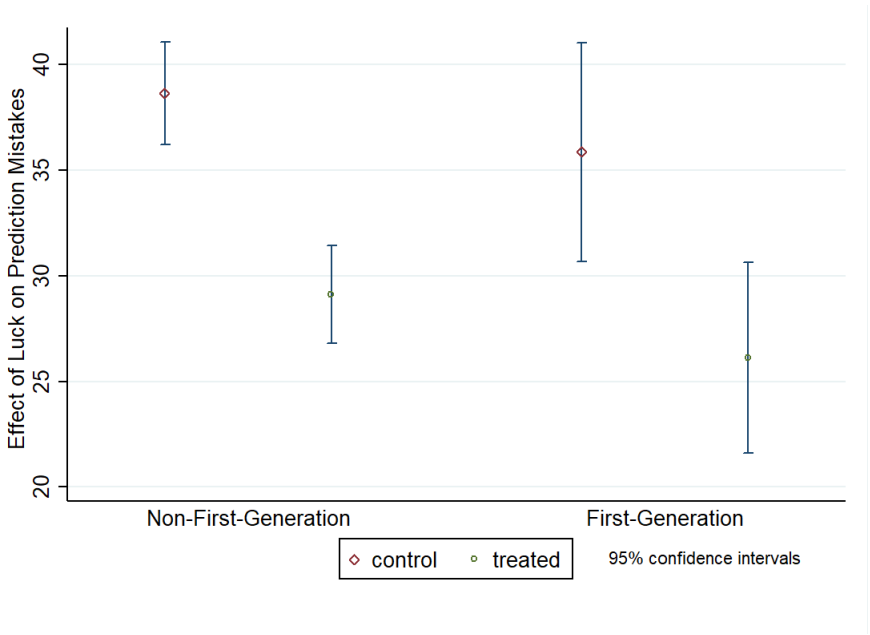


Figure 13: Belief Regarding Effect of Luck on Prediction Mistakes by First-Gen Status

A.3 Responsiveness to Information without Controls

We rerun regression (1) and (2) with only the β_0 to β_4 terms.

Table 6: Responsiveness to Information without Controls

	Ψ_i Prediction Error		Φ_i Average Deviation	
	Between	Within	Between	Within
β_1 control responsiveness	0.20^{***}	0.18^{***}	0.06	0.07^{*}
	(0.05)	(0.05)	(0.05)	(0.04)
β_2 treatment effect on responsiveness	0.14^{*}	0.16^{**}	0.10^{**}	0.09^{*}
	(0.07)	(0.08)	(0.05)	(0.05)
N	796	401×2	886	443×2

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

Note: Robust standard errors in parentheses

B Structural Estimation

B.1 Objective Model

B.1.1 Proposition 1

Given that η_{it}s are autocorrelated, we assume that η_{it} follow an AR(p) process. This implies that $\eta_{it} = \sum_{k=1}^{p} \beta_k \eta_{it-k} + \nu_{it}$ (3) where ν_{it} captures a measure of the individual's time-dependent shock to η_{it} which previous periods' values cannot linearly capture. We take $p = 2$ as we have 5 time periods, this then gives us 5 unknowns $(\sigma, \sigma_\eta, \sigma_\nu, \beta_1, \beta_2)$.

We obtain from the AR(p) formula that

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

as our first identifying equation.

We can recover $Var(\eta_{it} + \epsilon_{it}) = 9.5^2$ directly from our fixed effect regression where we estimate $g_{it} + \bar{g}_t = \theta_i + \eta_{it} + \epsilon_{it}$. The residual of the regression is $\eta_{it} + \epsilon_{it}$. This gives us our second identifying equation that $\sigma^2 + \sigma_\eta^2 = 9.5^2$.

Additionally, we can take the first difference of the residuals, $\eta_{it} + \epsilon_{it} - \eta_{it-1} - \epsilon_{it-1}$. The variance of this first difference is

$$Var(\eta_{it} + \epsilon_{it} - \eta_{it-1} - \epsilon_{it-1}) = \frac{2 - 2\beta_2 - 2\beta_1}{1-\beta_2}\sigma_\eta^2 + 2\sigma^2.$$

We estimate the variance of the first difference to be 14.6^2, giving us another identifying equation. We note this equality holds *only if* η is stationary, meaning $|\beta_2| < 1$, $\beta_1 + \beta_2 < 1$ and $\beta_2 - \beta_1 < 1$, which we have to verify once these terms are recovered.

Therefore, we currently have three equations and 5 unknowns: $(\sigma, \sigma_\eta, \sigma_\nu, \beta_1, \beta_2)$. One way to recover the βs would be to estimate the AR(2) process. However, we do not observe η_{it}, only $\eta_{it} + \epsilon_{it}$. Therefore we can only run the following equation,

$$\eta_{it} + \epsilon_{it} = \hat{\beta}(\eta_{it-1} + \epsilon_{it-1}) + \xi_{it}, \quad (4)$$

where ξ_{it} is a noise (note it is different from ν_{it}. Therefore, we can only recover a biased estimate of β_k. It can be however shown that asymptotically $\text{p-lim}(\hat{\beta}_k) = \beta_k \frac{\sigma_\eta^2}{\sigma_\eta^2 + \sigma^2}$. we recover that $(\hat{\beta}_1, \hat{\beta}_2) = (-0.24. - 0.32)$. The coefficient is statistically significantly different from 0 at any p-value with standard errors of 0.018 and 0.019, respectively. This gives us the final two equations.

We recover $(\sigma, \sigma_\eta, \sigma_\nu, \beta_1, \beta_2) = (3.75, 8.72, 7.9, -0.28, -0.37)$. Note that $\beta_1 = -0.28$ and $\beta_2 = -0.37$, so the recovered process is indeed stationary. Additionally, if instead we had assumed an AR(2) process we would recover $(\sigma, \sigma_\eta, \sigma_\nu, \beta) = (6.78, 6.79, 6.45, -0.31)$.

B.1.2 Residual Normality

Below we plot the residuals $\eta_{it} + \epsilon_{it}$ from a naive regression with just individual fixed effects and the demeaned grade on the left handside.

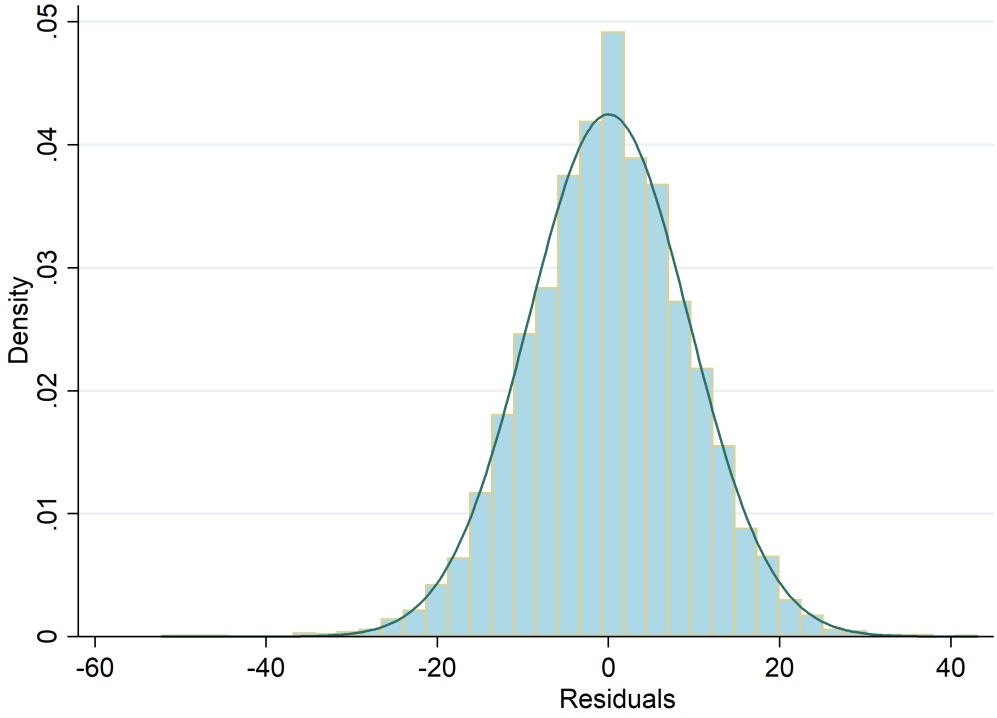


Figure 14: Residuals from objective model

C Survey Instructions

C.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.

C.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)

C.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.

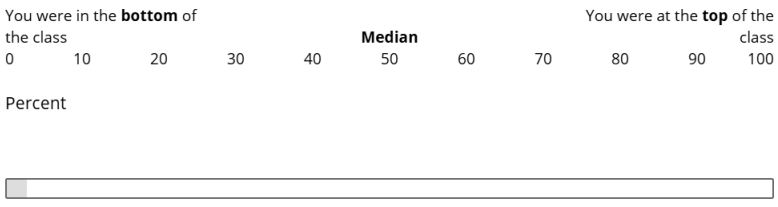
In the previous survey, you predicted the class average for test 4 would be 24.4 out of 40.

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

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

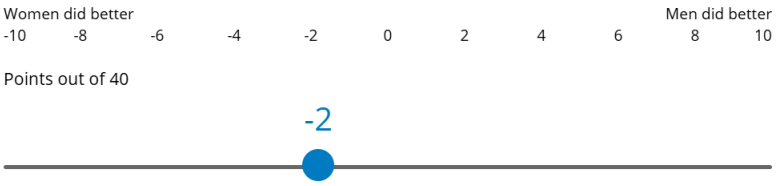
What percent of MAT137 students do you think that you actually outperformed in test 4 on March 24?

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



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 higher do you think male students actually scored in test 4 on March 24, compared to female students, on average?

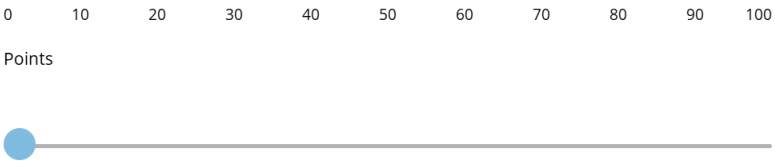


The following questions are about the final exam on Wednesday, April 19. In all of the questions, you can assume that the final exam is graded out of 100 points.

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)?

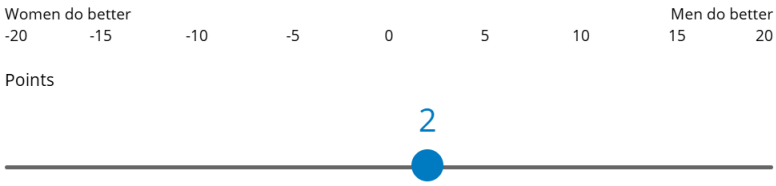
Your answer for **test 4**:60



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

0 10 20 30 40 50 60 70 80 90 100

Points

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

You will be in the **bottom** of the class — **Median** — You will be at the **top** of the class

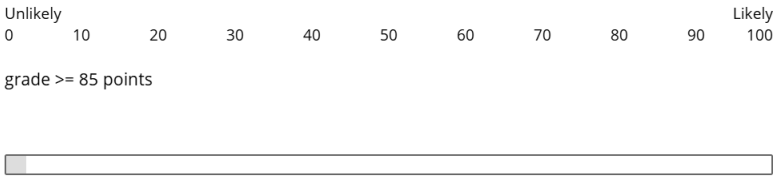
0 10 20 30 40 50 60 70 80 90 100

Percent

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

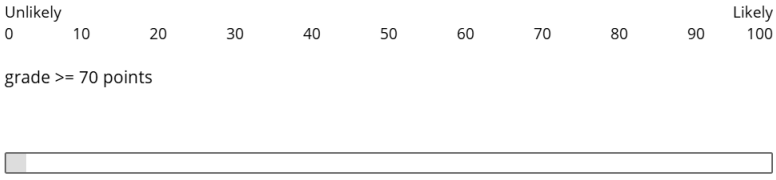
0 means it is impossible. 100 means it will happen with absolute certainty.

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



Done

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



C.4 Belief Elicitation: Testing Noise

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

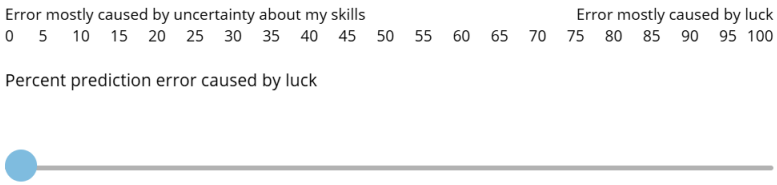
Making an accurate prediction is difficult because grades are determined by both **your MAT137 skills** and **luck.**

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?

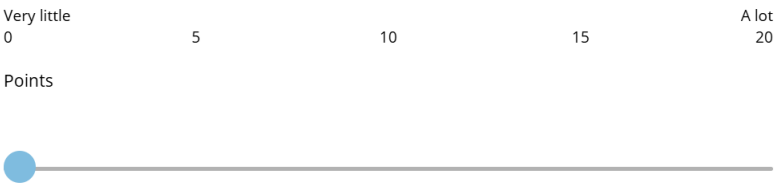
Your answer last survey: 60



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

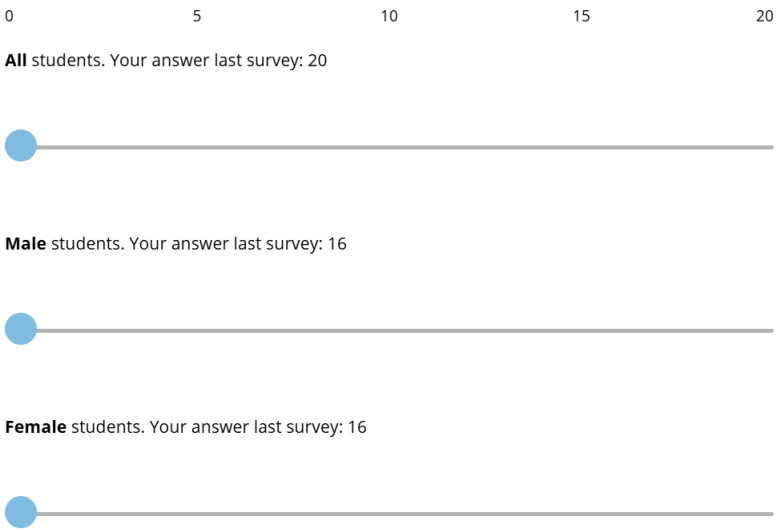
On average, luck increases their test score **slightly** : 0 to 3 points	0
On average, luck increases their test score **moderately** : 3 to 10 points	0
On average, luck increases their test score **dramatically** : more than 10 points	0
Total	0

C.5 Other Questions

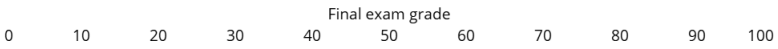
The following questions may be selected for payment. The reward is determined by the accuracy of your predictions. In order to maximize your payment, you should answer truthfully. Individual responses are strictly anonymous and will never be shared with instructors.

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.	○	○	○	○	○