Understanding Performance Declines*

Carl Meyer Germán Reyes Jason Somerville

September 2026

Abstract

Performance declines over the course of standardized tests affect individual scores, contribute to achievement gaps across demographic groups, and explain a substantial share of cross-country variation in PISA scores. Yet, despite their pervasiveness and importance, these declines remain poorly understood. We investigate their sources and consequences in an online experiment where participants complete a math exam. We randomly assign enhanced incentives ("bonus boosts") on one-third of the questions—either the first, middle, or last ten—and elicit preferences over this timing. Performance deteriorates substantially: the likelihood of a correct answer falls approximately 10 percentage points from beginning to end. Ex ante, participants fail to forecast this decline; ex post, they recognize only about half its magnitude. Choice data reflect this failure: nearly 60 percent of participants report no preference over when to receive bonus boosts, and willingness to pay does not differ significantly across timing options. This naivete is costly—receiving boosts early (vs. late) raises expected bonus earnings by around 9 percent. Falling cognitive effort accounts for one-third to one-half of the decline, and participants who most underestimate their fatigue show the steepest declines in effort and performance. Participants view this as suboptimal: when asked about optimal strategies, they report that effort should be roughly constant. These findings point to optimization failures as a novel explanation for within-exam performance declines, with implications for exam design, achievement gaps, and the broader set of settings where performance deteriorates over time.

*Meyer: Yale University, email: carl.meyer@yale.edu; Reyes: Middlebury College and IZA, email: greyes@middlebury.edu; Somerville: University of California, Santa Barbara, email: jasonsomerville@ucsb.edu. For helpful comments and suggestions, we thank Judd Kessler, George Loewenstein, Muriel Niederle, Dmitry Taubinsky, Roberto Weber, Alistair Wilson, and seminar participants at Middlebury College, Stanford University, the University of Pittsburgh, the University of Southern California, UC Santa Barbara, and the University of Zurich, as well as conference audiences at the 2025 Bay Area Behavioral and Experimental Economics Workshop and the 2025 Economic Science Association North American Meeting. The experiment reported in this paper was preregistered in the AEA RCT Registry in August 2025, under the ID AEARCTR-0016565. The experiment was approved by the Institutional Review Board at the University of California Santa Barbara under protocol number 3-25-0507.

1 Introduction

Students perform worse on later questions of tests (Borghans and Schils, 2018; Reyes, 2023; Brown et al., 2025), and these within-exam declines have consequences that extend well beyond individual test-takers. The magnitude of declines varies across demographic groups—for example, females show more sustained performance than males, contributing to gender gaps in test scores (Balart and Oosterveen, 2019). Because test scores play a central role in college admissions, these within-exam gaps can translate into disparities in educational attainment and labor market outcomes. Declines also vary substantially across countries and explain a substantial share of the cross-country variation in PISA scores (Zamarro et al., 2019). Therefore, understanding why these declines occur is important for interpreting both the ranking of countries on standardized assessments and score gaps across demographic groups within countries.

Yet the origins of these performance declines remain poorly understood. One view emphasizes human capital: individuals differ in their *cognitive endurance*, and performance declines reflect this heterogeneity (Brown et al., 2025). A complementary possibility—which we propose and test—is that performance declines partly reflect optimization failures: students may fail to anticipate how their performance will evolve and, as a result, misallocate effort over the course of an exam. Understanding which explanation applies has direct implications for whether the right policy response is investing in endurance-building interventions or redesigning task environments to account for predictable declines.

We conduct an online experiment where participants complete a 30-question mathematics exam with randomized question order. We vary the timing of "bonus boosts"—increased incentives for one-third of questions—and elicit participants' preferences over this timing (early, middle, or late) as well as their willingness to pay (WTP) for a randomly assigned option. For 80 percent of participants in the bonus-boost treatment, we randomize actual timing independently of preferences. This design allows us to identify whether participants anticipate performance declines, whether misperceptions lead to costly mistakes, and how participants reallocate effort in response to incentives.

We first document that performance deteriorates substantially as the exam progresses. The likelihood of correctly answering a question falls by 9.5 percentage points (pp) from beginning to end, and 82 percent of individuals exhibit a performance decline. Aggregating by thirds, performance drops by 6.7 pp from the first to the last third (from 66.2 percent to 59.5 percent). Yet participants' pre-exam forecasts imply constant performance, and post-exam hindcasts capture only about half of the actual decline. These patterns indicate ex ante naivete and ex post imperfect learning.

With constant expected performance, participants have no reason to prefer early boosts over late ones. Choice data reflect this naivete: 58.0 percent report no timing

preference, only 34.5 percent choose early boosts, and WTP is statistically indistinguishable across the three options. Timing choices also track stated performance trajectories: first-third choosers forecast a decline, no-preference participants forecast nearly flat performance, and middle/final choosers forecast improvement. This indifference is costly: assigning boosts to the first rather than last third raises expected bonus earnings by 9 percent.

Why don't participants anticipate that their performance will decline? We provide suggestive evidence by decomposing the performance decline into changes in effort and changes in the productivity of effort, and examining participant beliefs about each. Effort falls over the exam: time spent per question declines 24 percent, from 93 to 71 seconds. Regression estimates attribute one-third to one-half of the performance drop to declining effort. The residual is consistent with cognitive fatigue lowering the productivity of effort. Mental fatigue more than doubles over the course of the exam. Participants do not view the effort decline as optimal even ex post, and most fail to anticipate how tired they will become: only 27.5 percent accurately predict their post-exam fatigue, and those who most underestimate it show steeper declines in both effort and performance. Together, these patterns suggest that individuals' imperfect knowledge of how their own effort and fatigue evolve over time prevents them from anticipating their performance decline.

Beyond exams, our findings have novel implications for the design of workplace schedules and incentive structures (Bloom and Reenen, 2011; Buser and Peter, 2012; Dohmen and Shvartsman, 2023). Because the evidence points to miscalibrated self-knowledge rather than misaligned incentives, a natural first response is to design interventions that correct beliefs. However, such information interventions frequently fail or backfire (DellaVigna and Linos, 2022; Coffman et al., 2024), and few organizations track within-task performance with the granularity needed to provide tailored feedback. A potentially more promising approach is to design policies that mitigate the consequences of miscalibrated beliefs—for instance, scheduling high-stakes tasks earlier in shifts, adding review steps for late-day decisions, or structuring breaks rather than leaving them to worker discretion. Indeed, Schubert and Wang (2025) find that mandatory breaks raise productivity enough to more than offset forgone earnings, yet only 19 percent of workers voluntarily choose to rest when given the option.

We contribute to a literature on within-person performance declines that cuts across education (e.g., Ackerman and Kanfer, 2009; Reyes, 2023; Brown et al., 2025), healthcare (e.g., Chan et al., 2009; Linder et al., 2014; Dai et al., 2015; Steiny Wellsjo, 2025), labor economics (e.g., Pencavel, 2015; Collewet and Sauermann, 2017), sports (e.g., Archsmith et al., 2025), and law and criminal justice (e.g., Danziger et al., 2011; Meier et al., 2025; Ferrazares, 2025). While prior work focuses on documenting these declines and identifying their proximate causes, we study their consequences for the allocation of high-stakes work.

We show that individuals fail to account for these declines when they allocate effort and high-stakes opportunities, and that this failure seems to be rooted in miscalibrated beliefs.

More broadly, our findings complement the literature on assigning workers to tasks (e.g., Roy, 1951; Sattinger, 1993; Acemoglu and Autor, 2011). That literature asks which worker should do which task, given heterogeneous skills and task requirements. We ask a different question: when within a session a worker should do each task, given that productivity varies predictably. Whether workers solve this within-session assignment problem efficiently, and what frictions prevent them from doing so, are open questions. Our results suggest they do not: the binding friction is imperfect self-knowledge.

The remainder of the paper is organized as follows. Section 2 provides a simple framework that links performance beliefs, incentive timing, and WTP. Section 3 describes the experimental design. Section 4 documents the within-exam performance decline. Section 5 tests whether participants allocate high-value tasks to their most productive moments and quantifies the cost of misallocation. Section 6 examines whether the misallocation reflects miscalibrated beliefs about performance. Section 7 asks why participants fail to anticipate the decline, decomposing it into changes in effort and effort productivity. Section 8 concludes.

2 Conceptual Framework

We introduce a simple framework to organize the analysis and clarify our main hypotheses. The framework takes performance declines as given, introduces beliefs about those declines, and derives what those beliefs imply for both boost-placement choices and WTP. It makes the connection between performance beliefs and allocation decisions precise, clarifies the role of performance responses to bonus boosts in our setting, and defines the empirical objects we estimate in Sections 5 and 6.

Consider a worker who completes tasks over two periods, early ($t = 1$) and late ($t = 2$).[1] The tasks are of constant difficulty and each pays the baseline stake w, so any change in performance across periods reflects differences in worker performance rather than differences in task difficulty. Let $q_t \in (0, 1)$ denote the true probability that the worker successfully completes a period-t task at the baseline stake, and define the performance decline as $D_q \equiv q_1 - q_2$. A boost raises the per-task payoff in some period from w to $w + b$ and may also raise the period-t success probability by $r_t \geq 0$, which we refer to as the *performance response* to the boost.

[1]The experiment in Section 3 divides the task into three periods; the two-period case is sufficient to describe our main hypotheses, and the framework extends to any number of periods.

Rational benchmark. Under correct beliefs about both baseline performance and performance responses, the worker places the boost early rather than late if

$$b\pi(q_1 - q_2) + (w + b)\pi(r_1 - r_2) \geq 0, \tag{1}$$

where π is the probability that the high-value task's outcome determines the worker's payoff. The first term captures the value of placing the boost in the period with higher baseline performance ($D_q > 0$ favors early). The second term captures the performance response: if r_2 is large enough relative to r_1, late placement can be optimal even when baseline performance declines—for instance, if a boosted payment is expected to drive the success probability close to one in either period, it adds more value where baseline performance is lower.

The same setup also defines the optimal WTP under accurate beliefs. WTP for a boost in period t is approximately $WTP_t^* \approx b\pi q_t + (w + b)\pi r_t$, combining the direct value of the higher stake and the value of the induced performance response. When performance responses are similar across periods ($r_1 \approx r_2$), the rational WTP premium for early over late boosts simplifies to

$$WTP_1^* - WTP_2^* \approx b\pi(q_1 - q_2) > 0. \tag{2}$$

Under correct beliefs, the worker would therefore pay more for boosts in the period with higher baseline performance.[2] Section 5.2 shows that performance responses are indeed similar across periods in our experiment.

Beliefs. The worker forms perceived success probabilities $(\tilde{q}_1, \tilde{q}_2)$ and perceived performance responses $(\tilde{r}_1, \tilde{r}_2)$. Beliefs about success probabilities can be inaccurate in both level and slope:

$$\tilde{q}_1 - \tilde{q}_2 = \lambda D_q, \tag{3}$$

$$\tfrac{1}{2}(\tilde{q}_1 + \tilde{q}_2) = \tfrac{1}{2}(q_1 + q_2) + \mu. \tag{4}$$

The parameter λ measures the extent to which perceptions of the performance decline align with reality; $\lambda = 1$ corresponds to full calibration and $\lambda = 0$ to flat beliefs. The parameter μ captures level optimism or pessimism about average performance. Perceived performance responses $\tilde{r}_t$ may likewise differ from actual r_t. The worker's choice rule

[2] Our formulation of WTP captures only the expected earnings from a boost, abstracting from effort costs and other potentially utility-relevant factors. These components could bias the level of the WTP benchmark. However, provided that these components do not vary systematically across exam thirds, they will cancel out in the early-vs-late difference that we examine empirically.

replaces true quantities with their perceived counterparts:

$$b\pi(\tilde{q}_1 - \tilde{q}_2) + (w + b)\pi(\tilde{r}_1 - \tilde{r}_2) \geq 0, \tag{5}$$

and the perceived WTP premium for early over late boosts is approximately

$$\widetilde{WTP}_1 - \widetilde{WTP}_2 \approx b\pi(\tilde{q}_1 - \tilde{q}_2), \tag{6}$$

provided $\tilde{r}_1 \approx \tilde{r}_2$.

Implications. Equations (1) and (5) highlight two ways late placement can arise. First, late placement can be rational if true r_2 is large enough relative to true r_1—a performance decline does not by itself imply that early placement is optimal. Second, late placement can arise due to miscalibrated beliefs, either about baseline performance ($\tilde{q}_t \neq q_t$, especially $\lambda < 1$) or about performance responses ($\tilde{r}_t \neq r_t$). The empirical strategy allows us to distinguish between these channels. Randomized boost assignment identifies actual r_t and lets us test whether late preference is rational; if true r_t does not favor late, observed deviations from the benchmark reflect miscalibrated beliefs. Direct choice data and the WTP benchmark premium $WTP_1^* - WTP_2^* \approx b\pi(q_1 - q_2)$ identify miscalibration in $\tilde{q}_t$—under correct beliefs all workers should prefer early boosts, so observed variation in choice and WTP reveals miscalibration. In our experiment, we elicit each participant's WTP for a randomly assigned third (first, middle, or final) of the exam (Section 3). Elicited perceived performance responses $\tilde{r}_t$ reveal whether miscalibration also operates through expected performance responses. Together, these three measurements allow us to isolate the source of any observed misallocation.

Forecasts and hindcasts. Forecasts and hindcasts give the empirical analog of $\tilde{q}_t$. Before the exam, participants forecast how many questions they will answer correctly in each third. These forecasts provide the ex ante analogue of $\tilde{q}_1 - \tilde{q}_2$. Under the proposed framework, timing choices and WTP should track these forecasted changes rather than realized changes. Sections 6.1 and 6.2 examine forecast calibration and the link between forecasts and choices.

After the exam, participants report hindcasts of their performance in each third. These hindcasts measure what participants infer from the experience of taking the task, before any question-level feedback. Our framework predicts that hindcasted changes should move toward actual changes if experience corrects belief errors, and remain flat if it does not. Section 6.3 tests this. We next describe the experimental design that measures these choices, beliefs, and incentive responses.

3 Experimental Design

The experiment proceeds through seven stages: instructions, practice questions, preference elicitation for incentive timing, comprehension checks, pre-exam belief elicitation, a mathematics exam, and post-exam assessments.

3.1 Overview of the Experiment

Instructions. Participants first receive instructions about a 50-minute mathematics exam with 30 multiple-choice questions (five options each) covering algebra and geometry content similar to the SAT or ACT.[3] Participants learn that questions appear in random order and that all answers are final—they cannot revisit or change previous responses. They may use pen, paper, and an on-screen calculator, but not the internet, textbooks, or outside help. Participants are informed that they must remain on the exam screen throughout: switching tabs or leaving the screen may result in disqualification.

Bonus Payment Structure. Participants receive a \$12 participation payment plus a performance bonus. At the end of the exam, we randomly select one question; if participants answer the selected question correctly, they earn the bonus associated with that question, and otherwise they earn nothing. This structure incentivizes effort on every question because participants do not know which one will determine their bonus. We randomly assign participants to one of two conditions: in the control group (probability 4/19), all 30 questions offer \$2 bonuses; in the "bonus boost" treatment (probability 15/19), 10 questions offer \$10 bonuses and 20 offer \$2.

Practice Questions. Before the preference elicitation, participants complete four practice questions drawn from the exam pool to familiarize them with the format and difficulty. Participants in the bonus-boost treatment see the boost banner on two of these four questions, randomly chosen, so they recognize the format when it appears in the main exam; control participants see no banner.

Bonus Boost Timing and Preference Elicitation. Participants in the bonus boost treatment complete two incentivized decisions about when they would like boosted questions to appear. In *Decision 1*, participants directly choose their preferred timing—first, middle, or final 10 questions—or indicate no preference, in which case we randomly assign the timing. After selecting, participants explain their choice in an open-text response. In *Decision 2*, participants complete a multiple price list (MPL) in which they choose

[3] We draw the questions from Brazil's national college entrance examination (*Exame Nacional do Ensino Médio*), which provides standardized items that are unfamiliar to U.S. test-takers and have known difficulty.

between boosted bonuses for a randomly assigned timing option and guaranteed cash amounts ranging from \$0 to \$3.00 in \$0.25 increments. Direct choices in Decision 1 reveal participants' preferred timing but not the strength of that preference. The MPL fills this gap by eliciting WTP for each option.

After participants submit both decisions, we determine their treatment assignment by random draw: with 10 percent probability we implement Decision 1, with 10 percent we implement a randomly selected row from Decision 2, and with 80 percent we assign timing uniformly at random across the three options. This procedure ensures incentive compatibility and generates exogenous variation in boost timing, which we use to identify its effect on performance.

Comprehension Checks. After the preference elicitation and treatment assignment, participants answer four comprehension checks on exam rules, bonus mechanics, and their assigned timing. These questions test participants' understanding of the rules and screen out inattentive participants (see Section 3.2).

Pre-Exam Belief Elicitation. Before the exam, participants forecast their performance and fatigue. For performance, they use sliders (0–10 scale) to indicate how many correct answers they expect in each third of the exam (see Appendix Figure B2). For fatigue, they report their current mental tiredness and predict it after all 30 questions, using the same 0–10 scale (see Appendix Figure B3 for exact wording and anchors).

The Mathematics Exam. Participants have 50 minutes to complete the exam, with remaining time displayed throughout. Each question page shows the question text, five response options, and an on-screen calculator. To distinguish effort-driven performance changes from variation in question difficulty, we independently randomize question order for each participant. For boosted questions, a banner indicates the enhanced bonus (see Appendix Figure B4). We record response choices and time spent on each question. Participants receive no feedback during the exam and learn only their total correct answers at the end.

Post-Exam Questions. After the exam, participants report their current mental tiredness and estimate how many questions they answered correctly in each third. They then report how they allocated mental effort (0–100 scale) and time (seconds per question) across the three sections, and how they would ideally allocate them if they could retake the exam. Finally, participants give advice to future exam takers. We incentivize the advice with a bonus: with 5 percent probability, a future participant reads it before starting their exam, and if they answer their bonus question correctly, the advice-giver earns an additional \$2. The survey concludes with a demographic questionnaire.

3.2 Sample and Summary Statistics

We recruit 475 gender-balanced participants through Prolific, restricting eligibility to U.S.-based adults with a high approval rate (minimum 200 prior approvals, 99 percent approval rate).

We apply several pre-registered exclusion criteria to ensure data quality. During the exam, we disqualify participants who disengage on the first seven questions (averaging under 20 seconds per question with fewer than four correct, or leaving the exam screen on all seven). For our main analyses, we exclude participants who complete all 30 questions in under 20 minutes, leave the exam tab for extended periods, switch tabs excessively, or fail multiple comprehension checks. To maintain exogenous variation in timing, we also exclude participants whose stated preference or MPL choice was implemented (the 20 percent for whom Decisions 1 or 2 determined boost timing). The results are robust to including these observations (Appendix B.1). These restrictions leave a main analysis sample of 345 participants.

Baseline characteristics are similar across treatment groups. Table 1 reports demographics, education, income, employment, and baseline engagement for the control group and the three bonus-boost timing groups (columns 2–5). Joint balance tests do not reject that the observed covariates are unrelated to assignment to any treatment arm and an omnibus multinomial-logit test also does not reject joint balance across all arms ($p = 0.732$). The sample averages 41.3 years of age, with 48.4 percent male, 63.8 percent white, and 35.1 percent holding a bachelor's degree. Participants answer 62.9 percent of exam questions correctly on average, with substantial heterogeneity across the sample (Appendix Figure A1).

4 Performance Declines

For misallocation of high-value tasks to arise, performance must vary systematically within the exam. We estimate both the average performance decline and the distribution of individual-specific declines, and find that performance falls substantially over the course of the exam. We first estimate the average change in accuracy from the first to the last question, holding fixed question difficulty. We then estimate participant-specific performance slopes, using empirical Bayes shrinkage to separate persistent heterogeneity in trajectories from noise in each participant's 30 observations.

To measure the average within-exam performance change, we estimate the regression

$$\text{Correct}_{ij} = \alpha + \beta \cdot \text{NormPos}_{ij} + \gamma_j + \varepsilon_{ij}, \tag{7}$$

where Correct_{ij} indicates whether participant i answered question j correctly, NormPos_{ij} is position normalized to range from 0 (first question) to 1 (last question), and γ_j are

Table 1: Balance of Baseline Characteristics by Treatment Assignment

			Randomized Boost Timing		
	All Participants (1)	No Boost (Control) (2)	Boost First 10 (3)	Boost Middle 10 (4)	Boost Final 10 (5)
Panel A. Demographics					
Age (years)	41.310	40.056	41.975	41.068	42.239
Male	0.484	0.456	0.418	0.545	0.511
White/Caucasian	0.638	0.656	0.582	0.636	0.670
Born in USA	0.945	0.944	0.962	0.932	0.943
Native English speaker	0.971	0.989	0.962	0.966	0.966
Panel B. Education					
No college	0.142	0.122	0.177	0.136	0.136
Some college or associates	0.325	0.356	0.354	0.273	0.318
Bachelor's degree or higher	0.522	0.511	0.456	0.580	0.534
Panel C. Income and Employment					
Income: less than 50k	0.284	0.233	0.329	0.307	0.273
Income: 50k–99k	0.394	0.456	0.354	0.318	0.443
Income: 100k or more	0.284	0.256	0.278	0.341	0.261
Employed	0.696	0.733	0.658	0.693	0.693
Not working	0.188	0.133	0.190	0.250	0.182
Student as primary occupation	0.046	0.033	0.025	0.057	0.068
Panel D. Experimental Engagement					
Passed comprehension (first try)	0.907	0.967	0.899	0.864	0.898
Desktop/laptop user	0.986	0.967	1.000	0.989	0.989
Baseline fatigue (1-10)	2.493	2.456	2.367	2.750	2.386
N	345	90	79	88	88
Balance F-statistic	.	1.180	0.648	0.930	0.399
Balance p-value	.	0.278	0.853	0.538	0.985

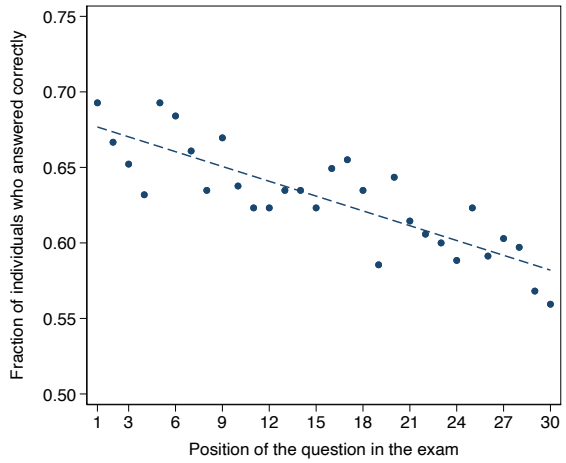
Notes: This table presents summary statistics for participants who passed the data-quality screens and, where relevant, whose bonus-boost timing was assigned at random. Column 1 reports all participants in this sample. The control group received standard $2 bonuses for all correctly answered questions. The three treatment groups received bonus boosts that increased payment from $2 to $10 for 10 questions, randomly allocated to either the first, middle, or final third of the 30-question exam. The balance F-statistics come from separate regressions of each assignment indicator on the covariates shown in Panels A–D; the corresponding p-values test whether those covariates jointly predict assignment to that arm. An omnibus multinomial-logit test using the same covariates does not reject joint balance across all randomized assignment arms ($p = 0.732$).

question fixed effects. Given this normalization, β represents the change in mean performance between the beginning and end of the exam. To examine heterogeneity, we estimate person-specific slopes $\hat{\beta}_i$ by regressing each participant's correct answers on NormPos. We apply an empirical Bayes shrinkage estimator to address measurement error in these individual-level estimates (see Appendix B.2); we denote the shrunk estimates β_i^s. We cluster standard errors at the participant level.

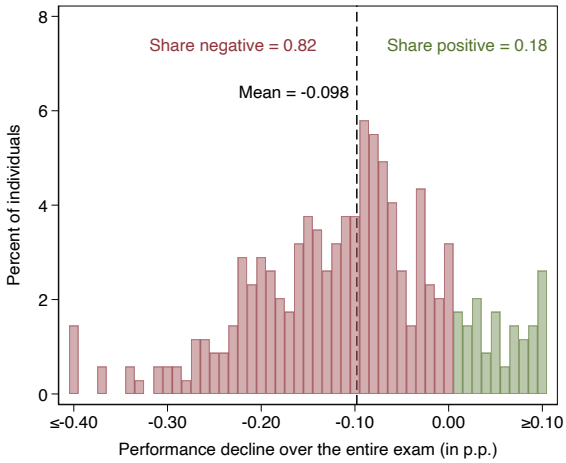
Average performance falls sharply over the exam. Figure 1, Panel A plots the share correct at each question position, along with a linear fit. The probability of a correct answer falls from 69.3 percent on the first question to 55.9 percent on the final question, a decline of 13.3 pp. The linear fit tracks the data closely, supporting the linear speci-

Figure 1: Performance Declines Over the Course of the Exam

Panel A. Average Performance by Question



Panel B. Individual Performance Declines



Notes: Panel A shows the fraction of participants answering correctly at each position in the exam. The dashed line represents predicted values from a linear regression of performance on question position. Panel B displays the distribution of individual-level performance slopes (β_i^s) from regressions of correct answers on normalized question position, adjusted using an empirical Bayes shrinkage estimator. For each participant i, we estimate $\text{Correct}_{ij} = \alpha_i + \beta_i \cdot \text{NormPos}_{ij} + \varepsilon_{ij}$, where NormPos_{ij} is normalized from 0 (first question) to 1 (last question). Negative values indicate performance declines over the course of the exam. The dashed vertical line indicates the mean slope.

fication in equation (7). The estimated coefficient is $\hat{\beta} = -0.095$ (Table 2, column 1), implying a 9.5 pp drop over the exam ($p < 0.01$), or 14.0 percent of initial performance (approximately 0.54 SD of overall scores). The estimate is unchanged with question and treatment fixed effects (column 2).[4]

Performance declines for most participants. Figure 1, Panel B plots the distribution of the shrinkage-adjusted slopes β_i^s. After accounting for measurement error, 82 percent of participants have negative slopes ($\beta_i^s < 0$). The mean shrinkage-adjusted individual-level decline is $\mathbb{E}[\beta_i^s] = -0.098$, indistinguishable from the pooled estimate of $\hat{\beta} = -0.095$ in equation (7). The standard deviation of individual-level slopes is $\hat{\sigma}_\beta = 0.124$, indicating substantial heterogeneity in the magnitude of the decline across participants.[5,6]

The decline is large within our sample and relative to standard benchmarks from labor and education. A participant performing at our first-question estimate would rank in the 57th percentile of our sample, while one performing at our last-question estimate would rank in the 39th percentile—a 18 percentile point drop. Mas and Moretti (2009) estimate

[4]Appendix B.5 shows that the decline is not driven by random guessing or broad disengagement. Rapid responses are rare, concentrated near the end of the exam, and excluding them leaves the main performance decline intact.

[5]These results are virtually identical when we exclude participants who received bonus boosts in the first third of questions, because their incentives drop from \$10 to \$2 after question 10, which could lower subsequent effort and mechanically generate a within-exam decline (Appendix Figure A3).

[6]Decline rates do not differ systematically across observable subgroups. We find no significant heterogeneity in pre-registered tests (by age, gender, and above-average education) or in exploratory regression-forest predictions from pre-treatment covariates (Appendix B.4).

Table 2: Estimates of Performance Declines over the Exam

	Outcome: =1 if answer is correct			
	(1)	(2)	(3)	(4)
Question position (normalized)	−0.095***	−0.094***	−0.062***	−0.054***
	(0.018)	(0.017)	(0.016)	(0.016)
Constant	0.677***	0.676***	0.352***	0.657***
	(0.012)	(0.012)	(0.029)	(0.011)
N (Question responses)	10,350	10,350	10,350	10,350
Question FE?	No	Yes	Yes	Yes
Treatment FE?	No	Yes	Yes	Yes
Time Controls?	No	No	Polynomial	Time-bin FE

Notes: This table presents estimates of β from equation (7). The dependent variable is an indicator for answering correctly. *Question position (normalized)* scales linearly from 0 (first) to 1 (last). Column 2 adds question and treatment fixed effects; column 3 adds flexible time controls (7th-order polynomial in minutes spent); column 4 adds time-bin fixed effects (ten second bins). Standard errors are heteroskedasticity-robust and clustered at the participant level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

a 21 percent P90–P10 gap in scan rates across grocery cashiers at a single supermarket chain, Chan et al. (2022) finds a 22 percent P90–P10 gap in diagnostic accuracy across emergency-room radiologists, and Lazear et al. (2015) estimate an 18 percent P75–P25 gap in output per hour across technology-services workers at a single firm. The median effect size of K–12 interventions is approximately 0.10 SD (Kraft, 2020). Our average within-exam decline of 0.54 SD exceeds this benchmark and is in line with typical cross-worker productivity gaps.[7]

5 The Allocation of High-Value Tasks

When performance declines, the timing of high-value tasks matters. In our experiment, high-value tasks are the ten questions in the boosted third of the exam, which pay $10 per correct answer rather than $2. The framework in Section 2 implies that participants who recognize the decline should value early placement, provided that performance responses to bonus boosts do not fully offset it. We study participants' stated preferences and WTP for the timing of bonus boosts, assess whether bonus boosts compensate for the underlying performance decline, and estimate the cost of misallocation.

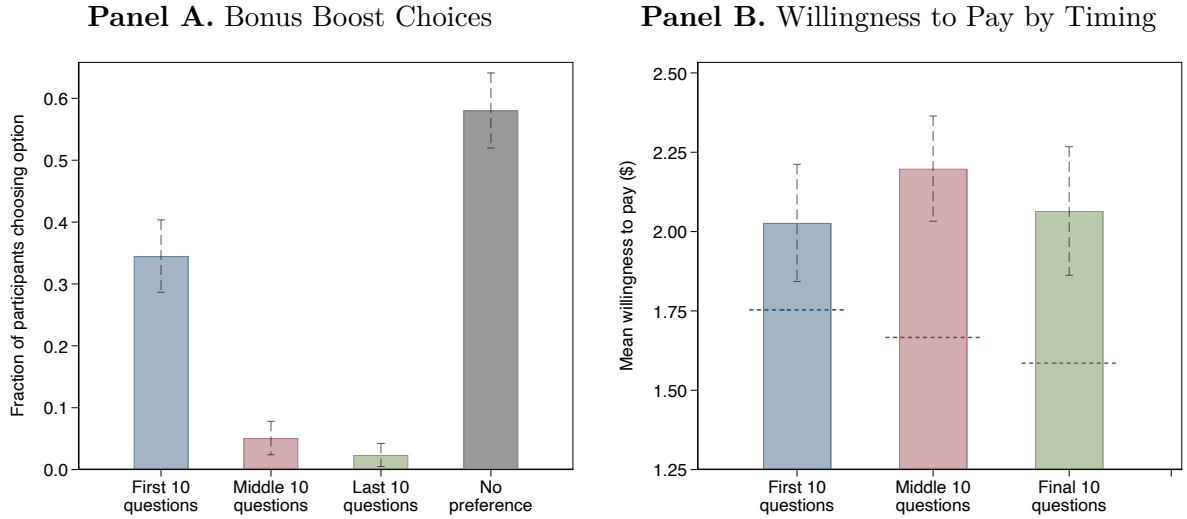
5.1 Bonus Boost Preferences

We measure participants' preferences over the timing of bonus boosts using two complementary measures: direct choice and incentivized WTP.

[7]The decline is not driven by our low-stakes online setting. Our questions are drawn from the same pool used in the Brazilian ENEM, a high-stakes college admissions exam. Our average within-exam decline of 9.5 pp (approximately 0.54 SD of overall scores) closely matches what Reyes (2023) documents for actual ENEM test-takers, who exhibit a cumulative within-exam decline of 7.1 pp (approximately 0.6 SD) in college-admission records for 15 million Brazilian high school students.

Direct Choice. Most participants do not choose early boosts, despite the documented performance decline. Figure 2, Panel A shows the distribution of timing preferences. Only 34.5 percent of participants chose to boost the first third of questions, and 7.5 percent chose the middle or last third. The modal response was indifference: 58.0 percent chose to let timing be randomly determined rather than selecting a specific third.

Figure 2: Bonus Boost Allocation Preferences



Notes: Both panels use bonus-boost participants who passed the data-quality screens and whose bonus-boost timing was assigned at random. Panel A shows the distribution of participants' choices for allocating bonus boosts that increase the payment from \$2 to \$10 if a question is answered correctly and randomly selected. Participants could choose to apply the 10 boosts to the first 10 questions, middle 10 questions, last 10 questions, or express no preference for random allocation. Panel B displays mean willingness to pay (in dollars) for a randomly selected third of the exam, derived from the multiple price list (MPL) task. Horizontal dashed lines show the predicted expected value of placing boosts in each third, calculated from baseline (non-boosted) success rates multiplied by the bonus stake (\$8) and the probability of random selection (1/3); baseline success rates are estimated from all observations where the third was at baseline incentive \$2 (pure control plus non-boosted thirds of treatment groups). Error bars in both panels represent 95 percent confidence intervals.

To understand what drives these choices, we ask participants to explain their decision in an open-text response. We then use an LLM-assisted coding procedure, with GPT-5 proposing the taxonomy and GPT-5 mini assigning labels, to classify each of the 255 responses in a multi-label scheme (Appendix Table B4).

Choice reasons track beliefs about the performance trajectory: those choosing the first third expect a decline, those choosing the middle or last third expect improvement, and those with no preference give reasons consistent with constant performance. Participants who chose the first third overwhelmingly cite expectations consistent with a decline—72.7 percent cite starting fresh and 56.8 percent cite fatigue. Those choosing the middle or last third typically express a belief that performance improves after warming up. Among no-preference participants, 51.4 percent state that timing does not matter (e.g., *the questions are random, so it does not matter*), 39.2 percent mention the difficulty of deciding, and

only 2.0 percent invoke fatigue or a performance decline. In Section 6, we assess the accuracy of these beliefs.

Willingness to Pay. Direct choices reveal which timing each participant prefers but not how strongly. To quantify the magnitude, we elicit WTP for a randomly assigned boosted third of the exam. Under the correct-beliefs benchmark, WTP should be higher for thirds with higher baseline success probabilities; the dashed lines in Figure 2, Panel B, plot this benchmark using the formula derived in Section 2.

Aggregate WTP does not differ across the three timings. Figure 2, Panel B, shows mean valuations of \$2.03 for the first third, \$2.20 for the middle, and \$2.06 for the last, with no pairwise difference statistically distinguishable from zero. Even among those who stated a preference, the higher fraction choosing the first third does not reflect stronger preferences for early boosts: among participants assigned to their preferred timing, mean WTP is \$2.16 for first-choosers, \$2.16 for middle-choosers, and \$3.00 for last-choosers (Appendix Figure A2). If anything, last-choosers express the strongest preferences.

Individual-level WTP supports the interpretation that the direct-choice measure captures meaningful preferences, but it does not overturn the weak aggregate preference result. Participants assigned to their preferred timing are willing to pay \$0.42 more than participants assigned to a non-preferred timing ($p = 0.025$), while participants who stated no preference have statistically flat valuations across assigned thirds (F-test $p = 0.37$; Appendix Figure A2). These patterns indicate that direct choices and incentivized valuations are internally consistent, even though neither measure shows a strong aggregate preference for early boosts.

Relative valuations also fall below the correct-beliefs benchmark. Comparing first-third valuations with middle- or final-third valuations, the observed first-minus-later WTP gap is −\$0.11 ($N = 255$), compared with a benchmark premium of \$0.13. This shortfall is significant against the one-sided alternative that the observed gap falls below the benchmark ($p = 0.022$).

Overall, direct choices and WTP point to weak preferences over the timing of bonus boosts, both in raw valuations and relative to the correct-beliefs benchmark. Whether these weak preferences are costly depends on whether the underlying performance decline is offset by increased effort on boosted questions. We test this possibility next.

5.2 Do Boosts Offset the Performance Decline?

The framework in Section 2 identifies a potential rationalization for indifference over boost timing: if performance responses to bonus boosts are sufficiently large for late versus early questions ($r_2 > r_1$), late placement can be rational even when baseline performance declines. We test this rationalization by estimating the effect of bonus

boosts on performance, using the random assignment of bonus-boost timing among the 80 percent of bonus-boost participants whose stated preferences were not implemented. Specifically, we estimate:

$$Y_{it} = \tau \cdot \text{Boosted}_{it} + \gamma_i + \delta_t + \lambda_{g(i,t)} + \epsilon_{it} \tag{8}$$

where Y_{it} is an outcome for participant i in exam third t, Boosted_{it} indicates whether participant i receives bonus boosts in third t, γ_i are participant fixed effects, δ_t are exam-third fixed effects, and $\lambda_{g(i,t)}$ are comparison-group fixed effects, where $g(i,t)$ indexes the (treatment-arm, exam-third) cell into which observation (i,t) falls within the boosted-versus-pure-control comparison.[8] We cluster standard errors at the participant level.

Bonus boosts increase time spent but not performance. Figure 3 compares time and performance during boosted thirds (pooled across treatment arms) with the pure control across exam thirds.[9] Participants spend 7.4 additional seconds per question on boosted items—a 9.3 percent increase ($p = 0.009$). Yet this increased effort translates into an estimated performance increase of only 2.3 pp (95 percent CI: -1.2 to 5.7 pp; $p = 0.195$). The confidence interval excludes a gain as large as the estimated 6.7 pp decline from the first to the final third. These effects do not vary significantly across exam thirds (Appendix Table A1, Panels B–C), and we find no evidence of spillovers to non-boosted questions (Appendix Table A4), indicating that bonus boosts do not affect overall performance trajectories.[10]

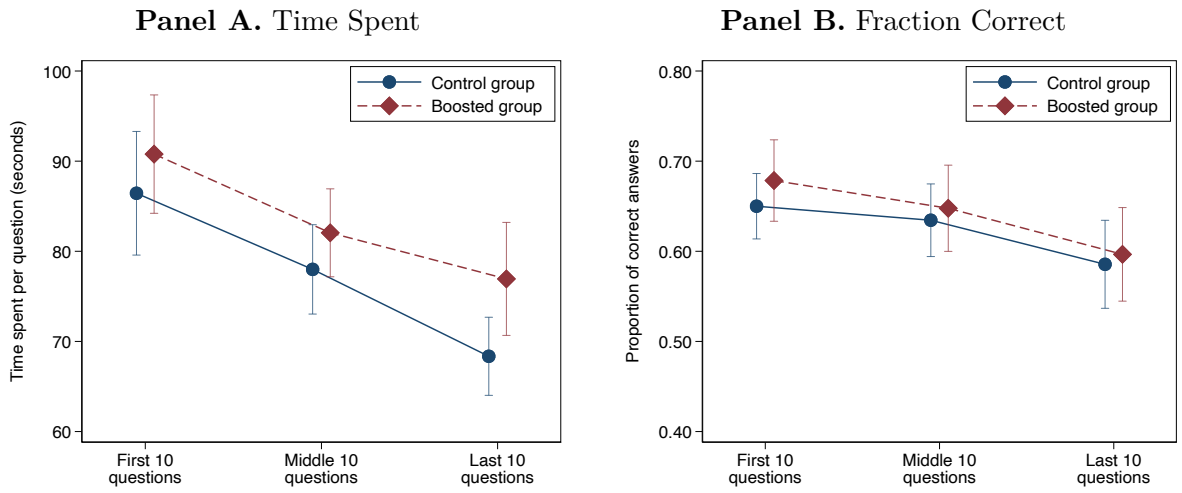
Thus, boosts do little to offset the performance decline, and the timing of bonus boosts translates directly into earnings differences. We quantify these earnings differences next.

[8] Comparison-group fixed effects address potential spillover effects: if receiving boosts in one third affects performance in other thirds (e.g., through fatigue from increased effort or effort substitution across exam thirds), comparing within treated participants would bias our estimates. By incorporating the pure control group—who face uniform incentives throughout—as a consistent baseline, we isolate the direct effect of bonus boosts from any spillover effects across exam thirds.

[9] See Appendix Figures A12 and A13 for additional outcomes and by-treatment-arm patterns.

[10] The same random variation could in principle support a cross-partial test in the spirit of Bronchetti et al. (2023): under a model of optimal effort allocation, incentive changes in one third should generate symmetric cross-third performance responses. In practice, the boost-induced performance responses and spillovers in our setting are too small for such a test to be informative.

Figure 3: Time Spent and Question Performance: Control vs. Bonus-Boosted



Notes: This figure compares outcomes between control participants (navy circles, solid lines) and participants who received bonus boosts in their assigned third (maroon diamonds, dashed lines). The boosted group pools participants across all three treatment arms, showing only their performance during incentivized periods. Panel A shows time spent per question; Panel B shows fraction correct. Error bars represent 95 percent confidence intervals. Appendix Figure A12 reports the same comparison for forecasted performance, retrospectively reported ideal time, perceived effort, and retrospectively reported ideal effort.

5.3 The Cost of Misallocating High-Value Tasks

Misallocation reduces participants' expected earnings, both within the boosted third and across the full exam. Among participants whose timing was randomly assigned, those assigned to early boosts earn \$2.26 in expected bonus from the first third, compared to \$1.99 for late-boost participants on their final third—14 percent higher in the early group ($p = 0.010$, one-sided). Across all 30 questions, total expected bonus payments are \$3.09 for the early group versus \$2.84 for the late group, 9 percent higher in the early group ($p = 0.035$, one-sided).[11]

Participants' relative valuations also fail to price the payoff consequences of timing. As shown in Section 5.1, the observed first-minus-later WTP gap falls significantly below the correct-beliefs benchmark.[12] Thus the earnings comparison shows that boost timing has payoff consequences, while the WTP comparison shows that participants do not price those consequences into relative valuations.

[11] We use one-sided tests for these comparisons because the allocation mistake has a directional prediction: early boosts should generate higher expected payoffs than late boosts. The smaller total-exam gap relative to the boosted-third gap reflects identical \$2 payments on middle-third questions.

[12] The benchmark follows from Section 2: with $r_1 \approx r_2$, the cross-period variation in WTP_t^* is approximately $b\pi q_t$, where $b = \$8$, $\pi = 1/3$, and q_t is the baseline success probability in period t, estimated from pooled baseline-incentive observations. These are the same benchmarks that appear as the dashed lines in Figure 2, Panel B. The first-minus-final comparison gives a similar but less precisely estimated benchmark test because it excludes participants assigned to value the middle third. The observed first-minus-final WTP gap is $-\$0.04$, compared with a correct-beliefs benchmark of \$0.17 ($p = 0.072$, one-sided; $N = 170$).

6 Miscalibrated Performance Beliefs

We document that participants fail to allocate high-value tasks to the periods when they are most productive, and that this misallocation is costly. Why do participants misallocate high-value tasks? The framework in Section 2 points to two belief channels that govern timing preferences: perceived baseline performance $\tilde{q}_t$ and perceived performance responses to bonus boosts $\tilde{r}_t$. We measure $\tilde{q}_t$ from pre-exam forecasts and post-exam hindcasts, and identify $\tilde{r}_t$ using the random assignment to boost timing. We then examine whether these beliefs predict timing choices, and ask whether experience corrects them.

6.1 Performance Forecasts

Before the exam, participants forecast their correct answers in each third ("forecasts"). We use these forecasts to construct implied beliefs about performance changes. Table 3 reports a descriptive calibration test: do forecasted changes line up with actual changes across participants? We implement a Mincer–Zarnowitz-style forecast-unbiasedness test (Mincer and Zarnowitz, 1969), regressing actual change on forecasted change:

$$\text{ActualChange}_i = \gamma_1 + \delta_1 \cdot \text{ForecastChange}_i + \eta_i, \tag{9}$$

where ActualChange_i is the difference between last-third and first-third correct answers, with ForecastChange_i defined analogously.[13] Accurate forecasts imply $\gamma_1 = 0$ (no bias) and $\delta_1 = 1$ (perfect calibration).

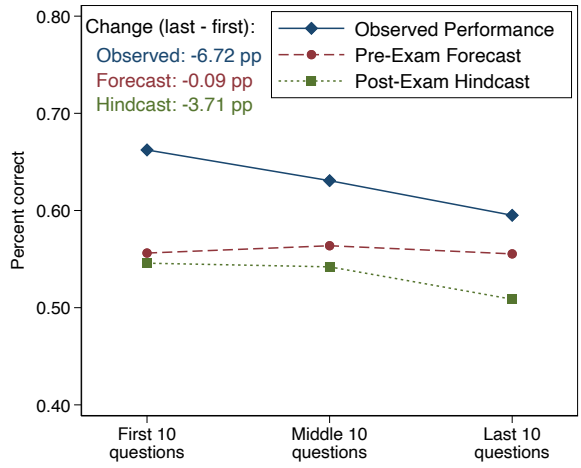
Participants do not anticipate their performance decline. The average forecast implies constant performance (55.6 percent correct in both the first and last thirds), while actual performance drops from 66.2 to 59.5 percent—a 6.7 pp decline (Figure 4, Panel A). The estimated forecast slope $\hat{\delta}_1 = 0.02$ ($p = 0.842$) is indistinguishable from zero and far below the perfect-calibration benchmark of $\delta_1 = 1$ (Table 3, columns 1–2). The estimated constant $\hat{\gamma}_1 = -0.67$ is significantly different from zero: participants forecasting no change actually decline by 0.67 questions on average. Through the lens of our conceptual framework, forecasts capture essentially none of the average decline: $\hat{\lambda}^F \approx 0.01$.[14] Appendix Figure A5 shows the same pattern nonparametrically by plotting forecasted and hindcasted changes against actual changes.

[13] This differs from our preferred performance measure (the $\hat{\beta}_i$ shrinkage estimates from Section 4), which uses all questions. We use the simpler last-minus-first measure here because it corresponds directly to our belief elicitation. The two measures are highly correlated ($\rho = 0.92$, Appendix Figure A4).

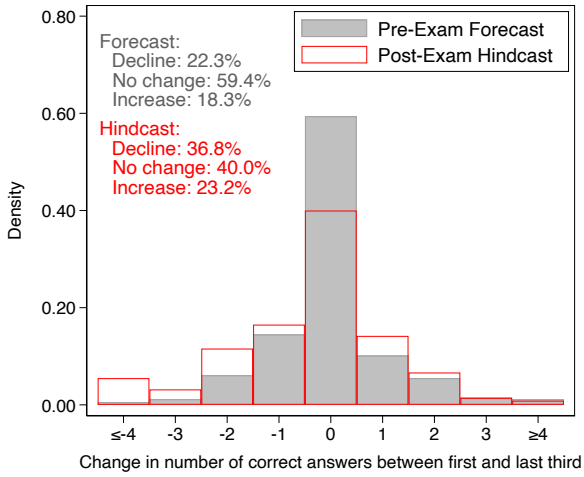
[14] Equation (3) defines λ as $(\tilde{q}_1 - \tilde{q}_2)/(q_1 - q_2)$. Averaging across the sample gives $\hat{\lambda}^F = \hat{E}[\tilde{q}_1 - \tilde{q}_2]/\hat{E}[q_1 - q_2]$. With an average forecasted change near zero and an average actual decline of 6.7 pp, this yields $\hat{\lambda}^F \approx 0.01$.

Figure 4: Beliefs About Performance Trajectories

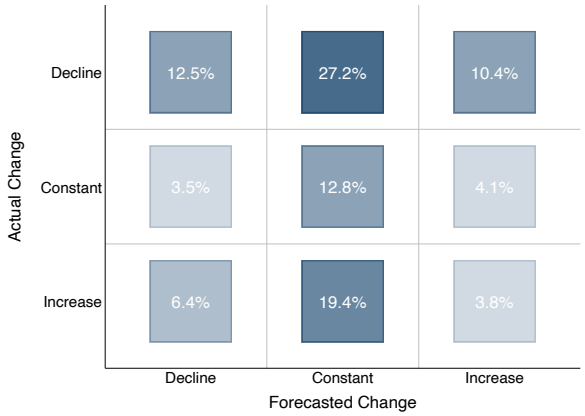
Panel A. Mean Performance by Exam Third



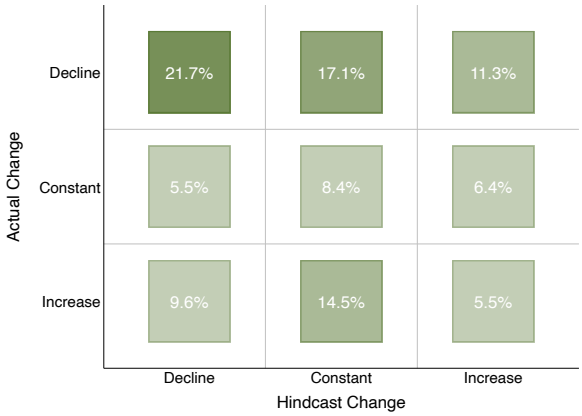
Panel B. Beliefs About Performance Change



Panel C. Actual vs. Forecasted Change



Panel D. Actual vs. Hindcasted Change



Notes: Panel A shows mean performance in the first, middle, and final third of the exam. The blue line with diamonds shows actual percent correct, the red line with circles shows pre-exam forecasts, and the green line with squares shows post-exam hindcasts. Panel B shows the distribution of beliefs about performance changes between the first and final thirds of the exam. Gray bars represent pre-exam forecasts and red outlined bars represent post-exam hindcasts. Panel C shows the joint distribution of actual performance changes (rows) and forecasted performance changes (columns) between the first and final thirds of the exam, with each cell showing the percentage of participants in that category. Panel D shows the joint distribution of actual performance changes and hindcasted performance changes between the first and final thirds. In Panels C and D, percentages sum to 100 percent across all nine cells.

Forecast errors are pervasive across participants. The distribution of forecasted changes concentrates at zero, with both the median and interquartile range at no change (Figure 4, Panel B). While 50.1 percent of participants actually experience a decline, only 22.3 percent forecast one. The most common pattern is an actual decline paired with a constant forecast (Panel C). Together, these results indicate a systematic bias toward expecting no performance decline, both on average and across most individuals. If participants expect performance to be flat, then many should see little reason to prefer early boosts over later boosts. With $\hat{\lambda}^F \approx 0$, the framework's perceived WTP premium

Table 3: Forecast and Hindcast of Performance Changes

	Outcome: Fraction of correct answers (final third - first third)			
	(1)	(2)	(3)	(4)
Change in forecast	0.022	0.023		
	(0.109)	(0.109)		
Change in hindcast			0.296***	0.302***
			(0.089)	(0.092)
Constant	−0.672***	−0.672***	−0.563***	−0.561***
	(0.128)	(0.128)	(0.124)	(0.124)
N (Question responses)	10,350	10,350	10,350	10,350
Question FE?	No	Yes	No	Yes

Notes: This table examines how well participants predict their own performance decline. The dependent variable is the actual change in the share of correct answers between the last third (questions 21-30) and first third (questions 1-10) of the exam. *Change in forecast* is the participant's predicted change in performance between these same thirds, elicited before taking the exam. *Change in hindcast* is the participant's retrospective estimate of their performance change, elicited after completing the exam. Columns 2 and 4 include question fixed effects. Standard errors are heteroskedasticity-robust and clustered at the participant level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

$\widetilde{WTP}_1 - \widetilde{WTP}_2 \approx b\pi(\tilde{q}_1 - \tilde{q}_2)$ from equation (6) collapses to approximately zero, which is consistent with the WTP indifference documented in Section 5.1.

We also find that participants are pessimistic about their level of performance. The average forecast is 55.8 percent correct, compared with an actual mean of 62.9 percent—an underestimate of about 2 questions out of 30 (Appendix B.3). Through the lens of our theoretical framework, this gap corresponds to $\hat{\mu}^F \approx -0.07$ in equation (4), where μ is the average gap between perceived and actual success probabilities. This level pessimism is distinct from the slope error $\hat{\lambda}^F$ documented above: timing decisions depend on $\tilde{q}_1 - \tilde{q}_2$ (governed by λ), not on the level (governed by μ).

Performance forecasts also identify the perceived performance response to bonus boosts, $\tilde{r}_t$. The framework in Section 2 shows that flat WTP can be rationalized by an asymmetric perceived performance response across periods ($\tilde{r}_2 > \tilde{r}_1$), even with correct beliefs about baseline performance. We measure $\tilde{r}_t$ by comparing pre-exam forecasts across treatment arms: each treatment-arm participant forecasts performance under their assigned boost timing (with the pure control forecasting under uniform stakes), so cross-arm contrasts within each third identify the perceived effect of a boost. Participants expect bonus boosts to raise per-question accuracy by only a small amount, with no systematic variation across exam thirds (Appendix Table A1). The implied $\tilde{r}_t$ values are similar across periods, so $\tilde{r}_1 \approx \tilde{r}_2$ in beliefs as well as in actual performance (Section 5.2). Miscalibrated performance-response beliefs therefore cannot rationalize the observed indifference; the binding belief friction is the flat $\tilde{q}_t$ documented above. We next test whether participants' timing choices line up with these stated beliefs.

6.2 The Link Between Performance Beliefs and Bonus Boost Choices

We next assess whether participants' timing choices line up with their stated forecasts. This is a key validation of our interpretation in Section 5—namely, that the misallocation pattern is not noise but a predictable consequence of the belief errors documented above.

The average forecasted change in performance differs systematically by timing choice. Participants who chose first-third boosts forecast a decline of 0.38 questions from the first to the final third ($p = 0.006$). Participants who expressed no preference forecast essentially no change, 0.09 questions ($p = 0.210$). Participants who chose middle or final boosts forecast an improvement of 0.68 questions ($p = 0.044$). All pairwise differences are significant at $p < 0.01$, except middle/final versus no-preference ($p = 0.065$).[15]

Individual-level consistency rates show the same pattern. Table 4 reports, by timing choice, the share of participants whose timing choice is supported by their own forecasts. Our preferred ("weak") measure counts an active choice as belief-consistent if the chosen third is tied for the highest forecasted performance, and counts a no-preference choice as belief-consistent only if the participant's forecasts are exactly flat.[16] Under this measure, 73.9 percent of first choices and 84.2 percent of middle/final choices are belief-consistent, and 56.1 percent of no-preference participants have exactly flat forecasts.

Table 4: Performance Beliefs and Choice Reasons by Bonus-Boost Timing Choice

		Direct timing choice			
	N	Forecast change	Strict belief	Weak belief	Text reason
	(1)	(2)	(3)	(4)	(5)
Chose first	88	-0.38	27.3	73.9	75.0
No preference	148	0.09	56.1	56.1	80.4
Chose middle/final	19	0.68	42.1	84.2	78.9

Notes: This table uses bonus-boost participants who passed the data-quality screens and whose bonus-boost timing was assigned at random. Forecast change is the stated final-minus-first-third change in correct answers out of ten. Strict belief support equals a strictly highest forecast for the chosen active third and exact flat forecasts for no preference. Weak belief consistency equals the chosen active third being tied for the highest forecast and exact flat forecasts for no preference. Text reason is coded conservatively. First choices are supported by start-advantage or fatigue reasons. No-preference choices are supported by indifference or cleaned expected-value reasoning. Middle and final choices are supported by warm-up and finish-strong reasons, respectively. Cleaned expected-value reasoning excludes responses also tagged as start advantage, warm-up, finish strong, fatigue, or perceived difficulty by position. Columns 3–5 report percentages.

A regression of the no-preference choice on forecasted performance change shows the same pattern. Relative to participants with flat forecasts (the omitted group), those

[15] Forecasts are elicited after the timing choice, so we interpret this evidence as consistency between choices and stated beliefs rather than as the causal effect of beliefs on choices.

[16] A flat forecaster's active choice is trivially consistent (their chosen third is tied for highest), but a non-flat forecaster's no-preference choice is not, since they should have preferred their forecast-best third. The table also reports a strict measure that requires an active choice to be uniquely best; we prefer the weak measure to avoid penalizing flat forecasters' active choices.

who expect a decline are 25.8 percentage points less likely to report no preference ($p = 0.001$); those who expect improvement are also less likely, though the estimate is not significant ($p = 0.465$). The result is robust to controls for realized performance and standard demographics (Appendix Table B5). Stated indifference is concentrated among participants whose forecasts imply little payoff difference across timings—exactly what the framework predicts when $\tilde{q}_1 \approx \tilde{q}_2$.

Participants' own explanations reinforce the link between timing choice and performance beliefs. Among first choosers, 75.0 percent cite being stronger at the start or fatiguing over time. Among participants that chose middle or final, 78.9 percent cite warming up or finishing strong. Among those stating no preference, 80.4 percent cite indifference or expected-value reasoning that treats timing as payoff-equivalent (coding details in Appendix Table B4). All four measures show that choices track beliefs. The misallocation of bonus boosts that we document in Section 5 is therefore consistent with belief errors. We next ask whether experience corrects those errors.

6.3 What Do Participants Learn from Experience?

After the exam, we again elicit participants' estimates of the number of correct answers in each third of the exam. They receive no question-level feedback during the task, and they do not learn their total score until after these hindcasts are elicited. The hindcasts therefore capture what participants infer from the experience of taking the exam rather than from direct feedback about specific mistakes or about overall performance. We do not elicit hindcasted performance responses to bonus boosts, so the learning analysis is restricted to perceived baseline performance $\tilde{q}_t^H$.

Participants only partially recognize performance declines after the exam. The average hindcasted decline is 3.7 pp, capturing 55 percent of the actual drop (Figure 4, Panel A). Through the lens of our conceptual framework, the implied post-task calibration is $\hat{\lambda}^H \approx 0.55$.[17] Section 2 predicts that hindcasted changes should move toward actual changes if experience corrects belief errors; we observe partial movement, with $\hat{\lambda}^H$ between the ex ante $\hat{\lambda}^F \approx 0$ and the calibrated benchmark $\lambda = 1$. Re-estimating equation (9) with hindcasted rather than forecasted changes gives a positive slope, $\hat{\delta}_1^H = 0.30$ ($p < 0.01$), but the relationship remains well below the one-to-one calibration benchmark. The estimated constant, $\hat{\gamma}_1^H = -0.56$, is again significantly different from zero, and thus ex-post beliefs remain pessimistic about the overall level of performance.

The individual-level patterns show the same incomplete learning. The share recognizing a decline rises from 22.3 percent ex ante to 36.8 percent ex post, but the modal pattern remains an actual decline paired with a constant hindcast (Figure 4, Panel D). Experience therefore moves beliefs in the right direction, but most participants still do

[17] Computed analogously to $\hat{\lambda}^F$: $\hat{\lambda}^H = \hat{E}[\tilde{q}_1^H - \tilde{q}_2^H]/\hat{E}[q_1 - q_2] \approx 3.7$ pp/6.7 pp.

not recognize the full decline in their own performance, even after being prompted before and after the exam to report performance separately by third.

Participants' advice to future test-takers offers a complementary test of learning.[18] If participants fully updated about the performance decline from experience, their advice should often warn future participants about declining performance, fatigue, or how to allocate effort and attention over the exam. We classify each open-text response using a single-best theme and a set of multi-label tags, then separately code categories tied to performance trajectories, fatigue, intertemporal tradeoffs, and effort-allocation rules.

The dominant advice themes are broad, generic test-taking heuristics rather than strategies that reflect an awareness of declining performance. The modal single-best theme is time management and pacing (40.0%), followed by careful reading (22.3%). The any-mention coding yields the same ranking: pacing (51.9%), careful reading (36.2%), and stress management (25.8%). These responses show that participants have no difficulty giving advice, but the advice usually concerns general exam behavior rather than the within-task performance gradient.

Explicit discussion of performance dynamics is rare. Only 9.6% state that performance declines over time, and just 3.5% attribute this decline to cognitive fatigue. Qualitative intertemporal reasoning is more common: 34.2% invoke tradeoffs such as avoiding burnout. But these observations rarely translate into concrete allocation rules. Only 11.3% endorse a flat effort path, and 8.4% provide a quantitative pacing rule (Appendix Table A6). Almost no respondents mention the random-selection incentive scheme, consistent with the earlier evidence that participants do not naturally frame the exam as an intertemporal allocation problem.

Advice themes are also broadly similar across treatment arms. We cannot reject equality of the single-best theme distribution ($p = 0.108$), and the dominant pacing and reading themes appear at comparable rates across groups (Appendix Table A5). The only notable differences are mechanical: bonus- and stakes-related advice appears among boost-arm participants but is essentially absent in the control group, which faced no differential stakes. Outside these mechanical differences, explicit references to performance decline, cognitive fatigue, and implementable pacing rules remain uncommon in all arms (Appendix Table A6). The advice evidence therefore reinforces the hindcast results: experience does not make the performance decline salient enough to become central in participants' advice to others.

[18] Hindcasts have two limitations as a learning measure: the question primes participants to think about performance over time, and the response is unincentivized. Advice removes both—it is incentivized and uses a generic prompt that does not reference timing—so participants must volunteer the decline themselves for it to appear.

7 Effort, Fatigue, and Performance Declines

Why do participants fail to anticipate their performance decline? We provide suggestive evidence on two margins they would need to forecast correctly: how their effort changes over the task, and how the return to that effort changes as fatigue accumulates. First, we use time spent per question as a proxy for effort and ask how much of the decline remains after accounting for changes in time investment. Second, we examine whether participants recognize the effort decline ex post and whether they anticipate the rise in fatigue ex ante.

7.1 Understanding Performance Declines

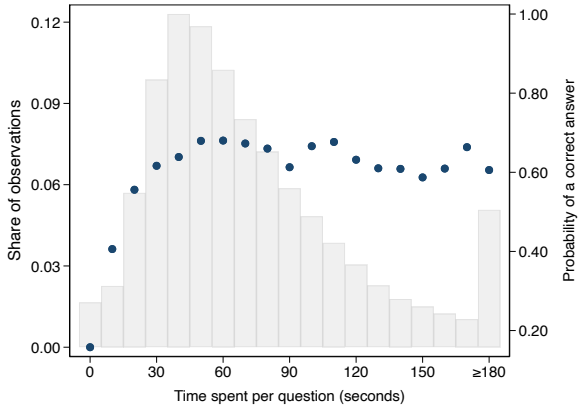
Performance can fall because participants invest less time and effort, because a given amount of time and effort produces lower accuracy later in the task, or both. We use time spent per question as a proxy for effort and validate this measure in two ways (Figure 5). First, time predicts performance: within the typical range of 0–60 seconds, the probability of a correct response increases from 15.8 percent at 0–10 seconds to 68.0 percent at 60–70 seconds, before plateauing (Panel A). Second, time aligns with self-reported effort: each additional 10 seconds corresponds to a 1.4-point increase in perceived effort on a 0–100 scale ($p < 0.01$, Panel B).

Time spent per question declines over the exam. Mean time falls from 93 seconds on the first question to 71 seconds on the final question—a 24 percent reduction (Figure 5, Panel C). The pattern holds at the individual level: 70.7 percent of participants exhibit declining time trajectories, with a mean reduction of 23.2 seconds from first to last question (Panel D).[19]

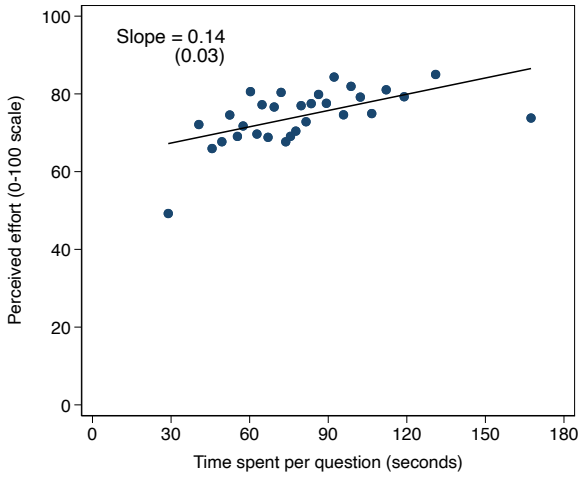
[19]The time decline is not driven by binding time constraints. The majority of participants finish with substantial time remaining (Appendix Figure A6, Panel A). Time investment declines smoothly over the exam rather than dropping sharply near the end, as would be expected under a binding time constraint (Appendix Figure A6, Panel B). Fewer than one percent of questions go unanswered, and missing responses occur predominantly at the very end of the exam (Appendix Figure A7). The decline also persists when we restrict the sample to participants who finished with at least three minutes remaining (Appendix Figure A10).

Figure 5: Time Allocation, Effort, and Exam Performance

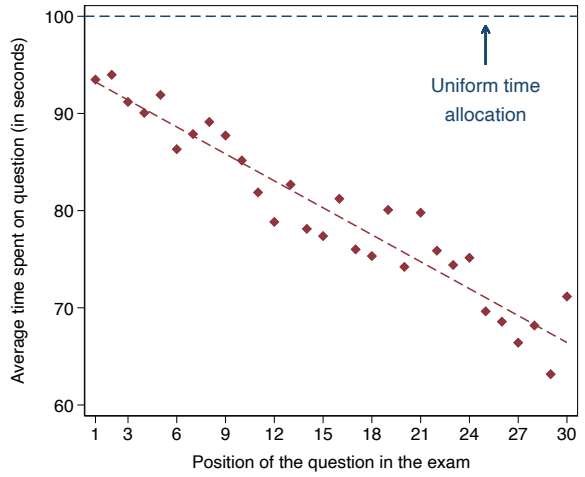
Panel A. Performance vs. Time Spent



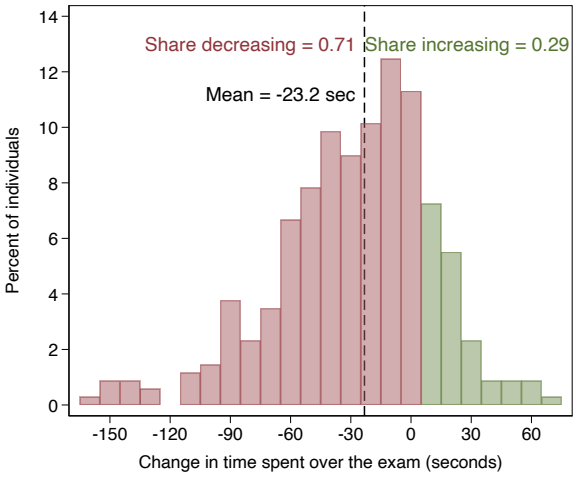
Panel B. Perceived Effort vs. Time Spent



Panel C. Time Spent Over the Exam



Panel D. Distribution of Time Changes



Notes: Panel A shows the relationship between time spent on questions and the probability of answering correctly. The scatter points (left y-axis) represent mean performance within 10-second time bins. The gray bars (right y-axis) show the distribution of observations across time bins. Time values are top-coded at 180 seconds. Panel B displays the relationship between actual time spent and perceived effort (0–100 scale), with each point representing mean perceived effort within 30 quantile bins of time spent. The regression slope with clustered standard errors quantifies how subjective effort assessments align with actual time allocation. Panel C shows average seconds spent per question at each position in the exam, with a linear fit demonstrating the declining trend. Panel D presents the distribution of individual-level time allocation slopes (δ_i^s) from regressions of time on normalized position, adjusted using empirical Bayes shrinkage. Negative values (red bars) indicate participants who spend progressively less time as the exam progresses.

Having documented the decline in effort, we now quantify how much of the performance decline it accounts for. We augment equation (7) with flexible controls for time spent, either as a seventh-order polynomial or as ten-second bin fixed effects:

$$\text{Correct}_{ij} = \alpha^T + \beta^T \cdot \text{NormPos}_{ij} + f(\text{Time}_{ij}) + \gamma_j^T + \varepsilon_{ij}^T. \tag{10}$$

The coefficient on normalized position falls from $\hat{\beta} = -0.095$ in the baseline to $\hat{\beta}^T = -0.062$ with polynomial controls and $\hat{\beta}^T = -0.054$ with time-bin fixed effects—a reduction of 34.2–42.6 percent (Table 2, columns 3–4). Declining time investment therefore accounts for roughly one-third to one-half of the performance decline. Appendix Figure A8 shows that time spent declines for both correct and incorrect responses. The remaining decline is visible within comparable time investments: even with 10-second time-bin fixed effects, the position coefficient remains negative and statistically significant. Appendix Figure A9 and Appendix Table A2 show the same pattern directly by exam third. Holding time bins fixed, correctness is 3.0 pp lower in the middle third and 4.6 pp lower in the final third relative to the first third. The decline therefore reflects two empirical margins: participants spend less time later in the exam, and later responses are less accurate even within the same time bins.

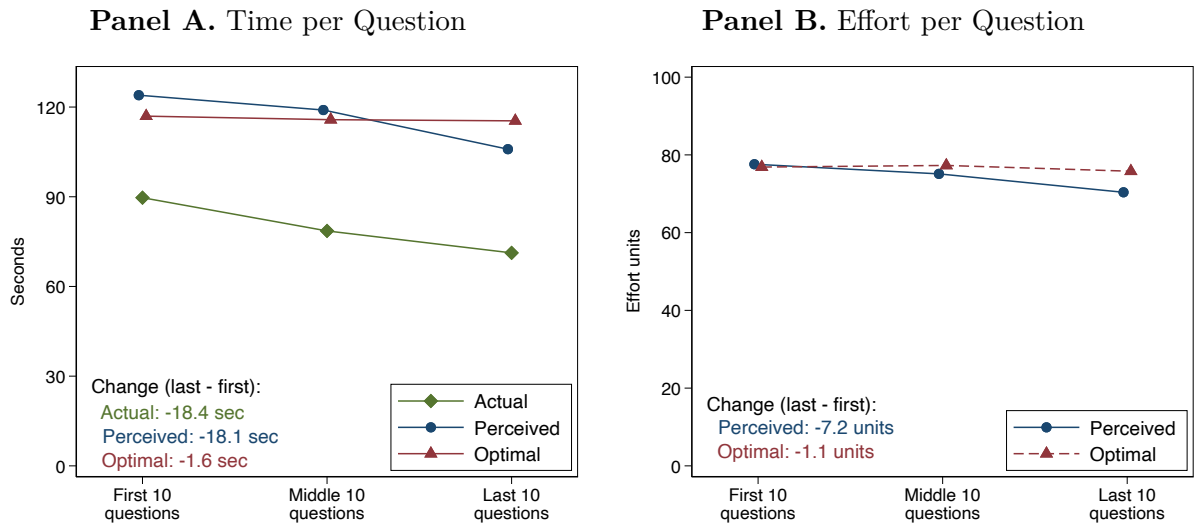
7.2 Perceived Effort Declines and Fatigue Forecast Errors

Perceived Versus Ideal Time and Effort Allocation. We first examine whether participants recognize the decline in their own inputs and whether they view that decline as optimal. After the exam, participants report both how they allocated time and effort and how they would ideally allocate them if they could retake the exam. If the decline was a deliberate strategy, perceived and ideal allocations should both fall; if it was unplanned, participants should perceive the decline but report a different ideal.

Participants recognize ex post that their inputs declined, but they do not view this decline as optimal. Figure 6 compares actual, perceived, and ideal allocations. Actual time spent declines from 90 seconds per question in the first third to 71 seconds in the last third. Participants accurately perceive this decline (reporting a 18-second reduction). Yet their stated optimal strategy is nearly constant, declining by only 2 seconds. The pattern for self-reported effort is analogous: perceived effort declines substantially, while stated optimal effort remains roughly constant at 76.9 units. The gap between perceived and ideal allocation thus suggests that part of the performance decline reflects suboptimal input allocation rather than deliberate pacing—a possibility participants likely did not account for when forecasting their performance.

Fatigue Forecast Errors and Performance Declines Why does effort decline over the exam, given that participants recognize ex post that this decline is not optimal? A natural hypothesis is that they misforecast how cognitively taxing the exam will become, in line with models of projection bias in effort choices (Kaufmann, 2022). To fix ideas, consider a test-taker beginning the exam. Feeling fresh, she expects the cost of effort to remain low and plans to work carefully on every question. Fatigue then sets in faster than anticipated: the questions feel more taxing than expected, and she exerts less effort than

Figure 6: Time and Effort Allocation: Actual, Perceived, and Optimal



Notes: Panel A compares actual time spent per question (green diamonds) with participants' perceived time allocation (blue circles) and their retrospectively reported ideal allocation (red triangles) across exam thirds. Panel B shows perceived effort (blue circles) and ideal effort allocation (red triangles) on a 0–100 scale across exam thirds. Error bars represent 95 percent confidence intervals.

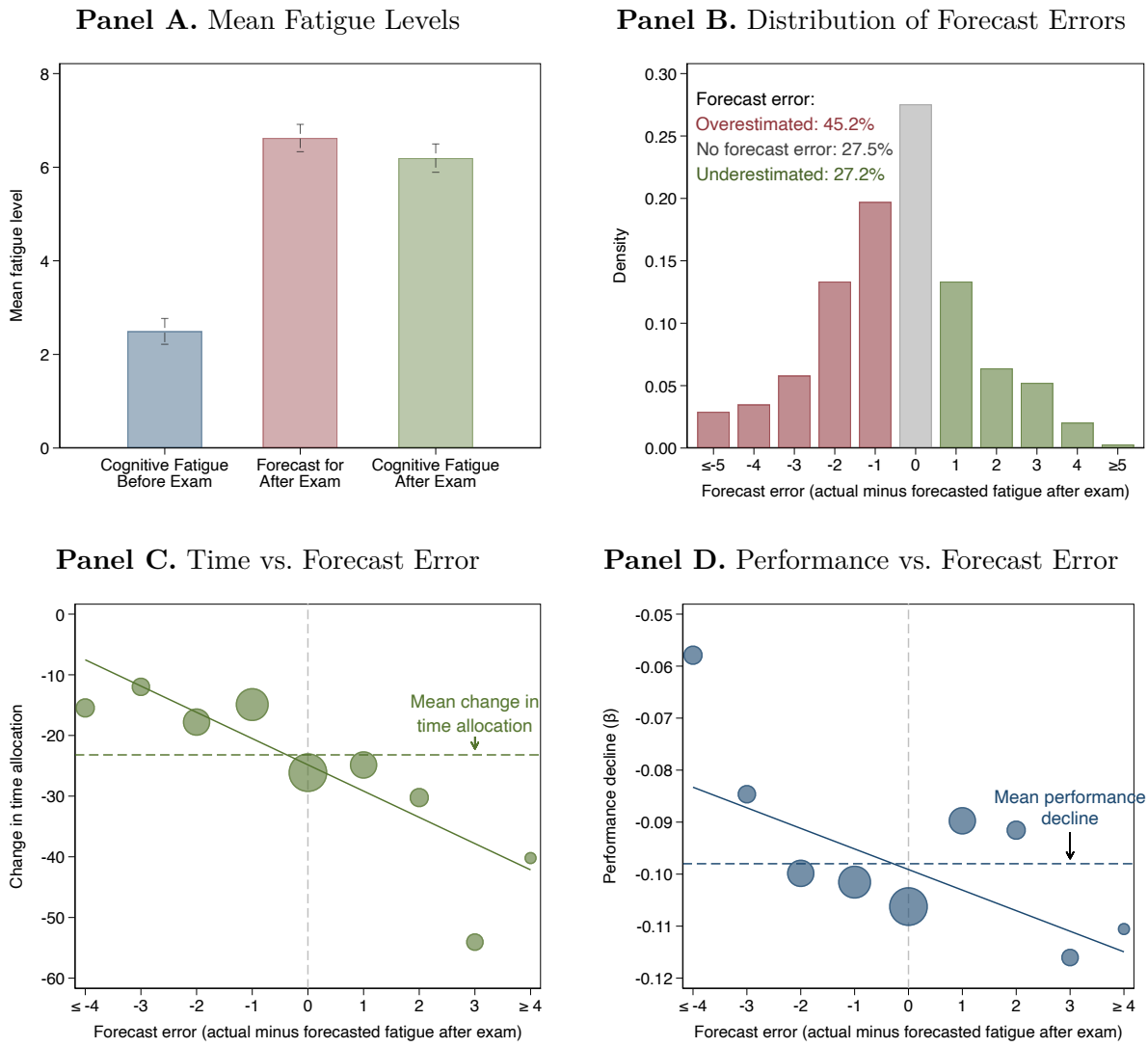
planned, especially on later questions. The result is a declining performance trajectory rooted in a fatigue forecast error. Under this account, participants who most underestimate their fatigue should exhibit the largest reductions in effort and performance. To test this prediction, we measure individual fatigue forecast errors as the difference between pre-exam predictions of post-exam fatigue and post-exam reports of actual fatigue.

On average, participants anticipate the increase in fatigue, but individual forecasts vary widely. Mental fatigue rises substantially over the exam, from 2.5 before to 6.2 after—a 3.7-point increase on a 10-point scale (Figure 7, Panel A). The mean forecast of 6.6 is close to actual post-exam fatigue. Yet this average masks substantial individual forecast errors: only 27.5 percent of participants accurately predict their fatigue, while 27.2 percent underestimate it and 45.2 percent overestimate it (Panel B).

Forecast errors predict heterogeneity in both effort and performance declines. Participants who underestimate fatigue exhibit larger reductions in time spent over the exam, while those who overestimate fatigue show smaller reductions (Panel C). A similar pattern emerges for performance: underestimating fatigue is associated with steeper performance declines (Panel D). This heterogeneity is consistent with the fatigue-forecast-error account.

Together, these results point to a fatigue-forecast-error account of the decline. Participants recognize ex post that time and effort declined and do not view this as optimal, and those who most underestimate their fatigue show the steepest declines in both inputs and performance. The evidence therefore links miscalibrated beliefs about cognitive fatigue to the input declines that accompany falling performance.

Figure 7: Cognitive Fatigue: Forecasts vs. Actual Experience



Notes: Panel A shows mean cognitive fatigue levels on a 0-10 scale at three time points: before the exam (baseline), forecasted level after the exam (predicted before starting), and actual level after exam completion. Error bars represent 95 percent confidence intervals. Panel B displays the distribution of individual forecast errors (actual minus forecasted post-exam fatigue). Values are winsorized at ±5 points. Panel C shows a binned scatterplot of the relationship between fatigue forecast errors and individual-level changes in time allocation across the exam (delta coefficients from time-on-position regressions). Panel D shows a binned scatterplot of the relationship between fatigue forecast errors and individual-level performance decline (β estimates from performance-on-position regressions). In Panels C and D, each point represents the mean within decile bins of forecast error, with the fitted line weighted by bin size. Horizontal dashed lines indicate population means. Positive forecast errors indicate participants underestimated their fatigue (felt more tired than expected).

8 Conclusion

In a controlled experiment, we show that individuals fail to allocate high-value work to high-productivity moments. Participants' performance declines sharply, yet most do not prefer to receive enhanced incentives early, and WTP is flat across timing options. This indifference is costly: assigning bonus boosts early rather than late raises expected

earnings. The evidence points to miscalibrated beliefs about within-task productivity as an important source of these mistakes. Participants forecast constant performance profiles before the exam, their timing choices line up with those forecasts, and they only partially recognize the decline after completing the task.

This mechanism likely extends beyond our experimental setting. Hospital nurses exhibit declining hand-hygiene compliance despite extensive training, and these declines reflect costly mistakes rather than strategic triaging of effort (Steiny Wellsjo, 2025). Physicians make more diagnostic errors over their shifts (Chan et al., 2009; Linder et al., 2014; Kim et al., 2018), and instructors give lower-quality feedback to assignments graded later (Wang et al., 2026). One might expect workers to mitigate these costs by taking breaks or other preventive measures, but recent evidence suggests these protective behaviors are themselves allocated suboptimally: workers avoid rest even when doing so harms their own performance (Schubert and Wang, 2025)—possibly because protective behaviors require the same self-knowledge our results suggest individuals lack.

A separate question is whether the performance decline itself reflects misoptimization. Three patterns are consistent with this interpretation. First, participants do not anticipate the performance decline ex ante, so they have no reason to plan around it. Second, most also misforecast their own fatigue, so even those who try cannot plan an optimal effort sequence. Third, ex post they report that ideal effort would have been roughly constant, even though their actual effort fell sharply. Whether these patterns reflect effort misoptimization or unavoidable cognitive fatigue hinges on how much of the effort decline is voluntary. Disentangling the voluntary from the involuntary share of the decline is an important direction for future work.

References

Acemoglu, D. and Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. In Ashenfelter, O. and Card, D., editors, *Handbook of Labor Economics*, volume 4, pages 1043–1171. Elsevier.

Ackerman, P. L. and Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied*, 15(2):163–181.

Archsmith, J., Heyes, A., Neidell, M., and Sampat, B. (2025). The dynamics of inattention in the (baseball) field. *The Economic Journal*, 135(671):2192–2219.

Balart, P. and Oosterveen, M. (2019). Females Show More Sustained Performance During Test-Taking than Males. *Nature Communications*, 10.

Bloom, N. and Reenen, J. V. (2011). Human Resource Management and Productivity. In Card, D. and Ashenfelter, O., editors, *Handbook of Labor Economics*, volume 4, pages 1697–1767. Elsevier.

Borghans, L. and Schils, T. (2018). Decomposing achievement test scores into measures of cognitive and noncognitive skills. *Working Paper. Available at SSRN 3414156*.

Bronchetti, E. T., Kessler, J. B., Magenheim, E. B., Taubinsky, D., and Zwick, E. (2023). Is Attention Produced Optimally? Theory and Evidence From Experiments With Bandwidth Enhancements. *Econometrica*, 91(2):669–707.

Brown, C., Kaur, S., Kingdon, G., and Schofield, H. (2025). Cognitive endurance as human capital. *The Quarterly Journal of Economics*, 140(2):943–1002.

Buser, T. and Peter, N. (2012). Multitasking. *Experimental Economics*, 15(4):641–655.

Chan, D. C., Gentzkow, M., and Yu, C. (2022). Selection with variation in diagnostic skill: Evidence from radiologists. *The Quarterly Journal of Economics*, 137(2):729–783.

Chan, M. Y., Cohen, H., and Spiegel, B. M. R. (2009). Fewer polyps detected by colonoscopy as the day progresses at a Veteran's Administration teaching hospital. *Clinical Gastroenterology and Hepatology: The Official Clinical Practice Journal of the American Gastroenterological Association*, 7(11):1217–1223; quiz 1143.

Coffman, L. C., Featherstone, C. R., and Kessler, J. B. (2024). A model of information nudges. Working paper, July 1, 2024.

Collewet, M. and Sauermann, J. (2017). Working hours and productivity. *Labour economics*, 47:96–106.

Dai, H., Milkman, K. L., Hofmann, D. A., and Staats, B. R. (2015). The impact of time at work and time off from work on rule compliance: the case of hand hygiene in health care. *Journal of Applied Psychology*, 100(3):846.

Danziger, S., Levav, J., and Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17):6889–6892.

DellaVigna, S. and Linos, E. (2022). RCTs to scale: Comprehensive evidence from two nudge units. *Econometrica*, 90(1):81–116.

Dohmen, T. and Shvartsman, E. (2023). Overexertion of Effort under Working Time Autonomy and Feedback Provision. *Journal of Economic Behavior & Organization*, 212:1255–1266.

Ferrazares, T. (2025). Shift structure and cognitive depletion: Evidence from police officers. Working Paper.

Kaufmann, M. (2022). Projection bias in effort choices. *Games and Economic Behavior*, 135:368–393.

Kim, R. H., Day, S. C., Small, D. S., Snider, C. K., Rareshide, C. A. L., and Patel, M. S. (2018). Variations in Influenza Vaccination by Clinic Appointment Time and an Active Choice Intervention in the Electronic Health Record to Increase Influenza Vaccination. *JAMA Network Open*, 1(5):e181770.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4):241–253.

Lazear, E. P., Shaw, K. L., and Stanton, C. T. (2015). The Value of Bosses. *Journal of Labor Economics*, 33(4):823–861.

Linder, J. A., Doctor, J. N., Friedberg, M. W., Nieva, H. R., Birks, C., Meeker, D., and Fox, C. R. (2014). Time of day and the decision to prescribe antibiotics. *JAMA internal medicine*, 174(12):2029–2031.

Mas, A. and Moretti, E. (2009). Peers at Work. *American Economic Review*, 99(1):112–145.

Meier, A., Levav, J., Meier, S., and Avnaim, L. G. (2025). Early release via parole and recidivism. Working Paper.

Mincer, J. A. and Zarnowitz, V. (1969). The evaluation of economic forecasts. In Mincer, J. A., editor, *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, chapter 1, pages 3–46. National Bureau of Economic Research.

Pencavel, J. (2015). The productivity of working hours. *The Economic Journal*, 125(589):2052–2076.

Reyes, G. (2023). Cognitive endurance, talent selection, and the labor market returns to human capital. *arXiv preprint arXiv:2301.02575*.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers*, 3(2):135–146.

Sattinger, M. (1993). Assignment models of the distribution of earnings. *Journal of Economic Literature*, 31(2):831–880.

Schubert, A. V. and Wang, J. M. (2025). Self-detrimental avoidance of rest. Working Paper.

Steiny Wellsjo, A. (2025). Simple actions, complex habits: Lessons from hospital hand hygiene. Working Paper.

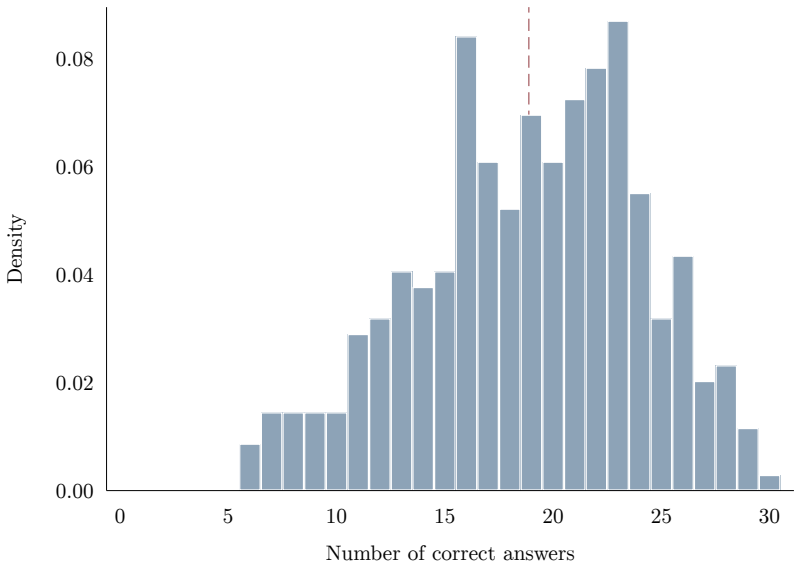
Wang, Z. H., Pei, J., and Li, J. (2026). 70 million canvas grading records reveal widespread sequential bias and system-induced surname initial disparity. *SSRN*. Last revised February 2026. SSRN Working Paper No. 4603146.

Zamarro, G., Hitt, C., and Mendez, I. (2019). When Students Don't Care: Reexamining International Differences in Achievement and Student Effort. *Journal of Human Capital*, 13(4):519–552.

Appendix

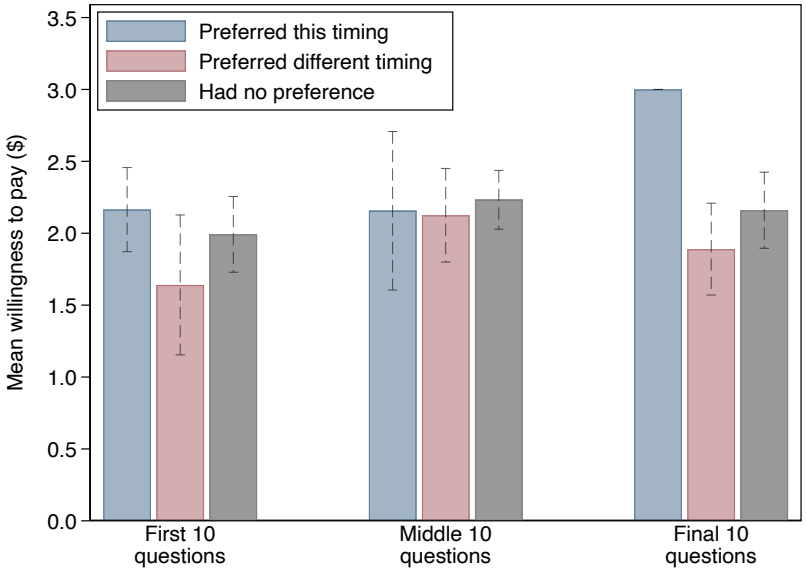
A Appendix Figures and Tables

Figure A1: Distribution of Number of Correct Exam Responses



Notes: This figure shows the distribution of total correct answers across all participants. The vertical dashed line indicates the mean (18.9 correct answers). The histogram uses bins of width one, ranging from 0 to 30 correct responses.

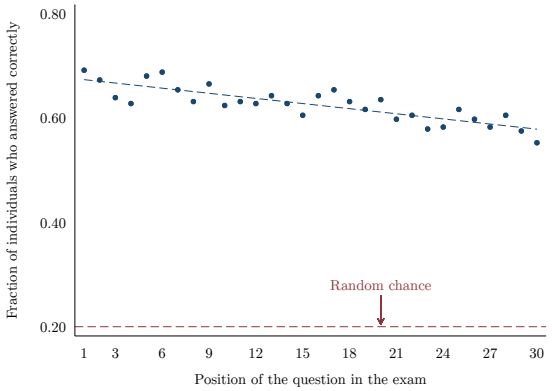
Figure A2: Willingness to Pay by Preference Match



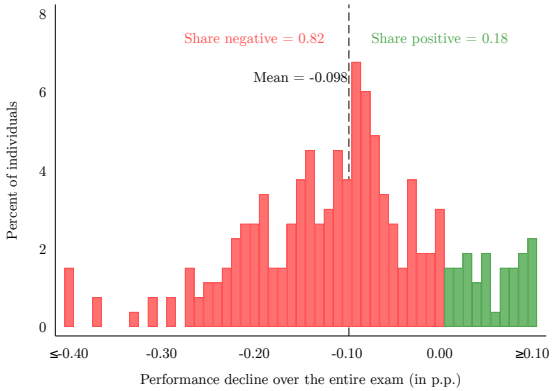
Notes: This figure uses bonus-boost participants who passed the data-quality screens and whose bonus-boost timing was assigned at random. It shows mean willingness to pay (in dollars) when participants were randomly assigned to value boosts in each third of the exam, separated by whether the assigned timing matched their Decision 1 preference (navy), differed from their preference (maroon), or they had expressed no preference (grey). Willingness to pay is derived from the multiple price list (MPL) task.

Figure A3: Performance Declines Excluding First-Third Boost Recipients

Panel A. Average Performance by Question Position

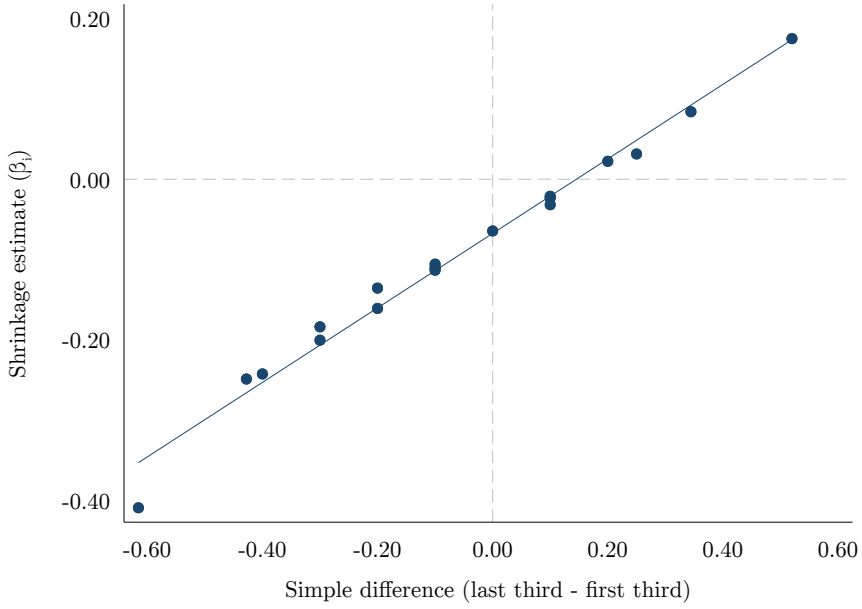


Panel B. Distribution of Individual Performance Declines



Notes: This figure replicates Figure 1 after excluding participants who received bonus boosts in the first third of questions. Panel A shows the fraction of participants answering correctly at each position in the exam. The dashed line represents predicted values from a linear regression of performance on question position. Panel B displays the distribution of individual-level performance slopes (β_i^s) from regressions of correct answers on normalized question position, adjusted using an empirical Bayes shrinkage estimator. The sample excludes participants who received a bonus boost in the first third of the exam to address concerns that the mechanical reduction in incentives after question 10 could induce performance declines.

Figure A4: Comparison of Performance Decline Measures

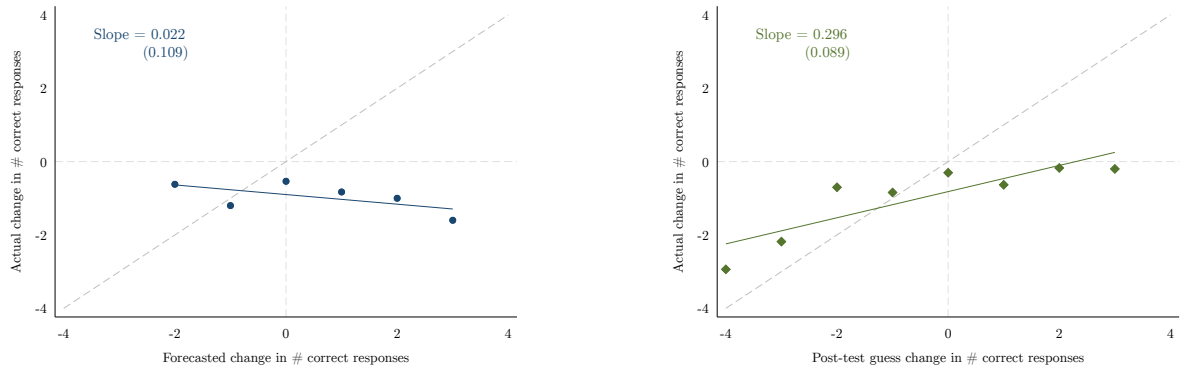


Notes: This figure presents a binned scatterplot comparing two measures of individual-level performance decline: (1) the simple difference in fraction correct between the last third (questions 21-30) and first third (questions 1-10) of the exam, shown on the x-axis, and (2) the empirical Bayes shrinkage estimate $\hat{\beta}_i$ from individual-level regressions of performance on normalized question position, shown on the y-axis. Each point represents mean values within 30 quantile bins of the simple difference measure. The solid line shows the best linear fit. The vertical and horizontal dashed lines at zero separate performance improvements (positive values) from declines (negative values).

Figure A5: Performance Change Forecasts vs. Actual Performance Changes

Panel A. Pre-Exam Forecast

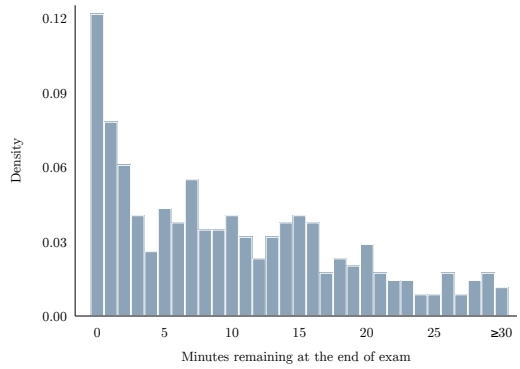
Panel B. Post-Exam Hindcast



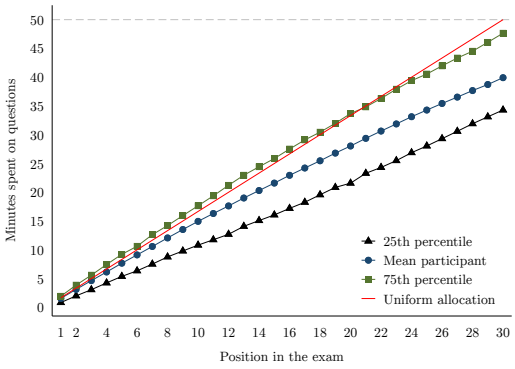
Notes: This figure shows binned scatter plots of participants' beliefs about performance changes against actual performance changes between the first and last thirds of the exam. Panel A compares pre-exam forecasted changes with actual changes in correct responses. Panel B compares post-exam hindcasted changes with actual changes. Each point represents the mean actual change for individuals with the same forecasted (Panel A) or hindcasted (Panel B) change value, excluding bins with fewer than five observations. The dashed 45-degree line indicates perfect prediction accuracy. The solid colored line shows the best linear fit between beliefs and actual changes, with slope coefficients and robust standard errors displayed.

Figure A6: Time Management in the Math Exam

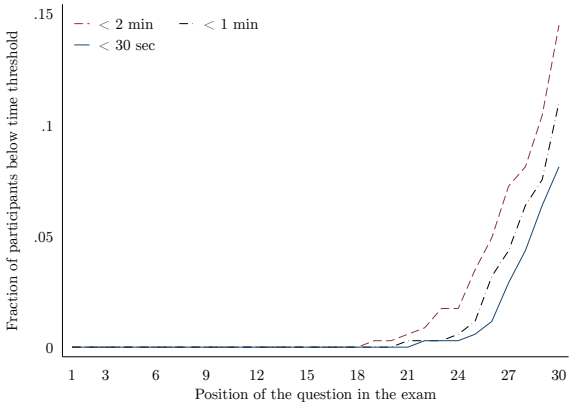
Panel A. Time Remaining at Exam Completion



Panel B. Cumulative Time Allocation

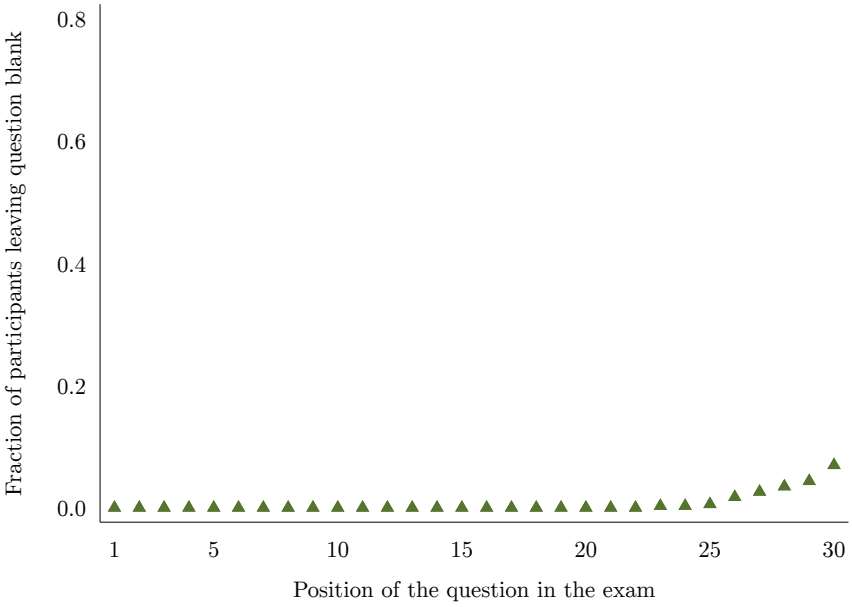


Panel C. Fraction of Participants Below Time Threshold



Notes: Panel A shows the distribution of time remaining when participants completed the exam in one-minute bins. Values are winsorized at 30 minutes. Panel B displays cumulative time spent across exam questions, showing the 25th percentile, mean, and 75th percentile of participants. The red line indicates uniform time allocation (50/30 minutes per question), and the horizontal dashed line marks the 50-minute time limit. Panel C shows the fraction of participants who start the nth question with less than two minutes, one minute, or 30 seconds remaining.

Figure A7: Unanswered Questions by Position in Exam



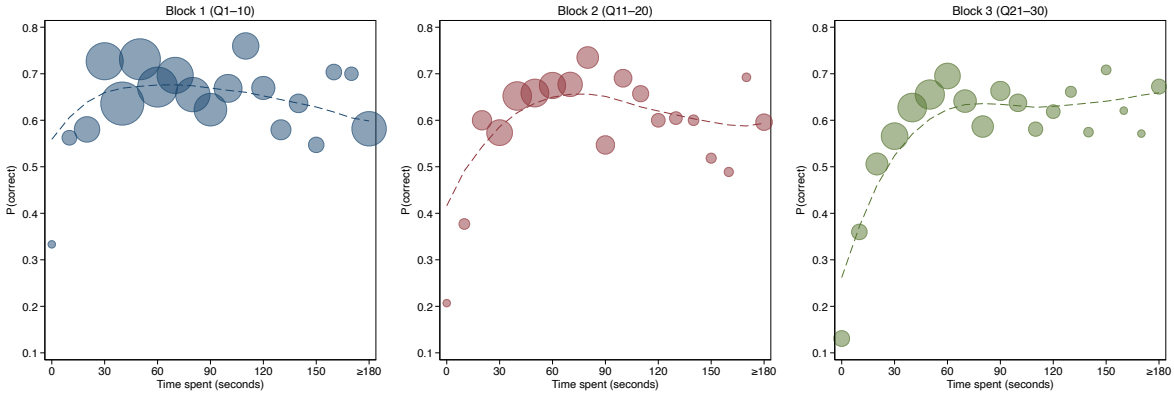
Notes: This figure shows the percentage of participants who left each question unanswered as a function of its position in the 30-question exam. Each point represents the fraction of missing responses for that question number.

Figure A8: Time Spent per Question: Correct vs. Incorrect Answers



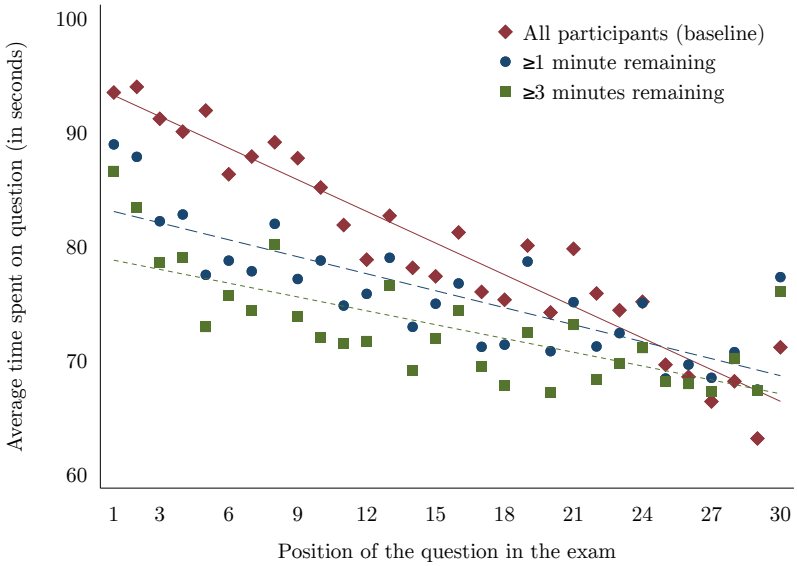
Notes: This figure shows the average time spent per question across the exam sequence, separately for correct responses (green triangles) and incorrect responses (red circles). Both series exhibit declines over the course of the exam. Dashed lines represent linear fits for each series.

Figure A9: Time–Performance Production Function by Exam Third



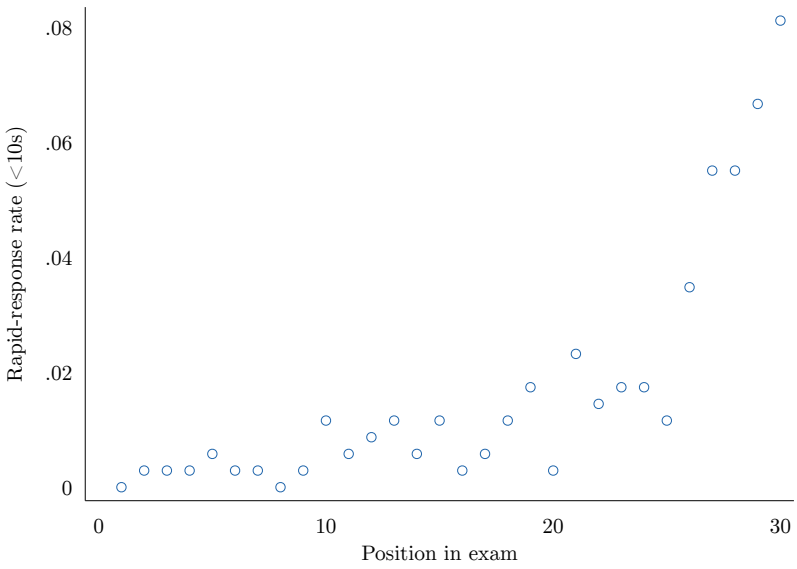
Notes: Each panel plots the fraction of questions answered correctly by 10-second time bins, separately for the first third (questions 1–10), middle third (questions 11–20), and final third (questions 21–30). Dot size is proportional to the number of observations in each bin. Dashed lines show weighted local-polynomial fits. Time is capped at 180 seconds. Question order is randomized, so systematic shifts across thirds are not driven by deterministic changes in question difficulty.

Figure A10: Time Decline Among Unconstrained Participants



Notes: This figure shows average time spent per question across the exam sequence for different subsamples. The red diamonds represent all participants (baseline). The blue circles exclude participants who finished with less than one minute remaining. The green squares exclude participants who finished with less than three minutes remaining. Dashed lines represent linear fits for each subsample.

Figure A11: Rapid Response Rate by Question



Notes: This figure shows the rapid response rate by question, where we define a rapid response as one taking less than 10 seconds. Participants who respond this quickly answer correctly approximately 20 percent of the time, equal to the chance rate on five-option questions. Rapid response rates remain below 3 percent for early questions but increase in the final five questions.

Table A1: Treatment Effects of Bonus Boosts

	Outcome					
	Forecast frac corr. (1)	Fraction correct (2)	Time Actual (3)	Time Optimal (4)	Effort Perceived (5)	Effort Optimal (6)
Panel A. Pooled Estimates						
=1 if boosted	0.017**	0.023	7.353***	12.068***	7.949***	11.051***
	(0.007)	(0.017)	(2.785)	(3.932)	(1.476)	(1.524)
Constant	0.554***	0.624***	78.018***	113.065***	72.409***	73.930***
	(0.002)	(0.004)	(0.686)	(0.969)	(0.364)	(0.375)
Panel B. Linear Heterogeneous Effects						
=1 if boosted	0.090*	0.097	−12.982	28.953	−14.014*	−1.869
	(0.047)	(0.101)	(18.744)	(33.526)	(7.955)	(8.625)
=1 if boosted × Exam Third	−0.029	−0.030	8.134	−6.754	8.785***	5.168
	(0.019)	(0.041)	(7.920)	(14.554)	(3.305)	(3.443)
Constant	0.551***	0.620***	78.950***	112.292***	73.415***	74.522***
	(0.002)	(0.006)	(0.864)	(1.134)	(0.441)	(0.523)
Panel C. Heterogeneous Effects by Exam Third						
=1 if boosted	0.010	0.005	−0.566	20.024***	9.811***	10.447***
	(0.012)	(0.026)	(4.176)	(6.520)	(2.247)	(2.057)
=1 if boosted × Middle	0.023	0.020	2.652	−5.254	−5.184*	−1.404
	(0.017)	(0.038)	(5.829)	(8.576)	(3.080)	(3.066)
=1 if boosted × Final	−0.010	0.008	10.980	−11.112	3.264	2.807
	(0.018)	(0.039)	(6.749)	(10.487)	(3.454)	(3.155)
Mean control	0.554	0.626	78.797	112.986	72.576	74.613
$p-$value: Middle = First	0.174	0.605	0.649	0.541	0.093	0.647
$p-$value: Final = First	0.592	0.842	0.105	0.290	0.345	0.374
$p-$value: Final = Middle	0.069	0.741	0.229	0.652	0.006	0.202
N	1,035	1,035	1,035	1,035	1,035	1,035
Participant FE	Yes	Yes	Yes	Yes	Yes	Yes
Pair-control FE	Yes	Yes	Yes	Yes	Yes	Yes
Third FE	Yes	Yes	Yes	Yes	Yes	Yes

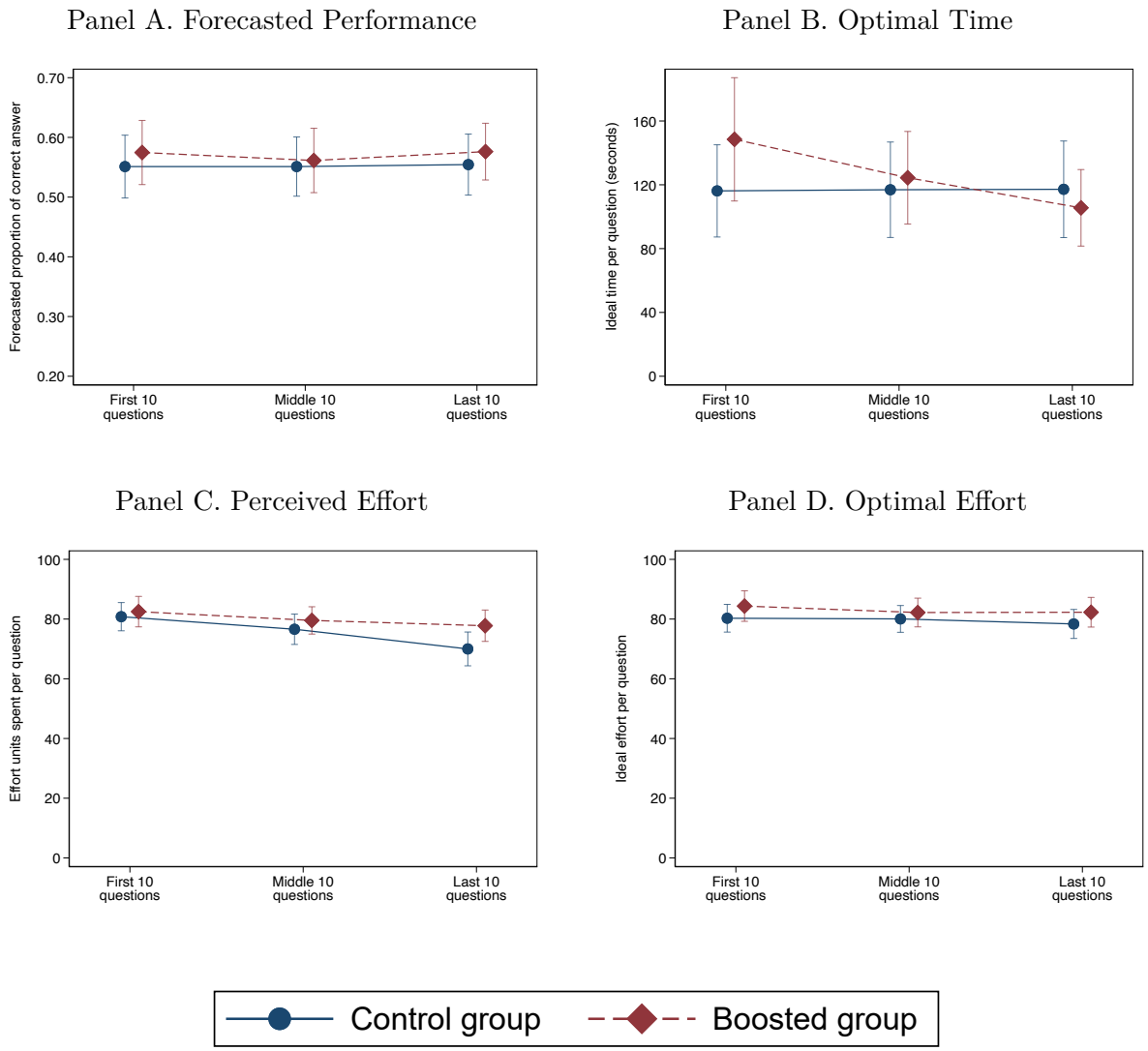
Notes: This table presents estimates of treatment effects from Equation 8. Panel A shows average effects where *Boosted* equals one if participant i receives bonus boosts in third t. Panel B includes interaction terms to test for heterogeneous effects across exam thirds, with the first third as the reference category. The coefficients on the interaction terms test whether treatment effects differ in the middle and final thirds relative to the first third. Bonus boosts increase payment from \$2 to \$10 for correctly answered questions. The unit of observation is participant-third. Standard errors are heteroskedasticity-robust and clustered at the participant level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A2: Production Function Shifts Across Exam Thirds

	Outcome: =1 if answer is correct
	(1)
Middle third (Q11–20)	−0.030**
	(0.012)
Final third (Q21–30)	−0.046***
	(0.012)
N (Question responses)	10,350
Time-bin FE?	Yes
Joint p-value	0.001

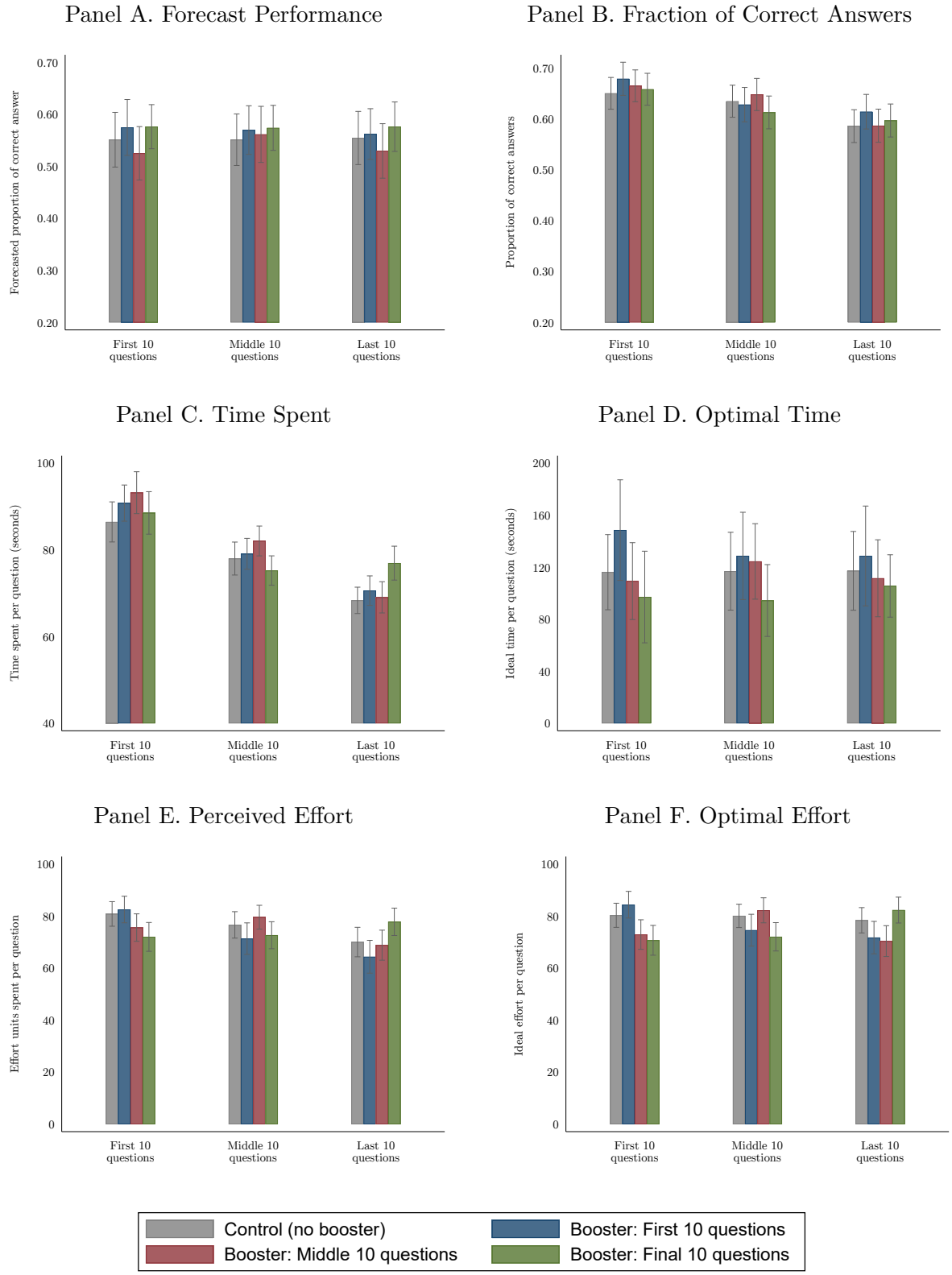
Notes: This table reports estimates from a question-level regression of correctness on indicators for the middle and final thirds of the exam, with the first third omitted. The specification includes 10-second time-bin fixed effects and clusters standard errors at the participant level. Coefficients therefore compare performance across exam thirds among responses with comparable time spent. Question order is randomized independently for each participant. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure A12: Boosted vs. Control Across Exam Thirds: Additional Outcomes



Notes: This figure reports additional outcomes from the pooled boosted–control comparison shown in main-text Figure 3. Control participants are shown as navy circles with solid lines; boosted participants (pooled across treatment arms, during their incentivized third) as maroon diamonds with dashed lines. Panel A shows forecasted performance, Panel B shows retrospectively reported ideal time allocation, Panel C shows perceived effort, and Panel D shows retrospectively reported ideal effort.

Figure A13: Mean Outcomes by Treatment and Exam Third



Notes: This figure shows the mean of the outcome listed in the panel title across exam thirds and for each treatment group. Error bars represent 95 percent confidence intervals. Panels A and B display forecasted and actual performance (fraction of correct answers) across exam thirds. Panels C and D show actual and retrospectively reported ideal time allocation. Panels E and F present perceived and retrospectively reported ideal effort levels. The control group (gray) received standard $2 bonuses for all questions. Treatment groups received bonus boosts increasing payment to $10 for 10 questions in their assigned third: First 10 (navy), Middle 10 (maroon), or Final 10 (green).

Table A3: Estimates of Performance Declines over the Exam: Robustness to Rapid-Response Exclusions

	Outcome: =1 if answer is correct				
	(1)	(2)	(3)	(4)	(5)
Q position (normalized)	−0.054***	−0.043***	−0.050***	−0.053***	−0.048***
	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)
Constant	0.657***	0.657***	0.661***	0.663***	0.666***
	(0.011)	(0.012)	(0.012)	(0.011)	(0.011)
N (Question responses)	10,350	9,090	9,990	10,179	9,945
Question FE?	Yes	Yes	Yes	Yes	Yes
Treatment FE?	Yes	Yes	Yes	Yes	Yes
Time-bin FE?	Yes	Yes	Yes	Yes	Yes
Exclusion	NA	End w/ >30 sec	Rapid Resp <4	<10 sec per Q	<20 sec per Q

Notes: This table presents estimates of β from equation (7). The dependent variable is an indicator for answering correctly. *Question position (normalized)* scales linearly from 0 (first) to 1 (last). In column 2 we restrict to those who finish the exam with more than 30 seconds; in column 3 we drop those who answer at least four questions in under 10 seconds; column 4 drops all answers that were given in under 10 seconds; column 5 drops all answers that were given in under 20 seconds. Standard errors are heteroskedasticity-robust and clustered at the participant level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A4: Spillover Effects: Performance in Non-Incentivized Thirds

	Outcome: Mean Performance in Non-Incentivized Thirds					
	Forecast frac corr. (1)	Fraction Correct (2)	Time Actual (3)	Time Ideal (4)	Effort Perceived (5)	Effort Ideal (6)
=1 if boosted	0.003	0.004	1.840	−5.754	−4.926*	−7.576***
	(0.028)	(0.021)	(2.248)	(17.670)	(2.861)	(2.784)
Constant	0.552***	0.623***	77.594***	116.748***	75.796***	79.567***
	(0.025)	(0.017)	(1.826)	(14.953)	(2.373)	(2.222)
Mean control	0.552	0.623	77.6	116.7	75.8	79.6
N	525	525	525	525	525	525

Notes: This table tests for spillover effects of bonus boosts on performance in non-incentivized thirds. The dependent variable in each column is the mean outcome across the two thirds where participant i did not receive enhanced incentives. For example, if a participant received boosts in the first third, the outcome is their mean performance in the middle and final thirds. The coefficient on *Boosted* tests whether receiving enhanced incentives in one third affects outcomes in other thirds. Pure control participants serve as the comparison group, where the outcome for controls in third t is also calculated as the mean across the two other thirds. This ensures that treated and control groups are compared using the same third combinations. The sample includes one observation per participant. Standard errors are heteroskedasticity-robust and clustered at the participant level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A5: Advice Categorization

	Overall	Treatment			
		Control	First	Middle	Final
Panel A. Single-Best Theme (%)					
Time management & pacing	40.0	42.2	36.7	38.6	42.0
Bonus/reward prioritization	4.6	0.0	10.1	3.4	5.7
Careful reading & extracting key info	22.3	24.4	25.3	18.2	21.6
Structured problem-solving	4.1	3.3	3.8	3.4	5.7
Use tools & external aids	4.9	1.1	5.1	8.0	5.7
Answer-choice strategy/guessing	3.2	3.3	6.3	2.3	1.1
Focus & testing conditions	8.4	5.6	5.1	13.6	9.1
Stress management & confidence	6.7	12.2	2.5	6.8	4.5
Check work & accuracy	1.4	3.3	0.0	2.3	0.0
Preparation & skill refresh	2.0	3.3	1.3	1.1	2.3
Non-substantive	2.3	1.1	3.8	2.3	2.3
Panel B. Multi-Label Theme Prevalence (%)					
Time management & pacing	51.9	56.7	57.0	43.2	51.1
Bonus/reward prioritization	15.9	0.0	21.5	17.0	26.1
Careful reading & extracting key info	36.2	35.6	32.9	39.8	36.4
Structured problem-solving	8.7	8.9	10.1	6.8	9.1
Use tools & external aids	20.6	14.4	21.5	29.5	17.0
Answer-choice strategy/guessing	17.4	21.1	17.7	10.2	20.5
Focus & testing conditions	16.5	13.3	13.9	21.6	17.0
Stress management & confidence	25.8	27.8	26.6	26.1	22.7
Check work & accuracy	8.1	10.0	10.1	8.0	4.5
Preparation & skill refresh	2.9	5.6	1.3	2.3	2.3
Non-substantive	1.4	0.0	1.3	2.3	2.3
N	345	90	79	88	88

Notes: This table reports the distribution of advice themes from open-ended responses collected after the exam. Panel A shows the single-best theme assigned to each response by LLM-assisted classification. Panel B shows the prevalence of each theme under multi-label coding, where each response may be assigned to multiple themes. Columns show percentages for the overall sample and by treatment arm (Control, First 10, Middle 10, Final 10 boosted questions). A chi-square test for equality of the single-best distribution across treatment arms yields $p = 0.108$.

Table A6: Second-Pass Advice Indicators

	Overall	Treatment			
		Control	First	Middle	Final
Trajectory Belief					
Declines over time	9.6	6.7	12.7	9.1	10.2
Constant over time	2.3	2.2	1.3	3.4	2.3
Improves over time	1.2	1.1	1.3	0.0	2.3
Trajectory Mechanism					
Time pressure	5.5	5.6	8.9	3.4	4.5
Cognitive fatigue	3.5	1.1	3.8	4.5	4.5
Warm-up / calibration	0.9	0.0	1.3	0.0	2.3
Effort Allocation					
Dynamic tradeoff reasoning	34.2	32.2	34.2	25.0	45.5
Uniform allocation	24.3	32.2	22.8	21.6	20.5
Front-load effort	8.4	10.0	15.2	2.3	6.8
Back-load effort	2.0	1.1	0.0	3.4	3.4
Adaptive (by difficulty)	13.9	13.3	13.9	14.8	13.6
Stakes-contingent	11.0	0.0	13.9	13.6	17.0
Flat path explicitly optimal	11.3	14.4	10.1	10.2	10.2
Bonus & Payment					
Recommend early placement	3.2	0.0	5.1	2.3	5.7
Recommend late placement	1.2	0.0	1.3	1.1	2.3
Timing irrelevant	0.9	0.0	1.3	0.0	2.3
Random selection understanding	3.8	2.2	1.3	1.1	10.2
Quantitative pacing rule	8.4	13.3	3.8	6.8	9.1
N	345	90	79	88	88

Notes: This table reports indicators from a second-pass LLM-assisted coding of the same open-ended advice responses. Each row shows the percentage of responses classified under the indicated category. Columns show percentages for the overall sample and by treatment arm.

B Empirical Appendix

B.1 Robustness (Exclusion Criteria)

We apply several exclusion criteria to ensure data quality and exogenous variation. The engagement screen excludes participants who left the tab frequently (on at least 10 questions) or for extended periods (more than 5 minutes), or who spent less than 20 minutes overall on the 30-question exam. The comprehension screen removes participants who failed multiple understanding checks. Finally, we exclude participants whose stated preference or MPL choice was implemented (i.e., non-random bonus timing), so that treatment timing remains exogenous in the main analysis.

Appendix Table B1 shows that the estimated performance decline is very similar across all exclusion variants, including models that apply each criterion in isolation and the specification with no exclusions at all. The distribution of bonus-boost timing choices and mean WTP by timing choice are likewise stable across samples (Tables B2 and B3), indicating that the timing-indifference result is not driven by any single exclusion rule. Thus, the core conclusions are robust to alternative sample definitions.

Table B1: Estimates of Performance Declines over the Exam: Robustness to Sample Exclusions

	Outcome: =1 if answer is correct				
	(1)	(2)	(3)	(4)	(5)
Q position (normalized)	−0.094***	−0.088***	−0.086***	−0.084***	−0.082***
	(0.017)	(0.017)	(0.016)	(0.017)	(0.016)
Constant	0.676***	0.672***	0.669***	0.661***	0.663***
	(0.012)	(0.011)	(0.011)	(0.012)	(0.011)
N (Question responses)	10,350	11,490	12,360	11,970	12,870
N (Participants)	345	424	453	399	477
Question FE?	Yes	Yes	Yes	Yes	Yes
Treatment FE?	Yes	Yes	Yes	Yes	Yes
Time-bin FE?	No	No	No	No	No
Exclusion	Main	Engage-ment	Compre-hension	Nonran-dom	No exclusions

Notes: This table reports robustness estimates of β from equation (7) using column (2) of Table 2 as baseline. The dependent variable is an indicator for answering correctly. *Question position (normalized)* scales linearly from 0 (first) to 1 (last). Column (1) reproduces the main analysis sample. Columns (2)–(4) include only a single exclusion criterion (engagement/attention, comprehension, or nonrandom assignment), while column (5) removes all exclusions. Standard errors are heteroskedasticity-robust and clustered at the participant level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B2: Distribution of Bonus Boost Choices by Exclusion Criteria

	(1)	(2)	(3)	(4)	(5)
First 10 questions (%)	34.5	30.2	32.3	34.4	30.8
	(3.0)	(2.5)	(2.5)	(2.7)	(2.4)
Middle 10 questions (%)	5.1	5.7	5.9	4.6	5.8
	(1.4)	(1.3)	(1.3)	(1.2)	(1.2)
Final 10 questions (%)	2.4	3.0	3.7	4.0	4.2
	(1.0)	(0.9)	(1.0)	(1.1)	(1.0)
No preference (%)	58.0	61.1	58.1	57.0	59.2
	(3.1)	(2.7)	(2.6)	(2.9)	(2.5)
N	255	334	356	302	380
Exclusion	Main	Engage-ment	Compre-hension	Nonran-dom	No exclusions

Notes: Each cell reports the share of participants selecting a timing option for the bonus boost in Decision 1, with standard errors in parentheses. Column (1) reproduces the main analysis sample. Columns (2)–(4) include only a single exclusion criterion (engagement/attention, comprehension, or nonrandom assignment), while column (5) removes all exclusions. The N row reports the sample size for each column.

Table B3: Willingness to Pay by Timing Choice: Robustness to Exclusion Criteria

	(1)	(2)	(3)	(4)	(5)
WTP: First 10 (USD)	2.03	1.96	2.05	2.09	2.02
	(0.09)	(0.08)	(0.08)	(0.09)	(0.08)
WTP: Middle 10 (USD)	2.20	2.18	2.17	2.10	2.13
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)
WTP: Final 10 (USD)	2.06	2.04	2.13	2.02	2.06
	(0.10)	(0.09)	(0.09)	(0.10)	(0.09)
N	255	334	356	302	380
Exclusion	Main	Engage-ment	Compre-hension	Nonran-dom	No exclusions

Notes: Each cell reports mean WTP (USD) from Decision 2 for the elicited timing option, with standard errors in parentheses. Column (1) reproduces the main analysis sample. Columns (2)–(4) include only a single exclusion criterion (engagement/attention, comprehension, or nonrandom assignment), while column (5) removes all exclusions. The N row reports the sample size for each column.

Table B4: Boosted Payment Choice Reasons

	Timing Choice				
	(1) Overall	(2) No Pref.	(3) First	(4) Middle	(5) Last
No preference / indifference	38.0	65.5	0.0	0.0	0.0
Probability / expected-value reasoning	32.5	52.0	3.4	23.1	0.0
Start-of-exam advantage (fresh)	25.9	1.4	72.7	0.0	0.0
Decision aversion / defaulting	23.5	39.2	2.3	0.0	0.0
Fatigue / declining performance	22.7	2.0	56.8	30.8	16.7
Perceived difficulty varies by position	8.6	4.7	14.8	7.7	16.7
Motivation / consistent effort	7.8	11.5	2.3	7.7	0.0
Self-assessed ability	7.5	6.1	8.0	23.1	0.0
Reduce stress / get it out of the way	7.1	6.1	9.1	7.7	0.0
Maximize bonus (general)	6.3	2.0	13.6	7.7	0.0
Middle-of-exam advantage (warm-up)	5.1	0.0	1.1	76.9	33.3
End-of-exam advantage (finish strong)	2.0	0.0	0.0	0.0	83.3
Present bias / reward sooner	1.2	0.0	3.4	0.0	0.0
N	255	148	88	13	6

Notes: This table summarizes participants' open-ended explanations for their preferred bonus-boost timing. We first used a large language model (OpenAI's GPT-5) to propose a taxonomy of response categories and then used GPT-5 mini to assign one or more categories to each response in a multi-label scheme (rows sum to more than 100 percent because many responses invoke multiple reasons). Categories with fewer than 0.5 percent of overall responses are omitted. Column 1 reports shares among all bonus-boost participants in this sample with classifiable explanations; columns 2–5 report shares by timing choice. The probability / expected-value row reports the raw category share; the text uses a cleaned no-preference statistic that excludes responses also tagged as start advantage, warm-up, finish strong, fatigue, or perceived difficulty by position.

Table B5: No-Preference Choice and Forecasted Performance Change

	Outcome: =1 if chose no preference	
	(1)	(2)
Expected decline	-0.269***	-0.258***
	(0.079)	(0.080)
Expected improvement	-0.089	-0.064
	(0.082)	(0.087)
First-final flat, not exact	-0.131	-0.098
	(0.107)	(0.114)
Omitted group	Exact flat	Exact flat
Controls	No	Yes
N	255	255

Notes: OLS regression of an indicator for choosing no preference on indicators for the participant's forecasted first-to-final-third change in correct answers. The omitted category is participants whose forecasted change is exactly zero. "First-final flat, not exact" indicates participants whose first- and final-third forecasts are equal but whose middle-third forecast differs. Column 1 reports the regression without controls; column 2 adds the participant's shrunk realized-performance slope from Section 4, total correct answers, age, gender, race, and education. Robust standard errors in parentheses. Sample: bonus-boost participants who passed the data-quality screens and whose bonus-boost timing was assigned at random. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B.2 Shrinkage Estimates of Performance Trajectory

Our analysis of individual performance trajectories faces a fundamental inference problem. Each participant answers only 30 questions, yielding individual slope coefficients that could be measured with substantial noise. Any measurement error would generate spurious heterogeneity: participants may exhibit extreme performance improvements or declines that reflect sampling variation rather than true differences in performance declines.

We address this challenge using an empirical Bayes shrinkage estimator that optimally combines individual-level estimates with population patterns. This approach recognizes that extreme individual estimates likely reflect noise rather than true heterogeneity.

The shrunk estimator β_i^s takes the form:

$$\beta_i^s = \omega_i \hat{\beta}_i + (1 - \omega_i)\bar{\beta}, \tag{B1}$$

where $\hat{\beta}_i$ is the individual estimate, $\bar{\beta}$ is the population mean, and the weight ω_i captures the signal-to-noise ratio in participant i's estimate.

The optimal weight balances the precision of the individual estimate against the extent of true population heterogeneity:

$$\omega_i = \frac{\text{Var}[\hat{\beta}_i] - \mathbb{E}[\text{SE}^2_{\hat{\beta}_i}]}{\text{Var}[\hat{\beta}_i] - \mathbb{E}[\text{SE}^2_{\hat{\beta}_i}] + \text{SE}^2_{\hat{\beta}_i}}. \tag{B2}$$

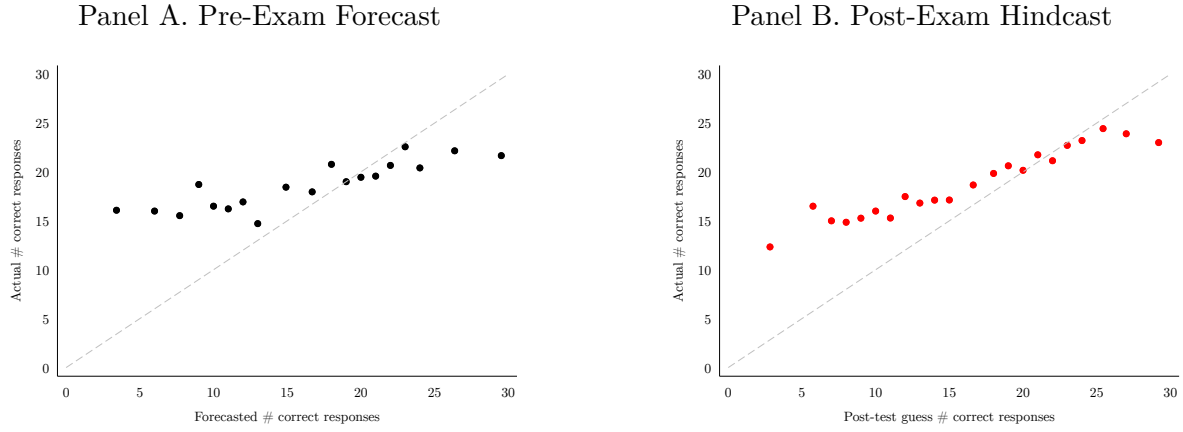
The numerator represents the variance of true performance trajectories (total variance minus average measurement error), while the denominator adds participant i's specific measurement error variance. This formulation ensures that precisely estimated trajectories receive minimal shrinkage toward the mean, while noisy estimates are substantially shrunk. When population heterogeneity is limited relative to measurement error, all estimates converge toward the common mean.

B.3 Overall Performance Forecasts

We elicited participants' beliefs about their performance at two points: before taking the exam (forecasts) and immediately after completion (hindcasts). Appendix Figure B1 compares these beliefs with actual performance using binned scatter plots. Panel A reveals that participants systematically underestimate their performance before taking the exam. They forecast answering only 16.8 questions correctly (55.8 percent), substantially below the actual mean of 18.9 questions (62.9 percent)—an underestimation of 2 correct questions. The binned averages cluster above the 45-degree line across most of the distribution, indicating that underestimation occurs at most levels of predicted performance.

Panel B shows that this pessimism remains after taking the exam. Participants' hindcasts average only 16.0 correct answers (53.2 percent), widening the gap between beliefs and actual performance to 2.92 questions. The standard deviation of hindcast errors (5.70) is smaller than that of forecast errors (6.99), suggesting that the exam experience leads to more systematic but less variable underestimation. The binned averages in Panel B remain consistently above the 45-degree line, with a tighter clustering pattern than in Panel A.

Figure B1: Performance Forecasts vs. Actual Performance



Notes: This figure shows binned scatter plots of participants' beliefs about their performance against actual performance. Panel A compares pre-exam forecasts with actual correct responses. Panel B compares post-exam hindcasts with actual correct responses. Each point represents the mean within bins of forecasted (Panel A) or hindcast (Panel B) values, with 30 bins total. The dashed 45-degree line indicates perfect accuracy. Points below the line indicate underestimation of performance.

B.4 Heterogeneity Analyses

Pre-registered heterogeneity. We pre-registered tests for heterogeneity in performance decline rates across demographic groups. Specifically, we estimate question-level accuracy as a function of normalized question position, including fixed effects for question, time-spent bins, and treatment assignment, and interact position with age, gender, and above-average education (BA+). We find no evidence that decline slopes differ systematically across these pre-registered covariates.

Exploratory heterogeneity. We also use regression forests to predict participant-specific decline slopes from pre-treatment covariates. Out-of-bag predictive performance is essentially zero, suggesting limited detectable heterogeneity in who declines more rapidly. The variables most frequently ranked as important include screen dimensions (a proxy for device type), age, gender, and start time. However, binned device-type comparisons and formal interaction tests reveal no robust differences.

Heterogeneous incentive effects. Finally, we estimate heterogeneous treatment effects of the incentive boost (bonus pay in a given third) using causal forests with pre-treatment covariates only. We find little evidence of systematic treatment effect heterogeneity beyond the small average effects documented in the main text.

B.5 Testing for Disengagement

The observed performance declines could result from participants giving up and guessing randomly due to time pressure, fatigue, or boredom. We consider time pressure and fatigue to be mechanisms through which performance declines operate rather than confounds. The remaining concern is that participants may disengage from the task entirely and resort to random guessing out of boredom. We find no evidence that this behavior drives our results.

To identify likely random guesses, we use response time. Figure 5, Panel B shows that participants who spend less than 10 seconds on a question answer correctly approximately 20 percent of the time—exactly the performance expected from random guessing on a five-option multiple choice question. In contrast, participants who spend 10–19 seconds perform significantly better than chance. We therefore classify responses under 10 seconds as "rapid responses" and treat them as proxies for random guesses.

Rapid responses are concentrated in later questions. Appendix Figure A11 shows that rapid response rates remain below 3 percent for the first 25 questions but increase substantially in the final five questions, reaching approximately 8 percent for question 30. Despite this pattern, excluding rapid responses from our analysis does not substantially alter the estimated performance decline after controlling flexibly for time spent on each question (Appendix Table A3).

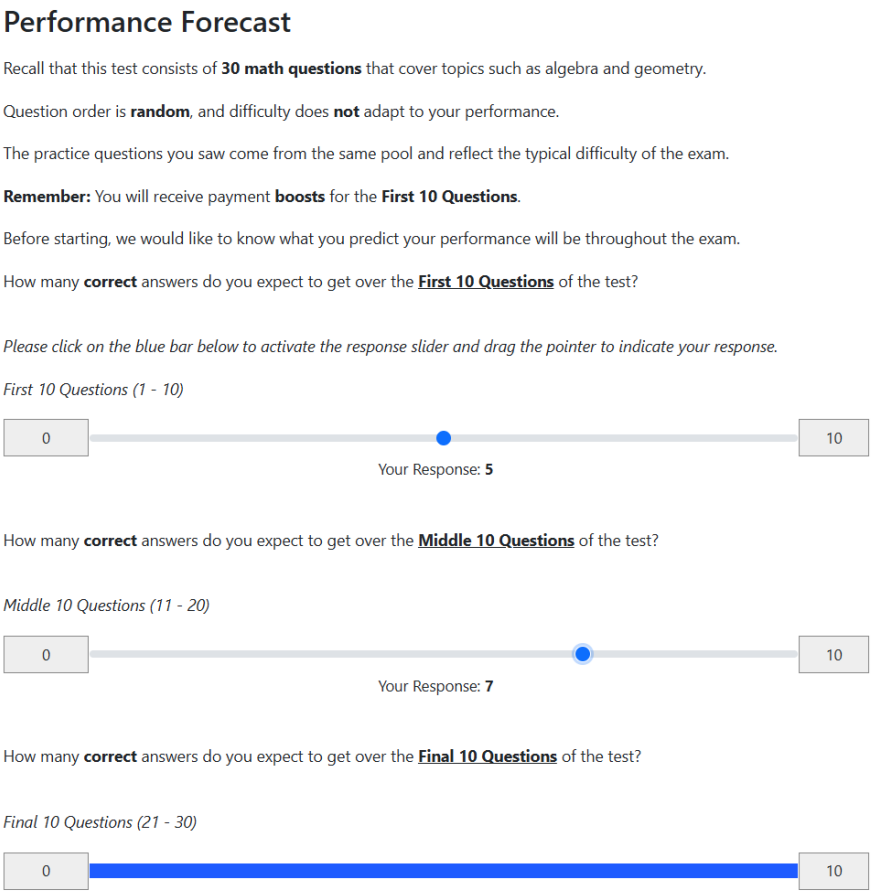
Most participants do not engage in systematic guessing. Over 75 percent of participants spend more than 10 seconds on all 30 questions. When we exclude participants in the 95th percentile of rapid response rates—those giving rapid responses on at least four questions—the position coefficient remains largely unchanged after controlling flexibly for response time (Appendix Table A3).

Time pressure explains only some of the rapid responses we observe. Among the 12 participants who responded within 10 seconds on at least four questions, only half finished with less than 30 seconds remaining, indicating severe time pressure. The remaining high-guessing participants had more time remaining on average (over 12 minutes) than the rest of the sample. Overall, only 12 percent of participants finish with less than 30 seconds remaining, and only 8 percent begin the final question with less than 30 seconds left. This indicates that most participants are not forced into rapid responses by time constraints. Even after excluding participants who finished with less than 30 seconds and controlling flexibly for response time, the position coefficient remains approximately half the baseline magnitude and is highly statistically significant (Appendix Table A3).

In sum, random guesses are relatively rare and appear to be driven primarily by time pressure, suggesting that some participants managed their time poorly rather than systematically disengaging from the task. A small subset of participants rushes through multiple questions without apparent time constraints, but the robustness checks in Appendix Table A3 demonstrate that these individuals do not drive our main results.

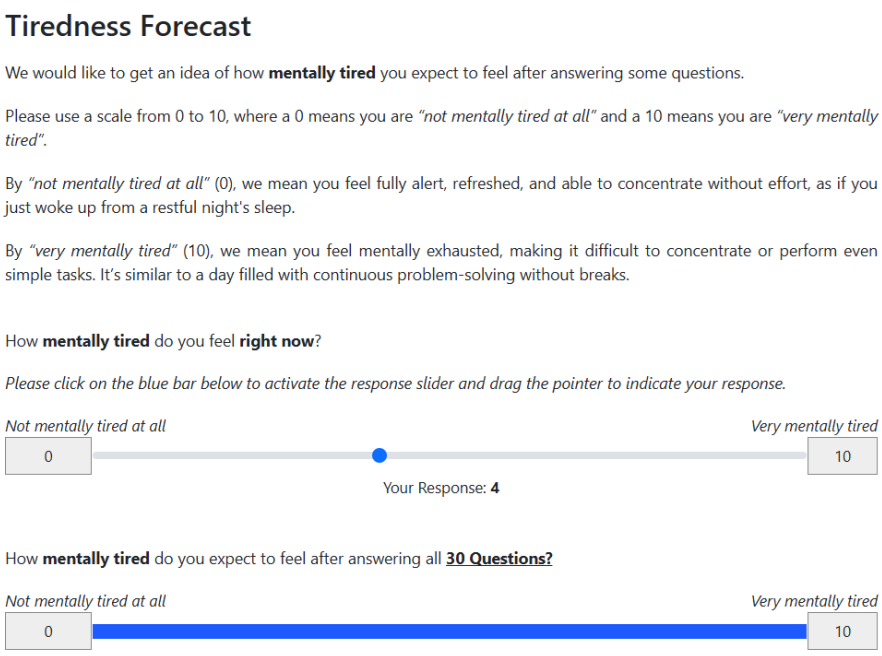
B.6 Experimental Instruments

Figure B2: Pre-Exam Performance Forecast Interface



Notes: This figure shows the interface for eliciting performance forecasts before the exam. Participants use interactive sliders to indicate how many questions (0–10) they expect to answer correctly in each third of the exam. The screen reminds participants that the test consists of 30 math questions covering algebra and geometry, that question order is random, and that difficulty does not adapt to performance. The example shown is for a participant assigned to receive payment boosts for the First 10 Questions.

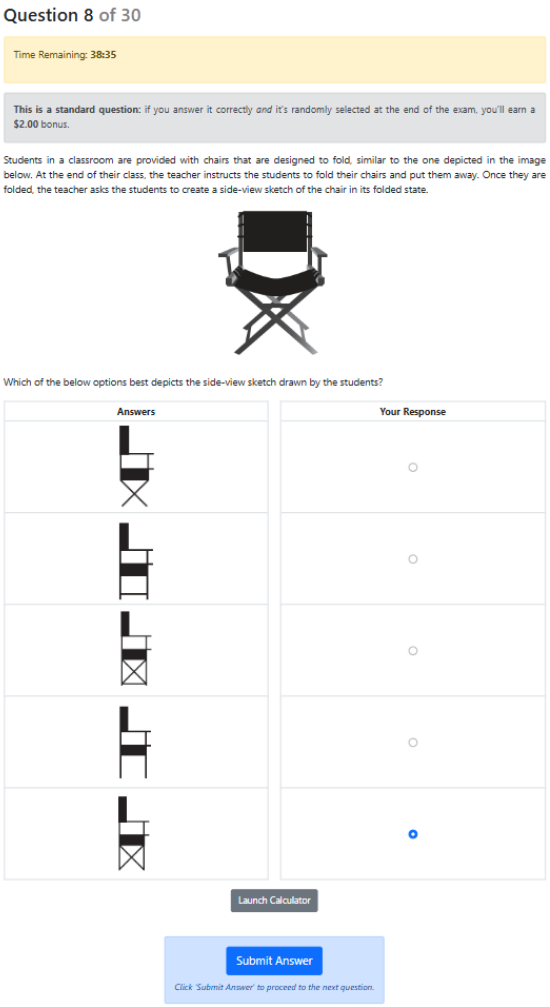
Figure B3: Mental Fatigue Elicitation Interface



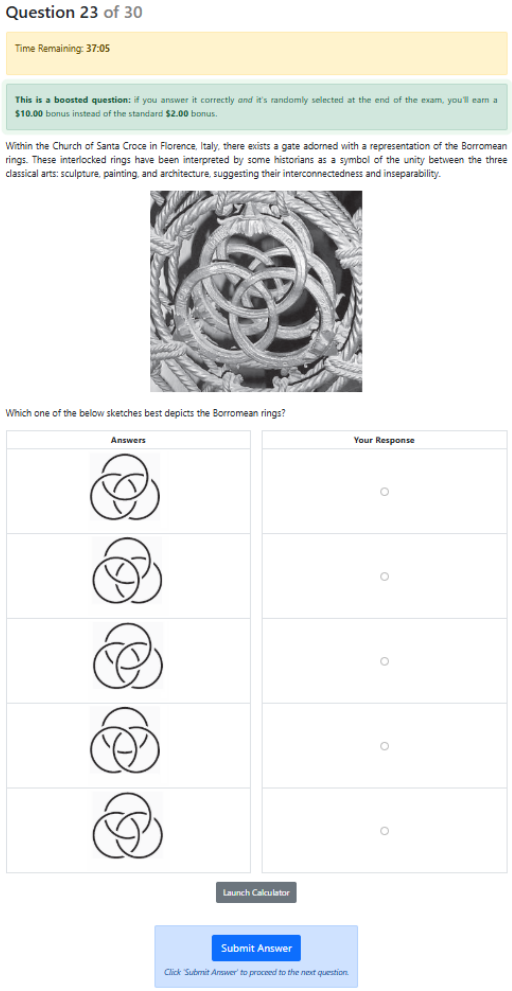
Notes: This figure shows the mental fatigue elicitation screen. Participants report their current mental tiredness and predict their tiredness after completing all 30 questions using a 0–10 scale with interactive sliders. The scale anchors define 0 as "not mentally tired at all" (fully alert, refreshed, and able to concentrate without effort) and 10 as "very mentally tired" (mentally exhausted, making concentration difficult, similar to continuous problem-solving without breaks).

Figure B4: Example Mathematics Exam Questions

Panel A. Standard Question ($2 Bonus)



Panel B. Boosted Question ($10 Bonus)



Notes: This figure shows the exam interface for mathematics questions drawn from Brazil's national college admission exam. Panel A displays a question offering a $2 bonus if correctly answered and randomly selected. Panel B displays a boosted question with the green banner notification indicating a $10 bonus if correctly answered and randomly selected. Both panels show the time remaining, question text with visual elements, multiple-choice options, a calculator launch button, and the "Submit Answer" button. Participants answer questions sequentially and cannot return to previous questions.