# Understanding Performance Declines*

Carl Meyer Germán Reyes Jason Somerville

March 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 is constant 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 roughly half of the decline, and suggestive evidence indicates that failing to anticipate fatigue leads participants to expend too much effort early. Participants view this as suboptimal: when asked about optimal strategies, they report that effort should be 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: Stanford University, email: cmeyer20@stanford.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, and Dmitry Taubinsky and seminar participants at Middlebury College, the 2025 Economic Science Association North American Meeting, the 2025 Bay Area Behavioral and Experimental Economics Workshop, and UC Santa Barbara. 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., 2024), 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., 2024). 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.

To assess whether these belief errors translate into costly mistakes, we examine participants' preferences over the timing of bonus boosts. Despite sizable performance declines, 58.0 percent of participants report no preference—delegating timing to chance—and only 34.5 percent choose early boosts. Open-text justifications confirm that most participants view timing as irrelevant. Consistent with this, WTP is independent of boost timing, averaging around $2.03 with no significant differences across thirds. But boost timing has real consequences: assigning boosts to the first rather than last third raises expected bonus earnings by 9 percent.

Why does performance decline? Using time investment as a proxy for effort, we find that falling effort accounts for roughly half the decline: most participants exhibit declining time trajectories, even though they finish with substantial time remaining. The remaining decline is consistent with fatigue reducing the productivity of effort—participants not only spend less time on later questions but derive less benefit from each unit of time. Participants who underestimate their fatigue show steeper declines in both effort and performance. Importantly, participants recognize their declining effort ex post, yet report that optimal effort allocation would be flat—consistent with misoptimization rather than deliberate pacing. Similarly, in incentivized open-ended advice to future test-takers, participants overwhelmingly emphasize generic heuristics (pacing and careful reading) and rarely mention within-exam performance decline or cognitive fatigue.

Our paper contributes to a literature on within-exam performance declines and their consequences for educational measurement. Researchers have documented substantial performance declines on low-stakes assessments like the PISA (Borghans and Schils, 2018; Zamarro et al., 2019; Balart and Oosterveen, 2019). Similar declines appear in high-stakes college entrance exams, where they affect score distributions and student-college sorting (Ackerman and Kanfer, 2009; Reyes, 2023). Recent experimental work shows that children's cognitive endurance can be trained, leading to less pronounced declines (Brown et al., 2024). We complement this human-capital perspective by providing the first experimental evidence that within-exam declines partly reflect optimization failures. This distinction matters for policy: if declines reflect not only limited endurance but also failures in effort allocation, then light-touch interventions—such as shorter exams, strategically timed breaks, or feedback on performance patterns—may reduce achievement gaps without requiring investments in endurance-building. Beyond standardized testing, organizations

could schedule high-stakes tasks earlier or add review steps for late-shift decisions; restructuring task environments around predictable declines—and individuals' naivete about them—may yield productivity gains without costly human capital investments.

Our research also speaks to the literature on inequality in standardized test performance (Rothstein, 2004; Ackerman and Kanfer, 2009; Bettinger et al., 2013; Hoxby et al., 2013; Bulman, 2015; Goodman, 2016; Goodman et al., 2020; Riehl, 2024), which documents substantial disparities in college admission exam scores by race and socioeconomic status. We propose a novel mechanism that may contribute to these gaps: systematic differences in how students from different backgrounds allocate effort over time. High-income students have greater access to resources such as tutoring and practice test-taking, which could help them better understand their own performance trajectories. These resources may also provide specific strategies that help manage fatigue—such as ways to prioritize questions or pace effort during long exams. In contrast, students without access to these resources may be more susceptible to performance declines. Our findings suggest that test score gaps may not only reflect differences in academic preparedness but also disparities in optimization strategies and fatigue management. More generally, our results imply that a given test score—and the gaps between demographic groups—should be interpreted in the context of exam design, since features like test length and question ordering can affect the scope for effort misallocation and thus the magnitude of score disparities.

More broadly, our findings contribute to a literature on within-person performance declines that cuts across 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). We provide evidence for a novel explanation rooted in optimization failures: participants fail to anticipate their performance decline, display indifference over incentive timing despite clear payoff consequences, and recognize ex post that their effort allocation was suboptimal.

## 2 Experimental Design

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

### 2.1 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.[1] 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 external resources including the internet, textbooks, or outside assistance. We warn that they must remain on the exam screen throughout and that navigating away, switching tabs, or stepping away may result in disqualification. To distinguish effort-driven performance changes from variation in question difficulty, we independently randomize question order for each participant.

**Bonus Payment Structure.** Participants receive a $12 participation payment plus performance bonuses. We randomly assign participants to either a control group (with probability 4/19) where all 30 questions offer $2 bonuses, or a "bonus boost" treatment (with probability 15/19) where 10 questions offer $10 bonuses and 20 offer $2. At the end of the exam, we randomly select one question. If participants answered the selected question correctly, they earn the corresponding bonus; incorrect answers earn nothing. This structure incentivizes effort on all questions since participants have no control over which determines their bonus payment.

**Bonus Boost Timing and Preference Elicitation.** Participants in the bonus boost treatment complete two incentivized decisions about when they would like to receive boosted questions. 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. If participants are aware of performance decline, they should optimally choose early boosts. 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 between receiving boosted bonuses for a randomly assigned timing option or guaranteed cash amounts ranging from $0.10 to $3.00. The MPL allows us to measure the strength of timing preferences—even if participants are nearly indifferent, direct choices in Decision 1 could still generate pronounced patterns favoring one option.

After participants submit both decisions, we determine their treatment assignment

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

through a random draw: with 10 percent probability we implement Decision 1, with 10 percent probability we implement a randomly selected row from Decision 2, and with 80 percent probability we randomly assign timing with equal probability across the three options. This procedure ensures incentive compatibility and generates exogenous variation in boost timing, allowing clean identification of its impact on performance.

**Practice Questions and Comprehension Checks.** Following treatment assignment, participants complete four practice questions drawn from the exam pool to familiarize them with the format and difficulty. Two practice questions display the boost banner. Participants then answer four comprehension checks on exam rules, bonus mechanics, and their assignment before proceeding. These questions reinforce the key aspects of the design and serve as exclusion criteria for highly inattentive participants in our analysis (see Section 2.2).

**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 their tiredness after completing all 30 questions using the same 0–10 scale, with anchors ranging from "not mentally tired at all" to "very mentally tired" (see Appendix Figure B3 for exact wording).

**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. 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 indicate how they would ideally allocate effort and time if retaking the exam. Finally, participants provide advice for future exam takers, incentivized by a potential bonus if a randomly selected future participant who receives their advice answers their bonus question correctly. The survey concludes with a demographic questionnaire.

### 2.2 Sample and Summary Statistics

We recruit 475 participants through Prolific, restricting to U.S.-based adults with high approval rates (minimum 200 prior approvals, 99 percent approval rate). The sample is gender-balanced by design.

We apply several pre-registered exclusion criteria to ensure data quality. During the exam, participants are disqualified if they show clear disengagement on the first seven questions (averaging under 20 seconds per question 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. We also exclude participants whose stated preference or MPL choice was implemented (the 20 percent for whom Decisions 1 or 2 determined boost timing) to maintain exogenous variation in timing. Appendix B.1 shows results are robust to including these observations.

Table 1 presents summary statistics. Participants average 41.3 years old; 49.1 percent are male and 66.4 percent are white. The sample is well-educated, with 35.1 percent holding bachelor's degrees. Participants attempt 29.9 questions on average, spend 40.3 minutes on the exam, and answer 62.9 percent correctly, with substantial heterogeneity across participants (Appendix Figure A1).

## 3 Performance Trajectories and Beliefs

This section documents performance declines over the exam and examines whether participants anticipate these patterns ex ante and update their beliefs ex post.

### 3.1 Performance Over the Course of the Exam

Figure 1, Panel A plots the fraction of participants answering correctly at each question position. Performance deteriorates substantially: 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.

To quantify the performance decline more systematically, Table 2 presents regression estimates from:

$$\text{Correct}_{ij} = \alpha + \beta \cdot \text{NormPos}_{ij} + \gamma_j + \varepsilon_{ij}, \tag{1}$$

where $\text{Correct}_{ij}$ indicates whether participant $i$ answered question $j$ correctly, $\text{NormPos}_{ij}$ is position normalized to range from 0 (first question) to 1 (last question), and $\gamma_j$ are question fixed effects. Standard errors are clustered at the participant level.

The coefficient $\hat{\beta} = -0.095$ (column 1) indicates that performance declines by 9.5 pp over the exam ($p < 0.01$), a 14.0 percent decrease from initial performance. This estimate is unchanged with question and treatment fixed effects (column 2). The magnitude of the decline is economically meaningful: evaluated at the first question, our estimate implies a participant would rank in the 57th percentile of our sample; evaluated at the last question, it implies they would rank in the 39th percentile—a 18 percentile point drop.

To examine if there is heterogeneity in performance trajectories across individuals, we estimate person-specific slopes:

$$\text{Correct}_{ij} = \alpha_i + \beta_i \cdot \text{NormPos}_{ij} + \varepsilon_{ij} \quad \text{for } i = 1, \ldots, N, \tag{2}$$

applying a shrinkage estimator to address measurement error in individual-level estimates (see Appendix B.2). Figure 1, Panel B shows the distribution of these adjusted slopes. After accounting for measurement error, 82 percent of participants exhibit declining performance ($\hat{\beta}_i < 0$). The standard deviation of individual-level declines is $\hat{\sigma}_\beta = 0.124$, suggesting substantial heterogeneity in the magnitude of the decline across participants.[2]

This analysis establishes our first result:

**Result 1.** *Performance declines substantially over the exam on average across participants and individually for most participants.*

### 3.2 Beliefs About Performance Trajectories

We next examine whether participants anticipate performance declines ex ante or recognize them ex post. Prospective accuracy matters because misperceptions could lead to suboptimal planning; retrospective accuracy matters for learning.[3] Before the exam, participants predicted correct answers in each third ("forecasts"); after, they estimated actual performance in each third ("hindcasts"). From these, we construct implied beliefs about

[2] Appendix Figure A2 shows that these results are virtually identical when we exclude participants who received bonus boosts in the first third of questions, which could mechanically induce a performance decline due to the reduction in incentives after question 10.

[3] Appendix B.3 analyzes forecasts of overall performance; here we focus on beliefs about performance changes between exam beginning and end.

performance changes.[4]

Figure 2 compares beliefs to actual performance. Panel A shows that participants forecast constant performance throughout the exam, predicting 55.6 percent correct in both the first and last thirds. In reality, performance drops from 66.2 to 59.5 percent—a decline of 6.7 pp. Post-exam hindcasts show only partial recognition: participants report a 3.7 pp decline, capturing roughly 55 percent of the actual drop.

Panel B shows that misperceptions are pervasive. The distribution of forecasted changes is concentrated at zero, with both the median and interquartile range at no change. While 50.1 percent of participants actually experience a decline, only 22.3 percent forecast one. Hindcasts show modest updating, with the proportion recognizing a decline increasing to 36.8 percent. The most common error is experiencing a decline while forecasting constant performance—a pattern that persists even in hindcasts (Panels C and D).

To formally test belief accuracy, Table 3 regresses actual performance changes on forecasted and hindcasted changes:

$$\text{ActualChange}_i = \gamma_1 + \delta_1 \cdot \text{ForecastChange}_i + \eta_i \tag{3}$$

$$\text{ActualChange}_i = \gamma_2 + \delta_2 \cdot \text{HindcastChange}_i + \mu_i \tag{4}$$

where $\text{ActualChange}_i$ is the difference in correct answers between the last and first thirds, and $\text{ForecastChange}_i$ and $\text{HindcastChange}_i$ are the corresponding predicted and retrospective beliefs. The constants $\gamma_1$ and $\gamma_2$ measure mean actual change when participants believe performance will remain constant. The slopes $\delta_1$ and $\delta_2$ measure how actual changes relate to beliefs: a one-question increase in forecasted (or hindcasted) change predicts a $\delta_1$ (or $\delta_2$) question change in actual performance. Accurate beliefs imply $\gamma_1 = \gamma_2 = 0$ (no bias) and $\delta_1 = \delta_2 = 1$ (perfect calibration).

Pre-exam forecasts have no predictive power: $\hat{\delta}_1 = 0.02$ ($p = 0.84$) is indistinguishable from zero, far below $\delta_1 = 1$ (see Appendix Figure A4 for visual evidence). The constant $\hat{\gamma}_1 = -0.67$ reveals systematic bias: participants forecasting no change actually decline by 0.67 questions on average, rejecting $\gamma_1 = 0$. Post-exam hindcasts perform better but remain miscalibrated. While $\hat{\delta}_2 = 0.30$ ($p < 0.01$) differs from zero, it falls well short of $\delta_2 = 1$—beliefs are systematically compressed, with each unit of actual change translating to

[4]For this analysis, we measure performance decline as the difference between last-third and first-third fraction correct. This differs from our preferred measure (the $\hat{\beta}_i$ estimates from our shrinkage estimator) which uses all questions. We use this simpler measure because it corresponds directly to our belief elicitation. Reassuringly, the two measures are highly correlated ($\rho = 0.92$), as shown in Appendix Figure A3.

less than half a unit of perceived change. The constant $\hat{\gamma}_2 = -0.56$ again shows significant bias, rejecting $\gamma_2 = 0$. This establishes our second result:

**Result 2.** *Participants fail to anticipate performance declines and only partially recognize them after the exam.*

This combination of ex ante naivete and ex post imperfect learning suggests that misperceptions about performance trajectories may lead to suboptimal decisions—a possibility we examine next.

### 3.3 Comparing Actual and Ideal Allocation of Effort

The preceding results show that participants fail to anticipate performance declines. This raises the possibility that they also misallocate effort. To assess this, we ask participants to report both their perceived allocation during the exam and their ideal allocation if they could retake it.

Figure 3 reveals a disconnect between actual behavior and stated ideals. 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 flat, declining by only 2 seconds across sections. The pattern for self-reported effort is analogous: perceived effort declines substantially, but stated optimal effort remains constant at approximately 76.9 units throughout. Participants thus recognize that their actual allocation deviated from what they believe would have been optimal.

This preference for steady effort is also reflected in participants' open-ended advice to another participant: pacing is the dominant theme, but only a small minority recommend a specific intertemporal pacing rule (see Section 4.4).

**Result 3.** *Participants recognize ex post that their effort allocation was suboptimal: they accurately perceive declining time and effort, yet report that optimal allocation would be constant.*

## 4 The Allocation of Bonus Boosts

Participants fail to anticipate performance declines and recognize afterward that their allocation was suboptimal. We next examine whether this naivete translates into costly mistakes by analyzing how participants allocate bonus boosts across the exam. Given the per-

formance declines documented above, participants should prefer early bonus placement—when their probability of answering correctly is highest—unless higher incentives fully offset the decline (a possibility we later reject). By eliciting both stated preferences and WTP for different timing options, we test whether participants optimally allocate higher-value tasks.

### 4.1 Bonus Boost Preferences

**Direct Choice.** Figure 4, Panel A shows the distribution of timing preferences. Despite the documented performance declines, only 34.5 percent of participants chose to boost the first third of questions. The modal response was indifference: 58.0 percent chose to let timing be randomly determined rather than selecting a specific portion.

To understand these choices, we analyze open-ended explanations collected immediately after the decision. We use an LLM to propose a taxonomy of reasons and classify each of the 255 responses under a single-label scheme (Table 4). The explanations reveal that most participants did not anticipate performance declines. Among those expressing no preference, over half cited indifference and 21.6 percent cited uncertainty about question difficulty—both consistent with believing performance would be stable. Participants who chose the first third overwhelmingly cited wanting to avoid fatigue (64.8 percent), though 18.2 percent offered reasons unrelated to performance expectations (e.g., stress reduction, earlier payout). Those choosing the middle or last third typically believed performance improves after warming up.

**Willingness to Pay for Timing.** Direct choices can magnify small valuation differences into apparently strong preferences. To measure the intensity of preferences, we elicited WTP for receiving bonus boosts in a randomly assigned third of the exam using a MPL.

Mean WTP is similar across exam thirds: \$2.03 for the first, \$2.20 for the middle, and \$2.06 for the last (Figure 4, Panel B). These valuations are not statistically distinguishable from each other, though all are significantly positive and large relative to the expected value of the booster. Thus, while participants value boosted payments, they do not prefer one timing over another.

Panel C disaggregates WTP by whether participants were assigned to their preferred timing, a different timing, or had expressed no preference. Participants assigned to their preferred timing show only marginally higher WTP than others, and the differences are not statistically significant. Even participants who chose early timing do not value it more

than those who expressed indifference. This establishes our fourth result:

**Result 4.** *Participants do not prefer early bonus placement despite its higher expected value, and most express indifference over timing.*

We next examine whether enhanced incentives can offset performance declines, and quantify the cost of suboptimal timing.

### 4.2 Bonus Boosts and Exam Performance

We exploit the randomization of bonus boost timing to identify the causal effect of enhanced incentives on question-level outcomes. We restrict this analysis to participants whose timing was randomly assigned (excluding the 20 percent whose preferences were implemented) and estimate regressions of the form:

$$Y_{it} = \tau \cdot \text{Boosted}_{it} + \gamma_i + \delta_t + \lambda_{p(i)} + \epsilon_{it} \tag{5}$$

where $Y_{it}$ is an outcome for participant $i$ in exam third $t$, $\text{Boosted}_{it}$ indicates whether participant $i$ receives bonus boosts in third $t$, $\gamma_i$ are participant fixed effects, $\delta_t$ are exam-third fixed effects, and $\lambda_{p(i)}$ are comparison-group fixed effects that pair each treatment group with the pure control group.[5] We cluster standard errors at the participant level. Table 5 presents results; Figure 5 displays the patterns graphically (Appendix Figure A10 shows results separately by treatment group).

Participants expend significantly more effort on boosted questions: they spend 7.4 additional seconds per question—a 9.3 percent increase—and report higher perceived effort (both $p < 0.01$). However, this increased effort translates into only a modest and not statistically significant performance increase of 2.3 pp ($p = 0.195$). These effects do not vary significantly across exam thirds (Table 5, Panels B–C), and we find no evidence of spillovers to non-boosted questions (Appendix Table A2), indicating that bonus boosts do not meaningfully alter overall performance trajectories. Consistent with these limited effects, participants' forecasts show they expect only small performance improvements from enhanced incentives. This establishes our fifth result:

[5]The inclusion of comparison-group fixed effects is important given 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 could 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.

**Result 5.** *Enhanced incentives increase effort but yield only modest performance gains, and cannot substantially offset underlying performance declines.*

### 4.3 The Cost of Suboptimal Timing

We quantify the financial consequences of bonus timing by comparing expected earnings between participants randomly assigned to early versus late boosts.

Participants receiving early boosts have expected bonus earnings of \$2.26 from the first 10 questions, compared to \$1.99 for late-boost participants on their final 10 questions—a 14 percent difference ($p = 0.010$, one-sided). Total expected bonus payments are \$3.09 for the early group versus \$2.84 for the late group, a 9 percent difference ($p = 0.035$). The smaller total difference reflects identical \$2 payments for middle questions. This earnings gap is a direct consequence of Results 3 and 4: because effort cannot offset performance declines, early placement is optimal—yet most participants fail to recognize this advantage.

**Result 6.** *Suboptimal bonus timing is costly: late placement reduces expected earnings relative to early placement.*

### 4.4 Advice to Future Test-Takers

To shed light on which strategies participants view as important, we analyze the incentivized open-ended advice they write immediately after the exam. We code each response in two steps. First, we assign a single-best theme and a set of any-mention (multi-label) tags capturing all themes present. Second, we code pre-specified indicators tied to our mechanism: whether advice references performance trajectories (declines/improves/constant), names a mechanism (fatigue, time pressure, warm-up), reasons about intertemporal tradeoffs, and recommends an effort-allocation rule (uniform, front-load, back-load, adaptive, or stakes-contingent).

Three patterns emerge. First, advice is dominated by broad test-taking heuristics rather than fatigue-based optimization. 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%).

Second, explicit discussion of within-exam performance dynamics is rare. Only 9.6% state that performance declines over time, and just 3.5% attribute this to cognitive fatigue. Qualitative intertemporal reasoning is more common—34.2% invoke tradeoffs such as avoiding burnout—but rarely translates into concrete plans: only 11.3% endorse a flat

effort path and 8.4% provide a quantitative pacing rule (Table 7). Relatedly, almost no respondents mention the random-selection incentive scheme, suggesting that participants do not naturally frame the task as an intertemporal optimization problem.

Third, advice themes are 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 (Table 6). 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.[6] Outside these differences, explicit references to performance decline, cognitive fatigue, or implementable pacing rules remain uncommon in all arms (Table 7).

Taken together, the advice data reinforce the view that performance declines reflect misoptimization rather than deliberate strategy: even after experiencing the exam, participants do not identify declining performance as a key challenge or recommend effort-allocation strategies to counteract it.

## 5 The Role of Time Allocation and Cognitive Fatigue

We investigate two mechanisms underlying performance declines: changes in effort allocation and the accumulation of mental fatigue.

### 5.1 Time Allocation Patterns

Performance declines may reflect declining effort over the exam. We use time spent per question as a proxy for effort and validate this measure in two ways (Figure 6). 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). Together, these patterns support time as a valid proxy for cognitive effort.

We next examine how time allocation varies over the exam. Figure 6, Panel C shows that mean time per question declines from 93 seconds on the first question to 71 seconds on

[6] Stakes-based allocation advice is absent in the control group (0.0%) but appears in 13.6–17.0% of boost-arm responses; similarly, bonus-prioritization advice is absent in the control arm (0.0%) but appears in 17–26% of boost-arm responses.

the final question—a 24 percent reduction. This pattern holds for most individuals: 70.7 percent exhibit declining time trajectories, with a mean reduction of 23.2 seconds from first to last question (Panel D).

Several pieces of evidence suggest that this decline does not reflect participants running out of time. The vast majority complete the exam with substantial time remaining (Appendix Figure A5, Panel A). Time investment declines smoothly rather than dropping sharply as would occur if participants suddenly faced binding constraints (Appendix Figure A5, Panel B). Fewer than one percent of questions go unanswered, and these occur predominantly at the very end (Appendix Figure A6). The decline holds for both correct and incorrect responses (Appendix Figure A7). Finally, we observe similar patterns when restricting to participants who finished with at least three minutes remaining (Appendix Figure A8). These patterns indicate that declining effort is a choice, not a constraint.

### 5.2 Decomposing Performance Decline

To quantify how much of the performance decline reflects changes in time allocation, we estimate an augmented version of equation (1) that flexibly controls for time spent, either as a seventh-order polynomial or a set of ten-second bin fixed effects:

$$\text{Correct}_{ij} = \tilde{\alpha} + \tilde{\beta} \cdot \text{NormPos}_{ij} + f(\text{Time}_{ij}) + \tilde{\gamma}_j + \tilde{\varepsilon}_{ij}. \tag{6}$$

The coefficient on normalized position falls from $\hat{\beta} = -0.095$ in the baseline to $\hat{\beta} = -0.062$ with polynomial controls and $\hat{\beta} = -0.054$ with time bin fixed effects—a reduction of 34.2–42.6 percent (Table 2, columns 3–4). Thus, declining time investment accounts for roughly one-third to a half of the performance deterioration in a regression sense. The remaining decline may reflect cognitive fatigue reducing the productivity of effort: participants not only spend less time on later questions but also derive less benefit from each unit of time invested. We do not find any evidence that our results are driven by disengagement or random guessing (see Appendix B.5).

### 5.3 Cognitive Fatigue and Forecast Errors

Why do participants allocate less time as the exam progresses? One possibility is that they fail to anticipate mental fatigue, leading them to expend too much effort early and too little later. We test this by examining whether fatigue forecast errors predict performance declines.

Figure 7, Panel A shows that mental fatigue increases substantially, rising from 2.5 before the exam to 6.2 after—a 3.7-point increase on a 10-point scale. On average, participants correctly anticipate this increase: their mean forecast of 6.6 is close to actual post-exam fatigue. However, individual forecast errors are common (Panel B). Only 27.5 percent of participants accurately predict their fatigue; the remainder split between underestimating (27.2 percent) and overestimating (45.2 percent).

These forecast errors predict exam behavior. Participants who underestimate fatigue exhibit larger reductions in time spent over the exam, while those who overestimate show smaller declines (Panel C). A similar pattern emerges for performance: underestimating fatigue is associated with steeper performance decline (Panel D). This establishes our final result:

**Result 7.** *Declining effort—proxied by time investment—accounts for roughly half of the performance decline, and this effort reduction is larger among participants who underestimate mental fatigue.*

## 6 Conclusion

A large literature documents that individual performance deteriorates over the course of cognitively demanding tasks. We propose and test a novel explanation for this decline: optimization failures. Participants in our experiment fail to anticipate performance declines, display indifference over incentive timing despite clear payoff consequences, and recognize ex post that their effort allocation was suboptimal. We find evidence for two mechanisms: declining effort accounts for roughly half of the performance deterioration, and the remaining decline is consistent with fatigue reducing the marginal product of effort. These mechanisms are linked to belief errors: participants who underestimate how fatigued they will become reduce effort more sharply once fatigue sets in, leading to steeper performance declines. Consistent with limited learning about within-task deterioration, participants' open-ended advice rarely references declining performance or cognitive fatigue and instead emphasizes generic heuristics like pacing and careful reading.

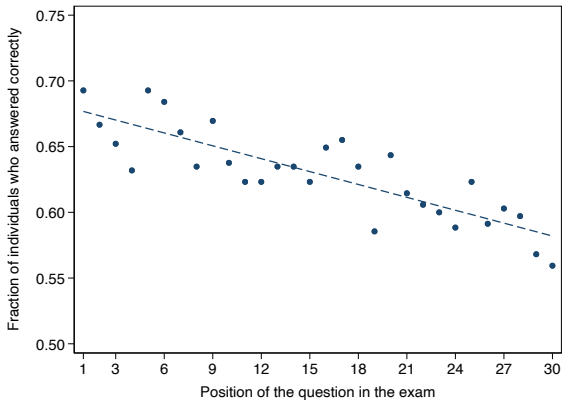
Our findings complement existing explanations that emphasize human capital as the origin of performance declines. We document a novel source that stems from the challenges in allocating effort over time in an optimal way. This mechanism likely extends beyond our experimental setting. For example, hospital nurses exhibit declining hand hygiene compliance despite extensive training and institutional protocols, and these declines appear

to reflect costly mistakes rather than strategic triaging of effort (Steiny Wellsjo, 2025).

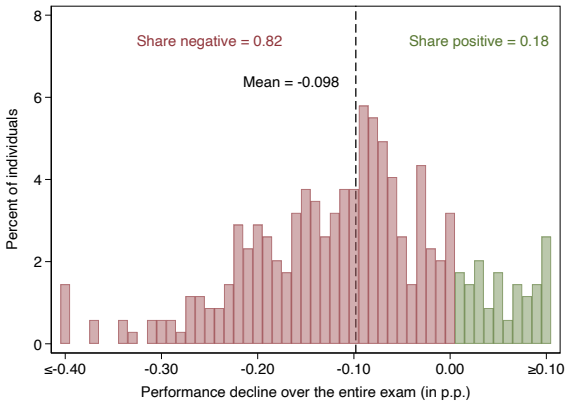
These results suggest opportunities for intervention. If performance declines partly reflect optimization failures, then light-touch interventions may help. For example, organizations could schedule high-stakes tasks earlier or implement strategic breaks. An open question is whether correcting beliefs alone improves outcomes or whether execution aids are also necessary.

Figure 1: Performance Declines Over the Course of the Exam

Panel A. Average Performance by Question Position



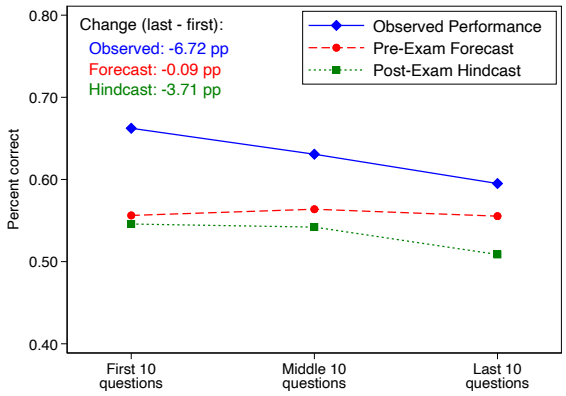

Panel B. Distribution of Individual Performance Declines



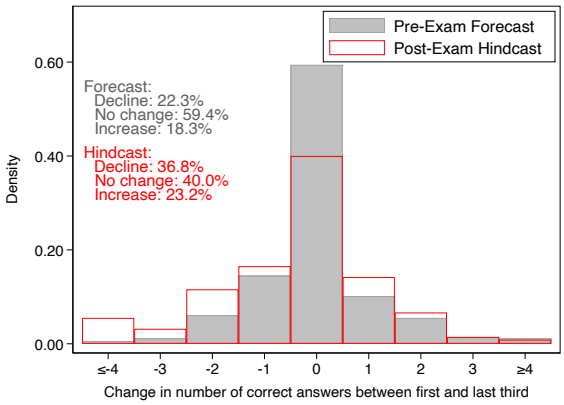

*Notes:* Panel A shows the fraction of participants answering correctly at each position in the exam. The dashed line represents predicted values from a linear regression of performance on question position. Panel B displays the distribution of individual-level performance slopes ($\beta_i^s$) from regressions of correct answers on normalized question position, adjusted using an empirical Bayes shrinkage estimator. For each participant $i$, we estimate $\text{Correct}_{ij} = \alpha_i + \beta_i \cdot \text{Position}_{ij} + \varepsilon_{ij}$, where Position 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.

Figure 2: Beliefs About Performance Trajectories

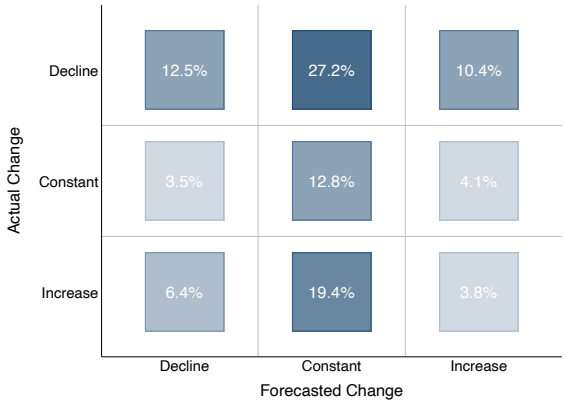
Panel A. Mean Performance Across Exam Thirds




Panel B. Distribution of Beliefs About Performance Change




Panel C. Joint Distribution of Actual and Forecasted Performance Change

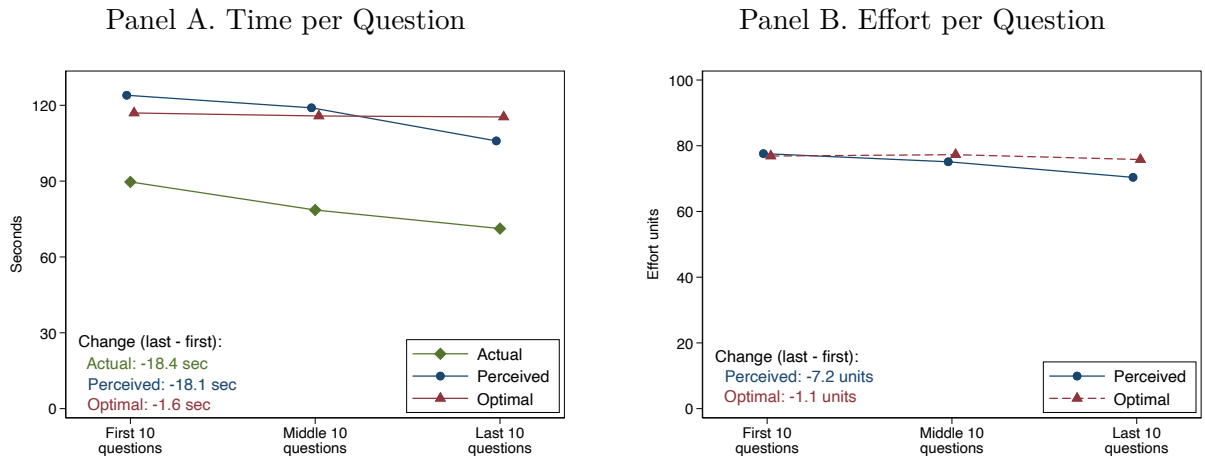



Panel D. Joint Distribution of Actual and Hindcasted Performance Change




*Notes:* Panel A shows mean performance in the first 10 exam questions, middle 10 questions, and last 10 questions. 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 last 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 last 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 last thirds. In Panels C and D, percentages sum to 100 percent across all nine cells.

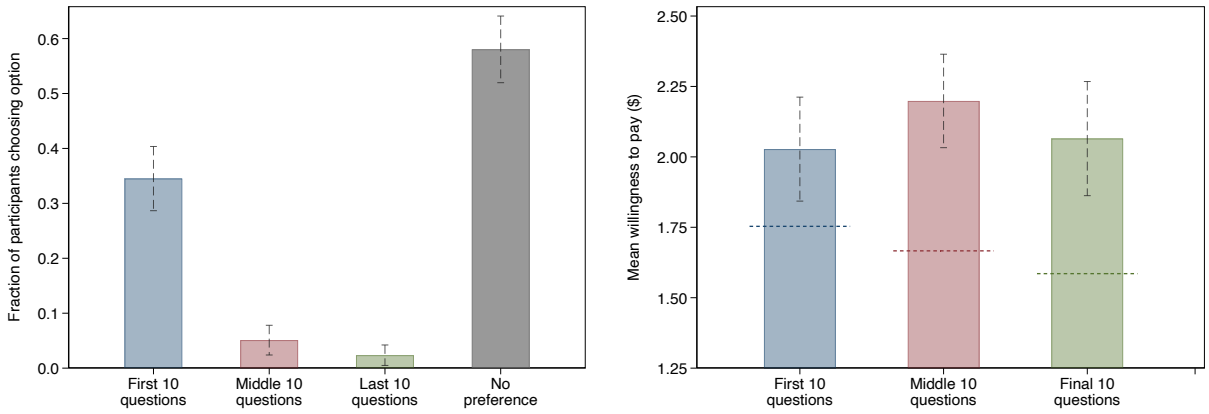
Figure 3: Time and Effort Allocation: Actual, Perceived, and Optimal



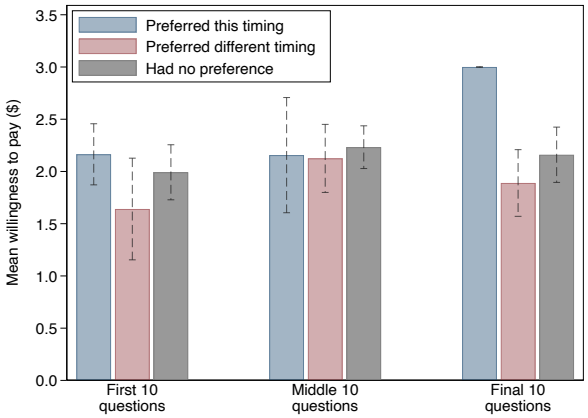

*Notes:* Panel A compares actual time spent per question (green diamonds) with participants' perceived time allocation (blue circles) and their retrospectively reported ideal allocation (red triangles) across exam thirds. Panel B shows perceived effort (blue circles) and ideal effort allocation (red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

Figure 4: Bonus Boost Allocation Preferences

**Panel A.** Distribution of Bonus Boost Choices **Panel B.** Willingness to Pay by Timing Choice

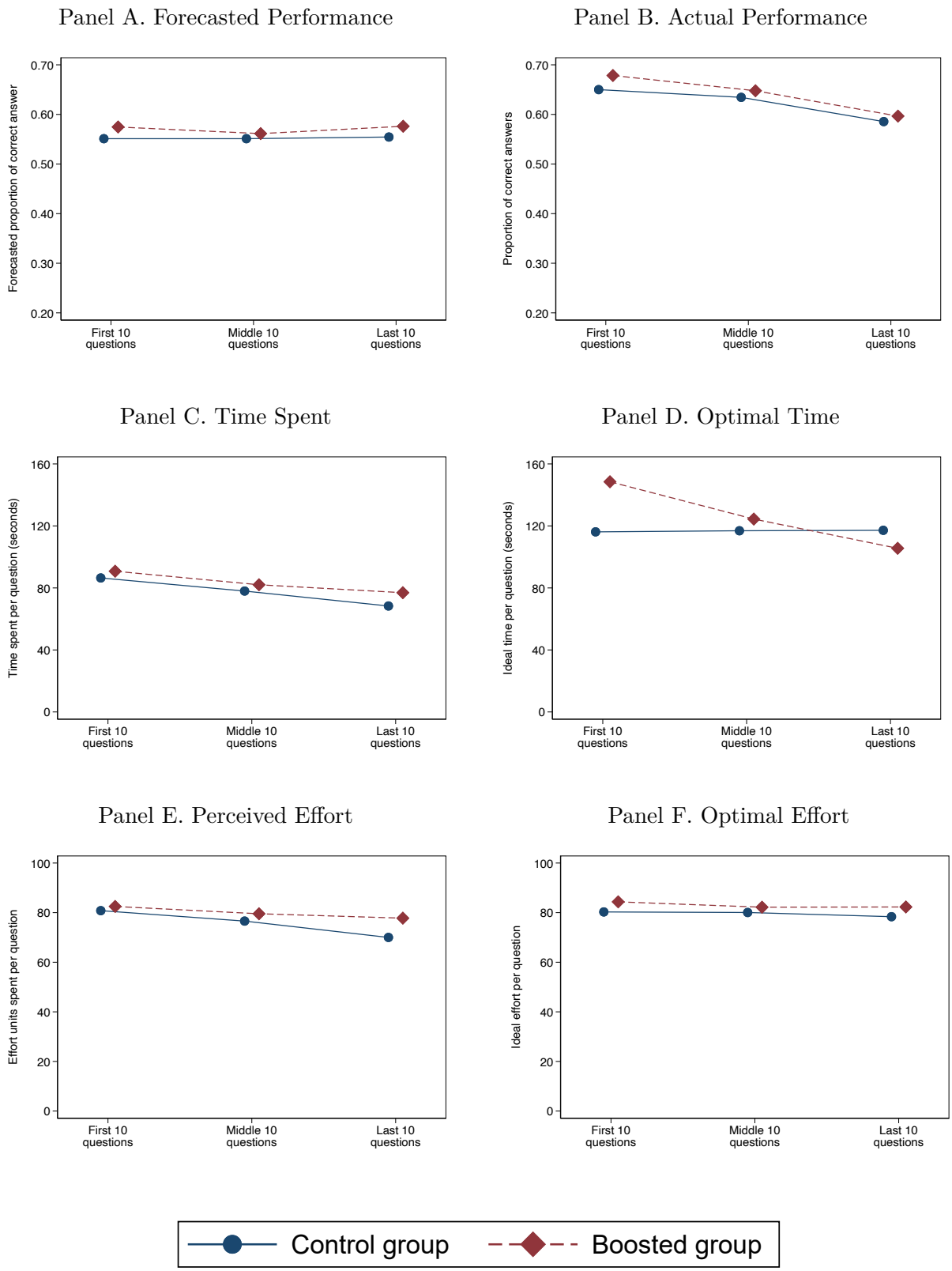



**Panel C.** Willingness to Pay by Preference Match




*Notes:* 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. Error bars represent 95 percent confidence intervals. Panel B displays mean willingness to pay (in dollars) by participants' chosen timing, derived from the multiple price list (MPL) task. Horizontal dashed lines show the predicted expected value of placing boosts in each third, calculated as the control-group baseline accuracy multiplied by the bonus stake (\$8) and the probability of random selection (1/3). Panel C shows mean willingness to pay 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 (green).

Figure 5: Average Outcomes Across Exam Thirds for Control and Bonus-Boosted Groups

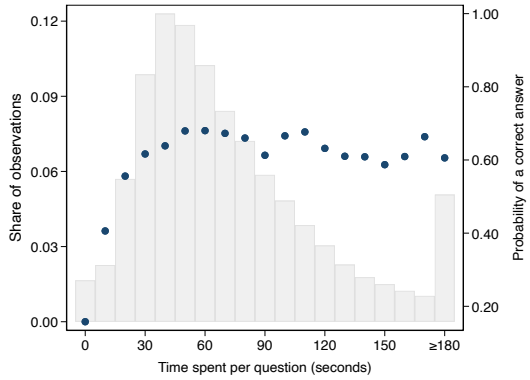




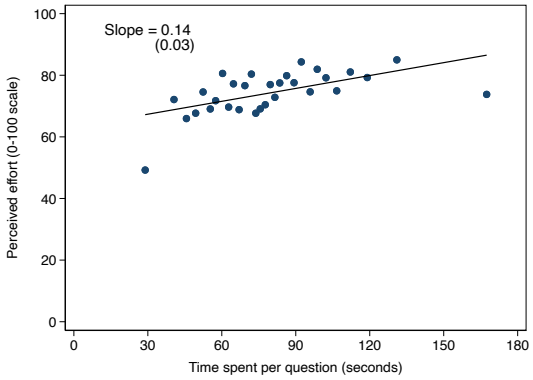

*Notes:* This figure compares outcomes between control participants (navy circles, solid lines) and participants who received bonus boosts in their assigned third (maroon diamonds, dashed lines). The boosted group pools participants across all three treatment arms, showing only their performance during incentivized periods. Panel A shows forecasted performance, Panel B shows actual performance, Panels C and D display actual and retrospectively reported ideal time allocation, and Panels E and F show perceived and retrospectively reported ideal effort levels.

Figure 6: Time Allocation, Effort, and Exam Performance

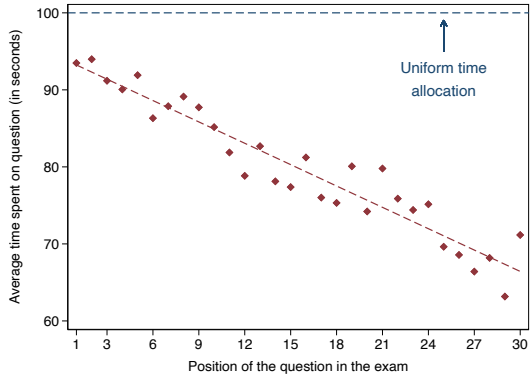
Panel A. Question Performance vs. Time Invested



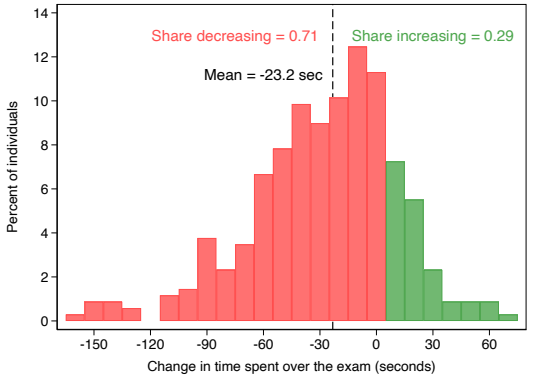

Panel B. Perceived Effort vs. Time Invested




Panel C. Time Spent Over the Exam



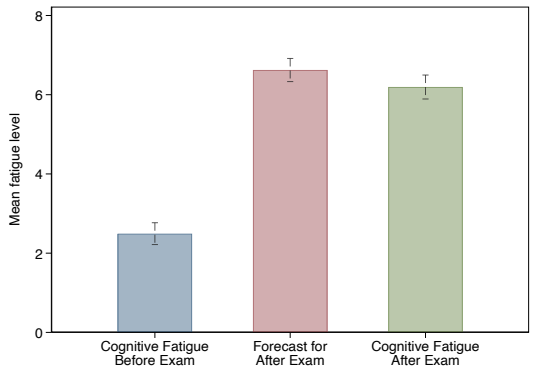

Panel D. Distribution of Time Changes



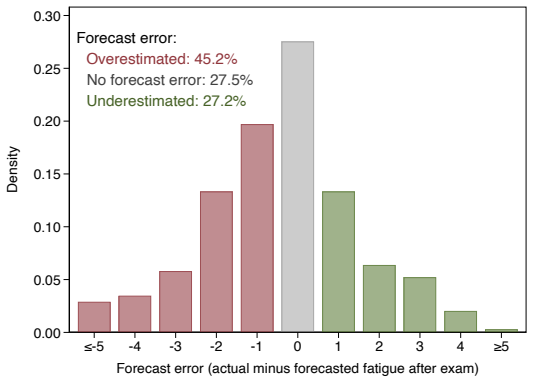

*Notes:* Panel A shows the relationship between time spent on questions and the probability of answering correctly. The scatter points (left $y$-axis) represent mean performance within 10-second time bins. The gray bars (right $y$-axis) show the distribution of observations across time bins. Time values 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 ($\delta_i^s$) 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.

Figure 7: Cognitive Fatigue: Forecasts vs. Actual Experience

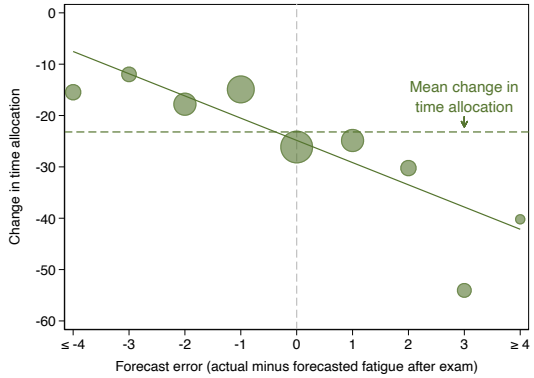
Panel A. Mean Fatigue Levels



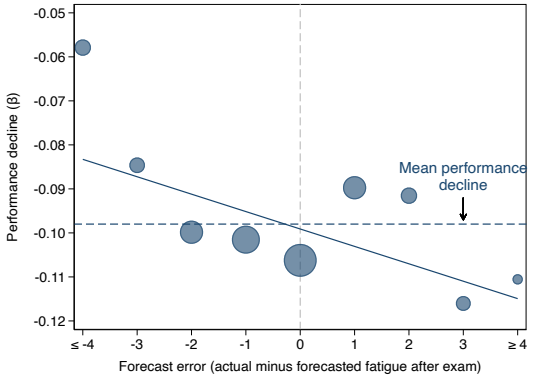

Panel B. Distribution of Forecast Errors




Panel C. Time Allocation and Fatigue Forecast Error




Panel D. Performance Decline and Fatigue Forecast Error




*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 ($\beta$ 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).

Table 1: Summary Statistics

| | | | 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 | 0.491 | 0.478 | 0.418 | 0.557 | 0.506 |
| White/Caucasian | 0.664 | 0.730 | 0.603 | 0.670 | 0.644 |
| Born in USA | 0.945 | 0.944 | 0.962 | 0.932 | 0.943 |
| Native English speaker | 0.962 | 0.978 | 0.962 | 0.966 | 0.943 |
| **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 | 0.351 | 0.333 | 0.241 | 0.477 | 0.341 |
| **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 | 0.130 | 0.159 | 0.043 | 0.182 | 0.130 |
| **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 |
| **Panel E. Exam Variables** | | | | | |
| Fraction correct | 0.629 | 0.623 | 0.640 | 0.633 | 0.622 |
| Total questions attempted | 29.864 | 29.911 | 29.899 | 29.875 | 29.773 |
| Time on questions (minutes) | 40.332 | 39.173 | 40.499 | 41.102 | 40.596 |
| Questions with tab switches | 0.719 | 0.878 | 0.620 | 0.807 | 0.557 |
| *N* | 345 | 90 | 79 | 88 | 88 |

*Notes:* This table presents summary statistics for the experimental sample across treatment conditions. 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.

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 $\beta$ from equation (1). 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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

Table 4: Boosted Payment Choice Reasons

| | | Timing Choice | | | |
|---|---|---|---|---|---|
| | (1) Overall | (2) No Pref. | (3) First | (4) Middle | (5) Last |
| Indifference / leave it to chance | 31.8 | 52.7 | 2.3 | 7.7 | 0.0 |
| Best early (freshness / avoid fatigue) | 22.7 | 0.0 | 64.8 | 0.0 | 16.7 |
| Uncertainty about difficulty | 12.9 | 21.6 | 0.0 | 7.7 | 0.0 |
| Belief about difficulty ordering | 5.9 | 0.7 | 15.9 | 0.0 | 0.0 |
| Probability / expected-value reasoning | 3.5 | 4.1 | 2.3 | 7.7 | 0.0 |
| Best after warm-up (middle/end) | 5.9 | 0.0 | 1.1 | 69.2 | 83.3 |
| Reduce stress / decision aversion | 7.5 | 8.8 | 6.8 | 0.0 | 0.0 |
| Motivation management | 1.2 | 0.7 | 1.1 | 7.7 | 0.0 |
| Want the bonus earlier (payout timing) | 1.6 | 0.0 | 4.5 | 0.0 | 0.0 |
| Delegate choice / convenience | 7.1 | 11.5 | 1.1 | 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-5.1-Pro) to propose a taxonomy of response categories and then assigned each response to a single best-fitting category using GPT-5.1-mini. Column 1 reports shares in the full sample; columns 2–5 report shares by timing choice.

Table 5: 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 5. 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 6: 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 7: 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.

## References


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.

Bettinger, E. P., Evans, B. J., and Pope, D. G. (2013). Improving college performance and retention the easy way: Unpacking the act exam. *American Economic Journal: Economic Policy*, 5(2):26–52.

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. (2024). Cognitive endurance as human capital. *The Quarterly Journal of Economics*.

Bulman, G. (2015). The effect of access to college assessments on enrollment and attainment. *American Economic Journal: Applied Economics*, 7(4):1–36.

Caplin, A., Csaba, D., Leahy, J., and Nov, O. (2020). Rational Inattention, Competitive Supply, and Psychometrics. *The Quarterly Journal of Economics*, 135(3):1681–1724.

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.

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.

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

Goodman, J., Gurantz, O., and Smith, J. (2020). Take two! sat retaking and college enrollment gaps. *American Economic Journal: Economic Policy*, 12(2):115–58.

Goodman, S. (2016). Learning from the test: Raising selective college enrollment by providing information. *Review of Economics and Statistics*, 98(4):671–684.

Hoxby, C., Turner, S., et al. (2013). Expanding college opportunities for high-achieving, low income students. *Stanford Institute for Economic Policy Research Discussion Paper*, 12(014):7.

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

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.

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

Milgrom, P. and Segal, I. (2002). Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601.

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*.

Riehl, E. (2024). Do less informative college admission exams reduce earnings inequality? evidence from colombia. *Journal of Labor Economics*, 42(4):1009–1047.

Rothstein, J. M. (2004). College performance predictions and the sat. *Journal of Econometrics*, 121(1-2):297–317.

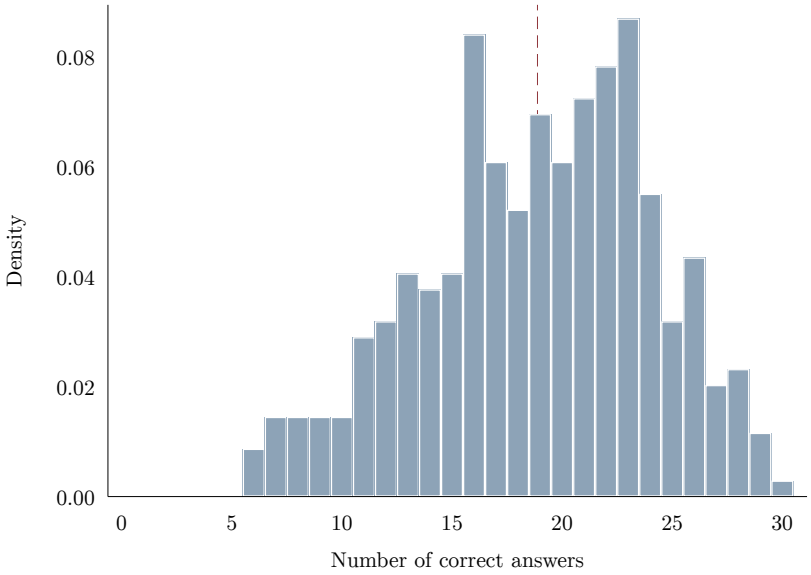
Steiny Wellsjo, A. (2025). Simple actions, complex habits: Lessons from hospital hand hygiene. Working Paper.

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

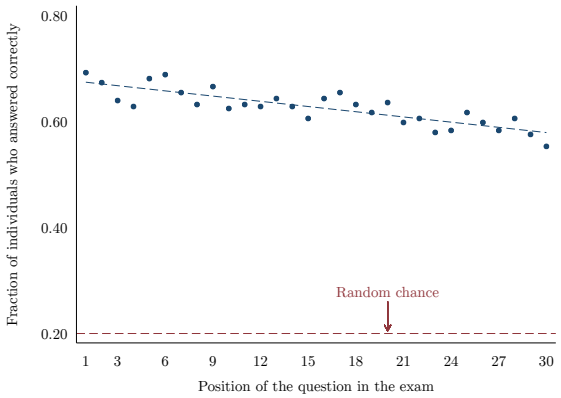
Figure A1: Distribution of Number of Correct Exam Responses



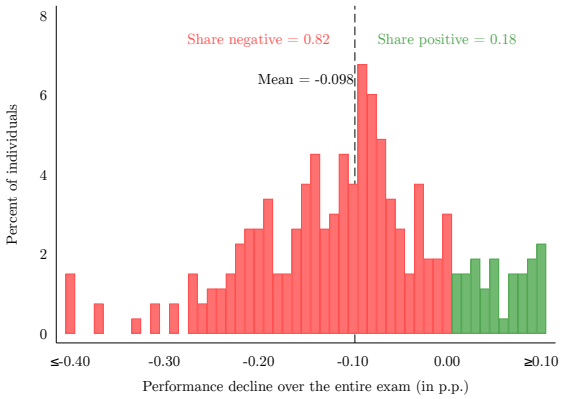

*Notes:* This figure shows the distribution of total correct answers across all participants. The vertical dashed line indicates the mean (18.9 correct answers). The histogram uses bins of width one, ranging from 0 to 30 correct responses.

Figure A2: Performance Declines Excluding First-Third Boost Recipients

Panel A. Average Performance by Question Position

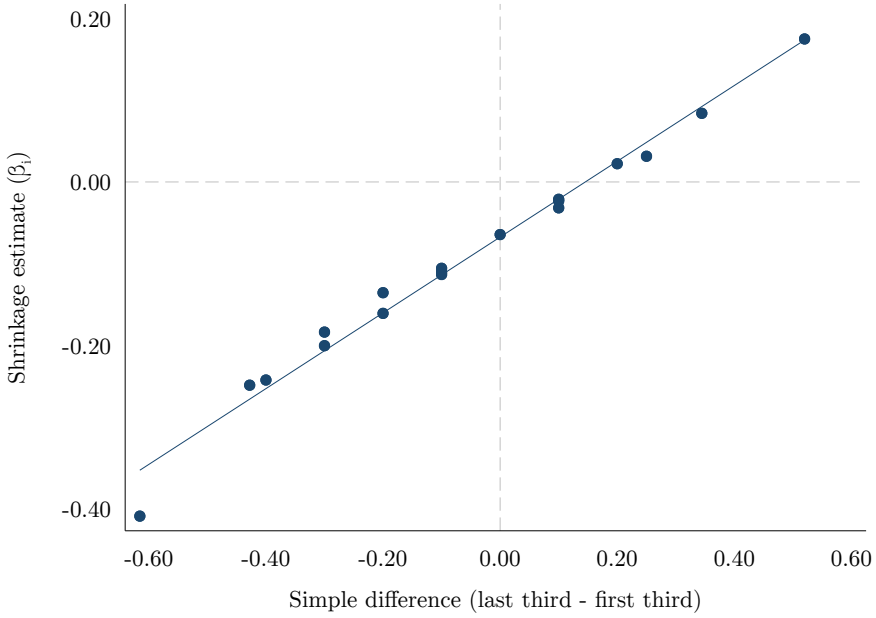



Panel B. Distribution of Individual Performance Declines




*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 ($\beta_i^s$) 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.

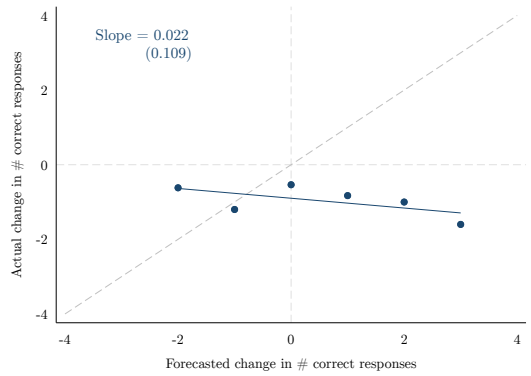
Figure A3: Comparison of Performance Decline Measures



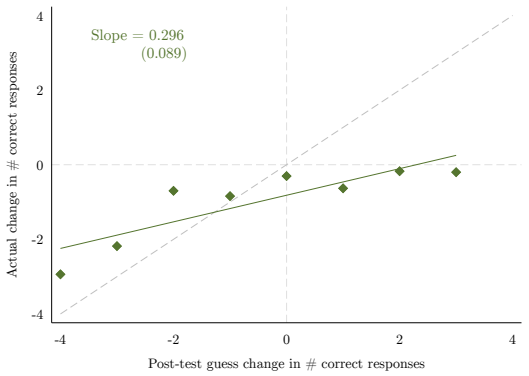

*Notes:* This figure presents a binned scatterplot comparing two measures of individual-level performance decline: (1) the simple difference in fraction correct between the last third (questions 21-30) and first third (questions 1-10) of the exam, shown on the $x$-axis, and (2) the empirical Bayes shrinkage 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 A4: Performance Change Forecasts vs. Actual Performance Changes

Panel A. Pre-Exam Forecast



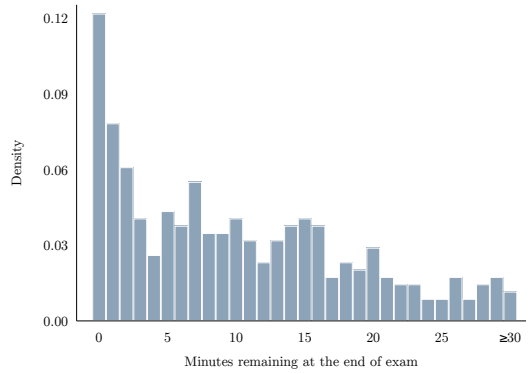

Panel B. Post-Exam Hindcast



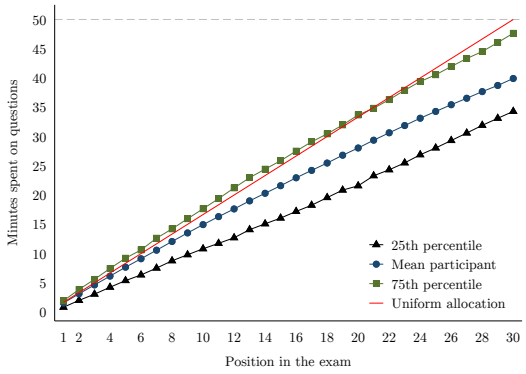

*Notes:* This figure shows binned scatter plots of participants' beliefs about performance changes against actual performance changes between the first and last thirds of the exam. Panel A compares pre-exam forecasted changes with actual changes in correct responses. Panel B compares post-exam hindcasted 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 A5: Time Management in the Math Exam

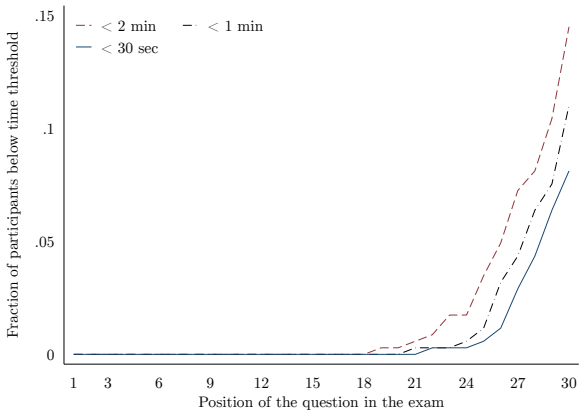
Panel A. Time Remaining at Exam Completion




Panel B. Cumulative Time Allocation

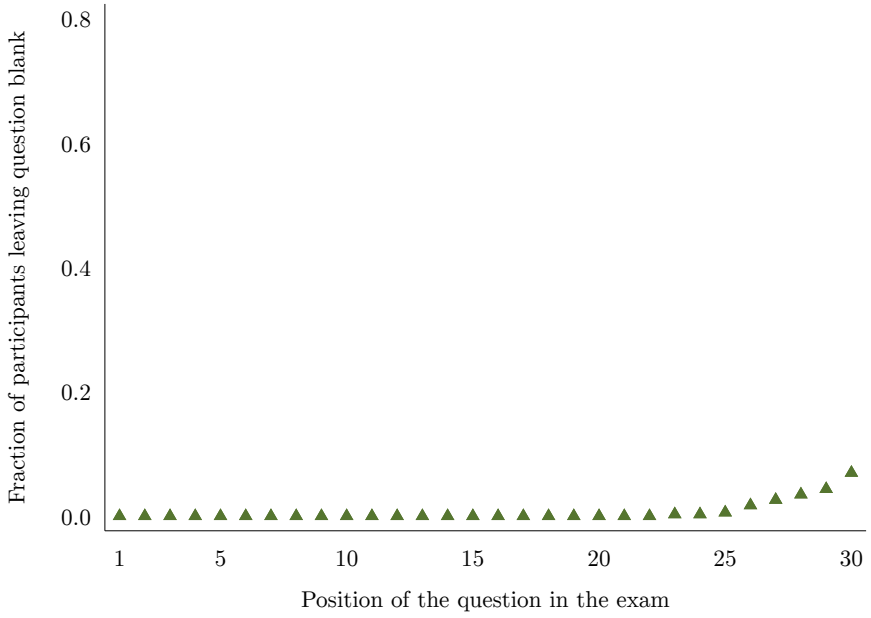



Panel C. Fraction of Participants Below Time Threshold




*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 A6: Unanswered Questions by Position in Exam



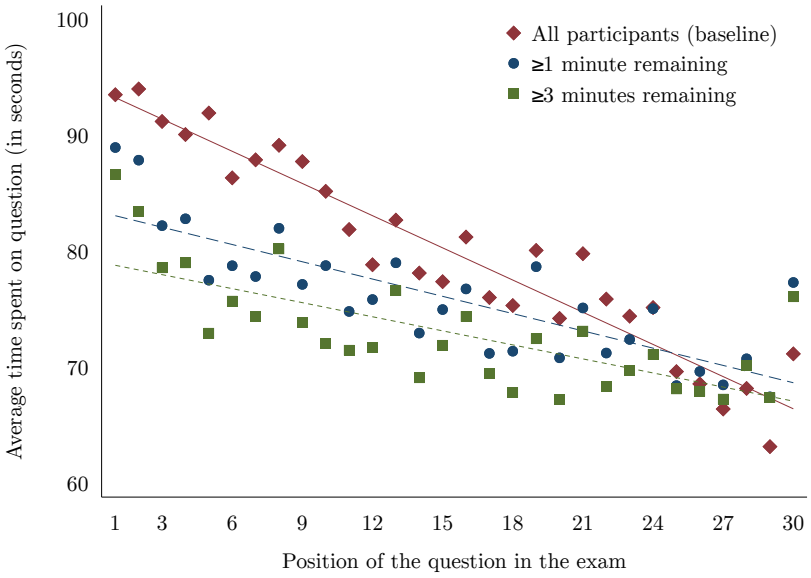

*Notes:* This figure shows the percentage of participants who left each question unanswered as a function of its position in the 30-question exam. Each point represents the fraction of missing responses for that question number.

Figure A7: Time Spent per Question: Correct vs. Incorrect Answers



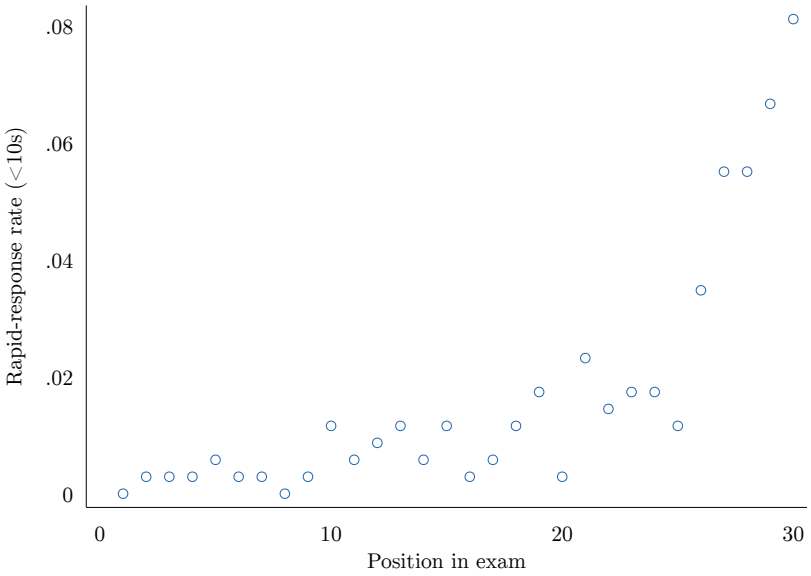

*Notes:* This figure shows the average time spent per question across the exam sequence, separately for correct responses (green triangles) and incorrect responses (red circles). Both series exhibit declines over the course of the exam. Dashed lines represent linear fits for each series.

Figure A8: Time Decline Among Unconstrained Participants



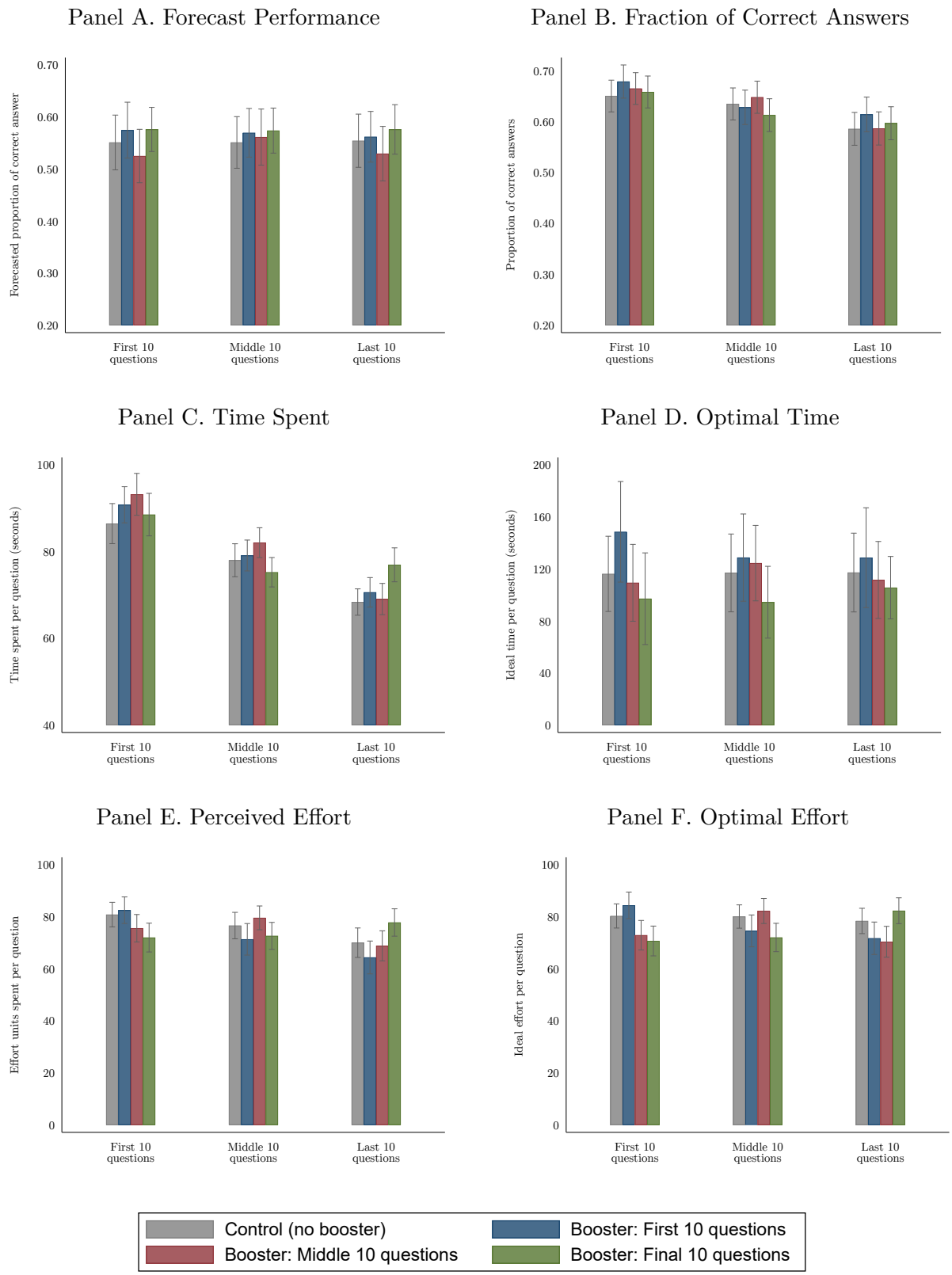

*Notes:* This figure shows average time spent per question across the exam sequence for different subsamples. The red diamonds represent all participants (baseline). The blue circles exclude participants who finished with less than one minute remaining. The green squares exclude participants who finished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

Figure A9: 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.

Figure A10: 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 A1: Estimates of Performance Declines over the Exam: Robustness to Exclusion Criteria

| | 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 $\beta$ from equation (1). 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 A2: Spillover Effects: Performance in Non-Incentivized Blocks

| | Outcome: Mean Performance in Non-Incentivized Blocks | | | | | |
|---|---|---|---|---|---|---|
| | 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 blocks. The dependent variable in each column is the mean outcome across the two blocks where participant $i$ did not receive enhanced incentives. For example, if a participant received boosts in the first block, the outcome is their mean performance in blocks 2 and 3. The coefficient on *Boosted* tests whether receiving enhanced incentives in one block affects outcomes in other blocks. Pure control participants serve as the comparison group, where the outcome for controls in block $t$ is also calculated as the mean across the two other blocks (e.g., when comparing to participants who received boosts in the first block, the control group's outcome is their mean performance in the middle and final blocks). This ensures that both treated and control groups are compared using the same block combinations. The sample includes one observation per participant. Standard errors are heteroskedasticity-robust and clustered at the participant level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

# 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 Exclusion Criteria

| | 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 $\beta$ from equation (1) 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.

### 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 from equation (2) 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 $\beta_i^s$ 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 $\omega_i$ 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:

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

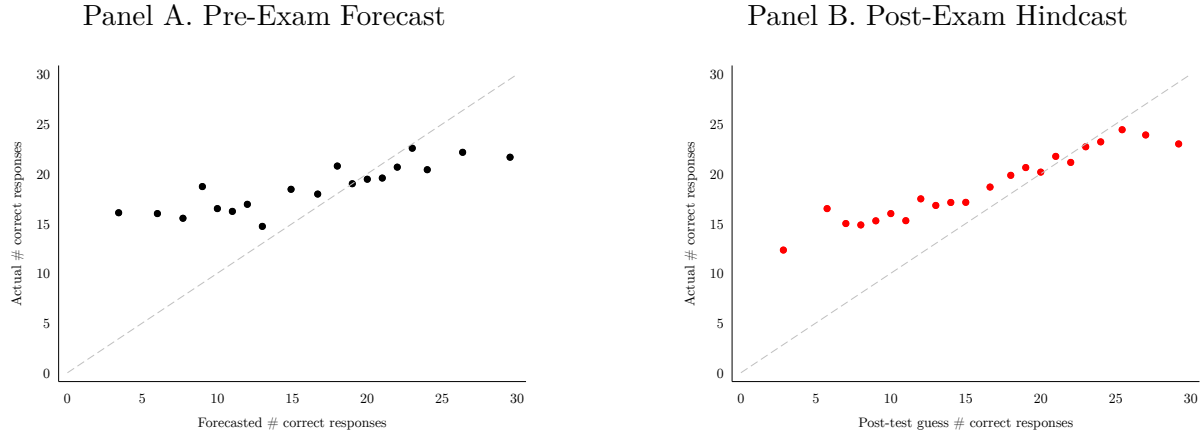
The numerator represents the variance of true performance trajectories (total variance minus average measurement error), while the denominator adds participant $i$'s specific measurement error variance. This formulation ensures that precisely estimated trajectories receive minimal shrinkage toward the 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 2.13 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 6, 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 A9 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 A1).

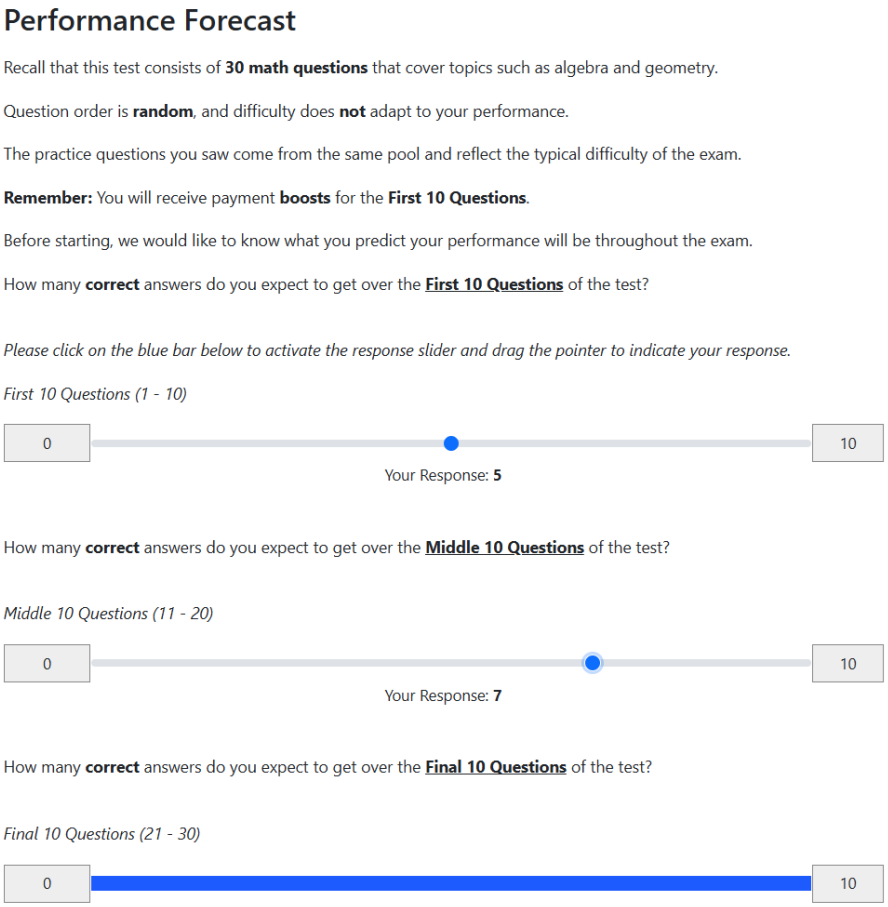
Most participants do not engage in systematic guessing. Over 75 percent of participants spend more than 10 seconds on all 30 questions. When we exclude participants in the 95th percentile of rapid response rates—those giving rapid responses on at least four questions—the position coefficient remains largely unchanged after controlling flexibly for response time (Appendix Table A1).

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 A1).

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 A1 demonstrate that these individuals do not drive our main results.

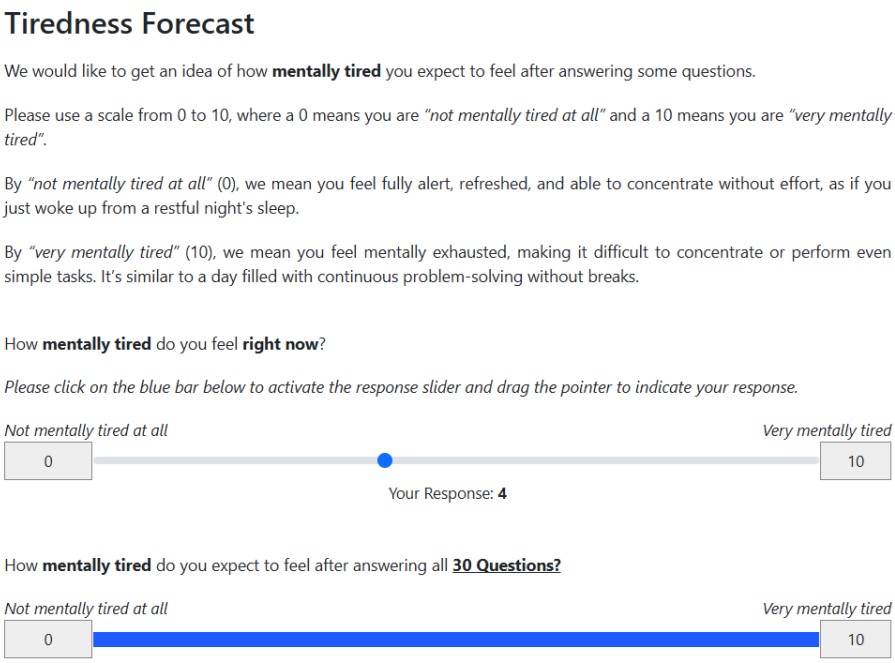
## B.6 Experimental Instruments

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.

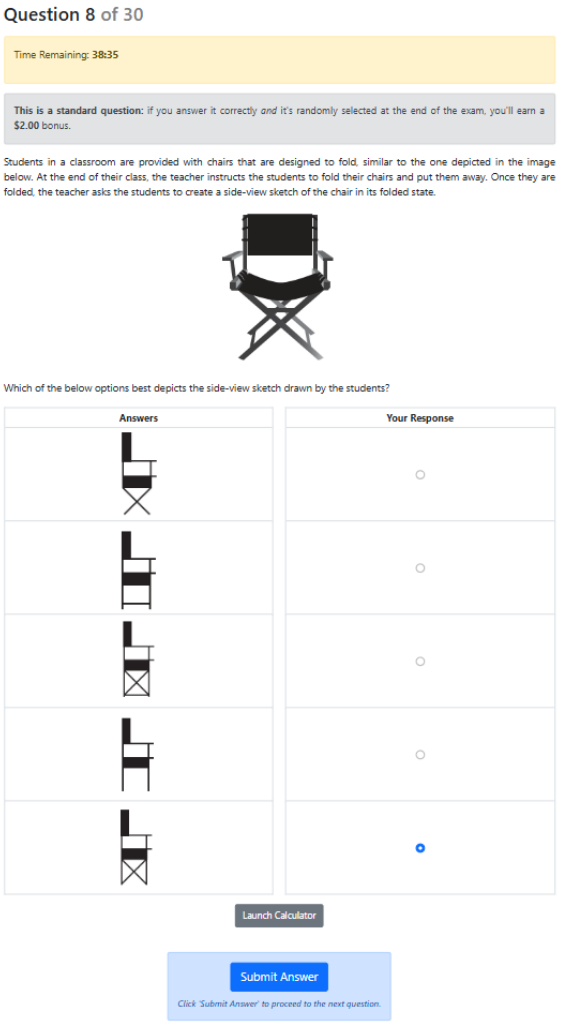
Figure B3: Mental Fatigue Elicitation Interface



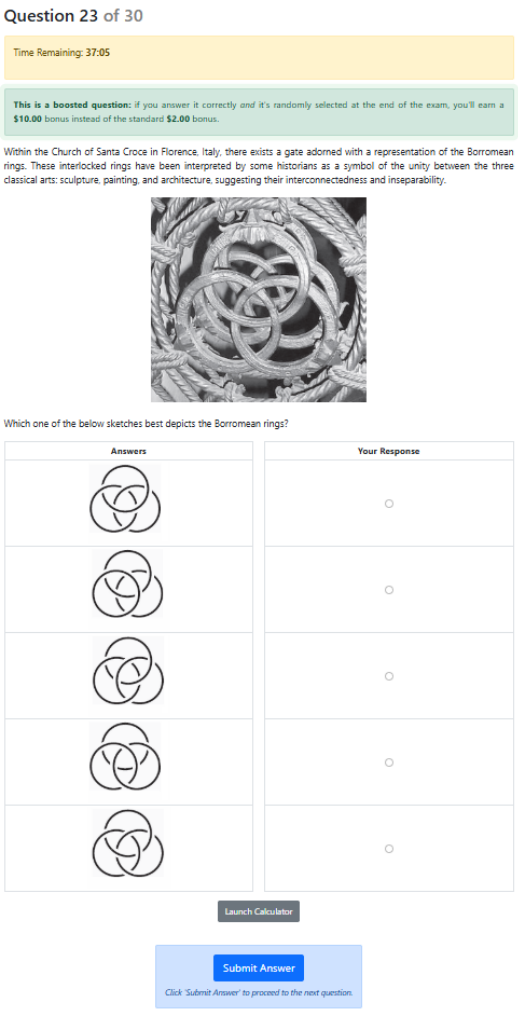

*Notes:* This figure shows the mental fatigue elicitation screen. Participants report their current mental tiredness and predict their tiredness after completing all 30 questions using a 0–10 scale with interactive sliders. The scale anchors define 0 as "not mentally tired at all" (fully alert, refreshed, 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.

# C Theoretical Appendix

We introduce a model of effort allocation with two periods: "early" ($t = 1$) and "late" ($t = 2$).[7] We apply the rational inattention approach outlined in Caplin et al. (2020) and Bronchetti et al. (2023) to derive testable predictions under the null of optimal effort allocation with varying task incentives. We show that willingness-to-pay (WTP) for different payment boost timing and baseline success probabilities under varying incentives are sufficient to test whether agents are allocating effort optimally.

## C.1 Environment

Consider a student who faces $q = 1, \ldots, N$ exam questions. The student chooses effort $e_q \geq 0$ for each question; the success probability on $q$ is $p_q(e_q) \in [0, 1]$ with $p_q'(e_q) > 0$, $p_q''(e_q) < 0$, and $\lim_{e_q \to \infty} p_q(e_q) = 1$. To allow for performance declines across the exam, we permit question-specific productivity. For a fixed effort level, performance declines imply that, $p_q(\bar{e}) > p_{q+1}(\bar{e})$.

At the end of the exam, one question is randomly selected and the student is paid a bonus if they answered the selected question correctly. Let $b(q) \in \{1, 2\}$ denote which period a question is in and define the set of questions in each period as $Q_t = \{q : b(q) = t\}$ for $t \in \{1, 2\}$ with sizes $N_t = |Q_t|$ and shares $\pi_t \equiv N_t/N$. Let $w_{b(q)} \geq 0$ denote the bonus payment for answering $q$. Denote the period-$t$ average success by

$$\bar{p}_t(\mathbf{e}) \equiv \frac{1}{N_t} \sum_{q \in Q_t} p_q(e_q), \qquad t = 1, 2.$$

Thus, the expected bonus can be written as

$$\frac{1}{N} \sum_{q=1}^{N} w_{b(q)}\, p_q(e_q) = \sum_{t=1}^{2} \pi_t\, w_t\, \bar{p}_t(\mathbf{e}).$$

where $\mathbf{e} = (e_1, \ldots, e_N)$ is a vector of effort levels.

Given this set-up, the student solves the following effort allocation problem,

$$\max_{\{e_q\}_{q=1}^{N} \geq 0} \left\{ \frac{1}{N} \left( \sum_{q \in Q_1} w_1\, p_q(e_q) + \sum_{q \in Q_2} w_2\, p_q(e_q) \right) - C(e_1, \ldots, e_N) \right\} \tag{C1}$$

[7] It is straightforward to extend this framework to an arbitrary $N$ periods. However, the simple two-period case is sufficient to illustrate the theoretical predictions that our experiment will test.

where $C(e_1, \ldots, e_N)$ is a convex effort cost function.

**Observation 1.** *For an agent who chooses their optimal effort allocation* $\{e_t^*\}_{t=1}^T$ *by solving equation* (C1)*, the optimal effort path can be increasing, decreasing, or non-monotonic.*

Observation 1 clarifies that we cannot infer anything about the optimality of effort allocation by simply examining its path over the course of the exam. For example, optimal effort could be decreasing if productivity is declining and costs are increasing. Conversely, decreasing costs and increasing productivity (e.g., "getting in the zone") could lead to an increasing optimal effort path. Finally, intermediate cases can generate cases in which optimal effort is non-monotonic.[8] Observation 1 thus highlights that we need additional data to assess the optimality of effort allocation and motivates key aspects of our experiment design: exogenous variation in payments $p_t$ and incentivized measures of willingness to pay for enhanced payments.

### C.2 Restating the Effort Allocation Problem

Following Bronchetti et al. (2023) (Lemma 1) and Caplin et al. (2020), we can restate this high-dimensional choice of effort in terms of an indirect utility function comprised of the aggregate success probabilities and the minimum costs associated with achieving that output. Define the minimal cost of achieving success in periods 1 and 2 as:

$$K(\bar{p}_1, \bar{p}_2) \equiv \inf_{\{e_q\}_{q=1}^N} \Big\{ C(\{e_q\}) \ : \ \bar{p}_t(\mathbf{e}) \geq \bar{p}_t \text{ for } t = 1, 2 \Big\}.$$

where $C$ is a convex cost of effort function. By construction, $K$ is the minimal subjective cost of producing accuracy $(\bar{p}_1, \bar{p}_2)$.

We can therefore define the student's indirect utility function as

$$V(w_1, w_2) \equiv \max_{\bar{p}_1, \bar{p}_2 \in [0,1]} \Big\{ \pi_1 w_1 \, \bar{p}_1 + \pi_2 w_2 \, \bar{p}_2 \ - \ K(\bar{p}_1, \bar{p}_2) \Big\},$$

where $V$ is convex, continuous, and weakly increasing in each $w_t$. When $K$ is differentiable, the first-order conditions are $\partial K / \partial \bar{p}_t = \pi_t w_t$ for $t = 1, 2$, which parallels the competitive-supply characterization in Caplin et al. (2020).

[8] Relatedly, Kaufmann (2022) considers a continuous-time effort allocation problem in which agents fail to anticipate increasing effort costs due to projection bias. He shows that this failure to internalize future costs ex-ante leads agents to plan for effort levels that are too high relative to the optimal. However, agents in this framework are able to adjust their effort supply instantaneously once the true effort costs are realized, leading them to ultimately adhere to the optimal effort path.

### C.3 Willingness to Pay for Boosted Payments

Following Bronchetti et al. (2023), we can apply the envelope theorems in Milgrom and Segal (2002) to analyze the impact of an increase in the bonus payment boost, $w_t$:

$$\frac{\partial V}{\partial w_t} = \pi_t \bar{p}_t^*(w_1, w_2).$$

In other words, the value of a marginal increase in the bonus payment is simply the probability of a period $t$ question being selected for payment times the average likelihood of getting a question from that period correct. Now suppose participants face baseline bonuses $w_1 = w_2 = w$, and have the potential to boost the payment to $w + \Delta$ in one period. Willingness to pay for a boosted payment in period 1 is,

$$\begin{aligned} WTP_1(\Delta|w) =& V(w+\Delta, w) - V(w, w) \\ =& \int_0^\Delta \pi_1 \bar{p}_1^*(w+s, w) ds \end{aligned}$$

with $WTP_2$ defined analogously. Applying Corollary 1 in Bronchetti et al. (2023), for small $\Delta$ we can approximate this difference as:

$$WTP_t(\Delta|w) = \pi_t \bar{p}_t^*(w)\Delta.$$

where $\bar{p}_t^*(w) \equiv \bar{p}_t^*(w, w)$. Intuitively, the student is willing to pay the expected increase in the bonus in period $t$, which is simply the bonus payment $\Delta$ times probability of period $t$ being selected for payment times the average baseline probability of answering a question from that period correctly.

Nuisance costs or framing effects could bias the *levels* of these WTP measures in practice. It is therefore instructive to consider WTP to shift a bonus payment from later to earlier. Formally,

$$\begin{aligned} WTP_{2\to1}(\Delta|w) &= V(w+\Delta, w) - V(w, w+\Delta) \\ &= \int_0^\Delta \left[\pi_1 \bar{p}_1^*(w+s, w) - \pi_2 \bar{p}_2^*(w, w+s)\right] ds \end{aligned}$$

which for $\Delta$ small can be approximated by

$$WTP_{2\to1}(\Delta|w) \approx \Delta \left[\pi_1 \bar{p}_1^*(w) - \pi_2 \bar{p}_2^*(w)\right].$$

In other words, the value of shifting a bonus to the early period is simply the expected increase in the net payoff. This provides a key test of whether individuals are allocating effort optimally in an exam setting: does the relative valuation of receiving boosted payments early versus late align with the associated success probabilities? More concretely, for our experimental setting with $\pi_1 = \pi_2$ and performance declines ($\bar{p}_1^*(w) > \bar{p}_2^*(w)$), this implies that $WTP_{2\to1}(\Delta|w) > 0$.[9]

[9] Bronchetti et al. (2023) also suggest an additional test of optimal effort allocation based on the cross elasticity of performance in a given period due to a change in incentives in the other period. However, since we do not observe any significant effort response spillovers in our context, such a test is uninformative.