Do Time-Constraints Matter? How, Why, and For Whom? *

Esma Ozer†, Kala Krishna‡, and Pelin Akyol §

April 9, 2025

Exams are designed to rank students objectively by their abilities, including elements such as time limits, the number and difficulty of questions, and negative marking policies. Using data from a lab-in-field experiment, we develop and estimate a model of student behavior in multiple-choice exams that incorporates the effects of time constraints and use it to conduct policy analyses for designing more efficient exams in sorting students. We find that additional time benefits men more than women. Our estimated model shows that this is driven by gender differences in the signal production function for the correct answers. Time has a smaller impact on women, while ability and difficulty play a larger role. Risk aversion, in contrast to what is suggested in the literature, does not differ significantly by gender, and confidence rises with more time. Our policy experiments find that exams more effectively rank students when ranking is gender-specific and that time pressure, question difficulty, and student ability have non-monotonic effects on sorting.

First version: Nov 11, 2024. We would like to thank Michael Gechter, Chloe Tergiman, Andres Aradillas-Lopez, Ewout Verriest, James Key, Ömer Faruk Koru, and participants of Penn State Applied Micro Brownbag and Women's Workshop, CIREQ PhD Students' Conference, 2023 EWMES, 2024 AGEW, and 2025 MEA for helpful comments. We also thank Zeynep Yildiz for helping us run our experiment. Financial support from Penn State University's Economics Department is gratefully acknowledged for the experiment. The Institutional Review Board approval was obtained from Penn State University, and the AEA RCT registry ID is 0010719. We also extend our thanks to M. Zahid Ozturk for his valuable insights and discussions that contributed to this work.

†Department of Economics, Penn State University, e-mail: esmaozer@psu.edu

‡Department of Economics, Penn State University, e-mail: kmk4@psu.edu

§e61 Institute and Department of Economics, Bilkent University, e-mail: pelin.akyol@e61.in

1 Introduction

Multiple choice examinations are commonly used worldwide to sort examinees by their proficiency. For example, college entrance exams rank students by score, with allocation mechanisms giving preference to higher-ranking students. English proficiency and similar certification exams verify competence. Millions of people take such exams every year. In 2023, China's college entrance exam, the Gaokao, saw a record registration number of more than 12 million candidates, while Turkey's college entrance exam had more than three million candidates registered.[1] These exams are high stakes, as performance in them determines important outcomes such as college admissions, job opportunities, or entry into a profession. The design of an exam includes various elements, including the number and difficulty of the questions, whether there is negative marking for incorrect answers, and the allotted time.

In this paper, we focus on a relatively poorly studied feature, namely how time constraints, as a specific design feature, impact individual performance and the efficiency of the exam in sorting students by ability. We also examine how time pressure interacts with the other elements of the exam's design. Our work highlights the importance of exam design in shaping student performance and sorting quality. The strength of our approach comes from the structural modeling of decisions and a sorting analysis based on the estimated model. Our work is the first to develop and estimate a model (using data from a field experiment) that examines how time pressure affects the ability of heterogeneous students to identify correct answers to questions of varying difficulty.

In our model, students decide on the answer to each multiple-choice question based on the signals they get. The signal quality "production function" and risk aversion can differ by gender. Inputs into this "production function" are a constant term, time constraints, question difficulty, and student ability. We find that signal processing production functions differ significantly by gender. The constant for this production function is lower for women, consistent with the finding in the literature that women underperform in such tests. Having more time matters less for women, while their ability helps them more, although question difficulty hurts them more. We also find that once we allow for gender differences in the signal processing "production function," women are not more risk-averse/confident, contrary to what the literature suggests. Due to these differences, ranking students independently of gender results in a far worse sorting by ability. The ranking within gender does much better in this dimension. This provides an unexpected plus for policies that set aside seats for women, with the remaining being for men.[2]

We begin by showing some facts from a natural experiment. In 2010, Turkey implemented a policy change that increased the time per question in the college entrance exam without significantly changing its difficulty level. In 2009, the time per question averaged 0.85 minutes; in 2010, this was increased to approximately 1.6 minutes. Using publicly available data on the mean and standard deviation of performance in each high school.[3] We compare student performance

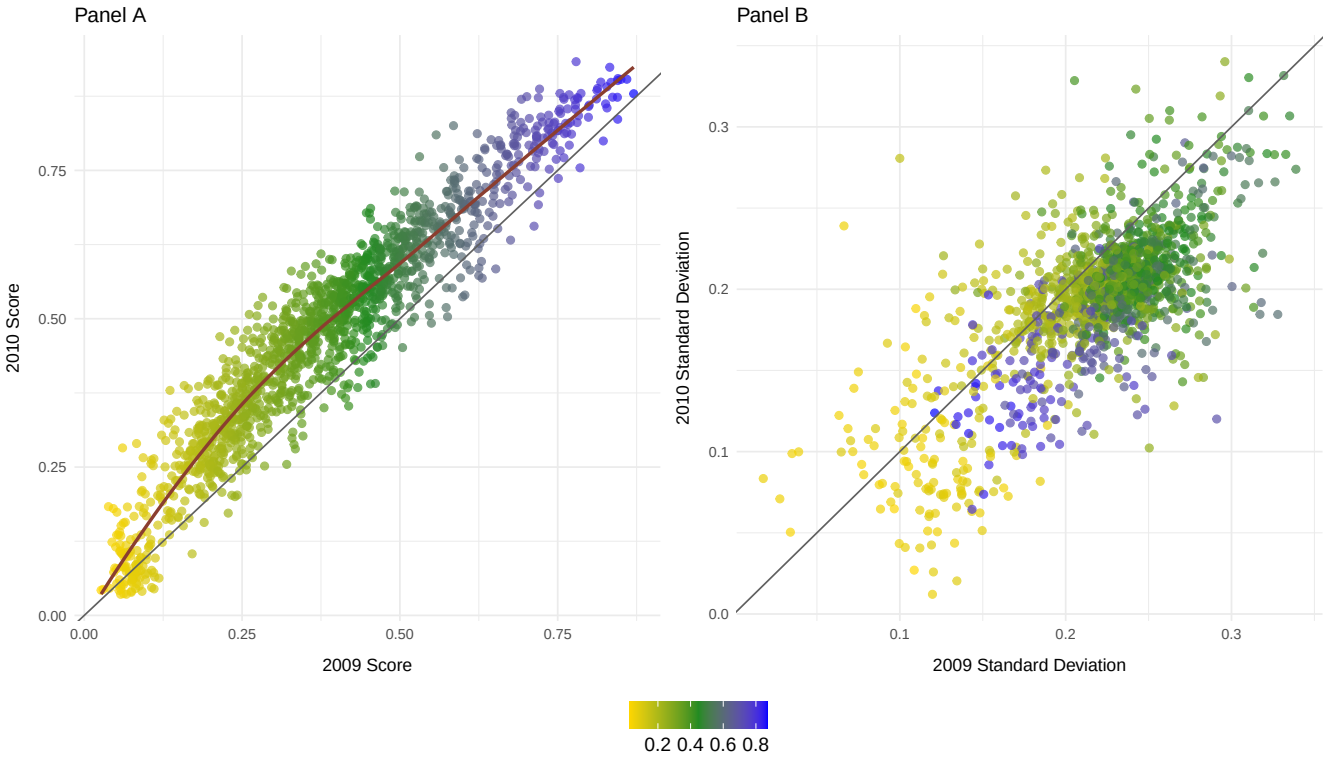
[1] Registration numbers sourced from Global Times for China and OSYM for Turkey. Other examples of such examinations include the SAT, TOEFL, and GRE.

[2] If the number of seats set aside for women were large enough, these would count as affirmative action in their favor.

[3] These are available in a hard copy format published by The Center for Student Selection and Placement (OSYM). We digitized these copies for our data analysis.

before and after this policy change. Figure 1 presents mean scores and standard deviations for schools in 2009, when there was more time pressure, and in 2010 when there was less.[4] The scatter plot in Panel A represents the mean scores at the *school* level in 2009 and 2010. The dots in the figure are colored differently according to the school's average score in 2009.[5] Two key patterns emerge from this figure. First, the (nonparametrically) fitted curve that depicts the relationship between mean scores at the school level in 2009 and 2010 lies above the purple 45-degree line.[6] This indicates that average school scores increased after the policy change. Second, the slight hump shape of the curve, relative to the 45-degree line, suggests a more pronounced improvement for mid-performing schools. This trend makes sense: on average, students in lower-achieving schools should not benefit as much from extra time because they lack the knowledge needed (especially when faced with difficult questions) to get the correct answer, while students in top-scoring schools should not need the additional time.[7]

Figure 1: Score and Standard Deviation Comparison Between 2009-2010



Note: The difference in the 45-degree line and the curve is statistically significant with $p < 0.05$. The non-linearity of the relationship in Panel A is confirmed by running a second-degree polynomial specification. The squared term is significantly different from zero at the 5 percent level.

Panel B of Figure 1, using school-level data, shows a reduction (on average) in the standard deviation of scores at the school level after the policy change. Most of this comes from better

[4] The scores represent the average performance across quantitative track subjects: Mathematics, Physics, Chemistry, and Biology.

[5] We dropped observations where the change across two years were outliers, as these were likely due to OCR errors. We also dropped schools where the percentage of quantitative students was less than 0.3, ensuring that we focus on cases where quantitative/science track students are not in the minority.

[6] A Lowess smoother with a bandwidth of .75 was used.

[7] This logic also suggests that the effect of pressure is likely to be heterogeneous in the abilities and difficulty of the questions.

schools, as the blue and green dots (for middle- and high-performing schools) tend to lie below the 45-degree line. More time would move the distribution of scores to the right in each school. However, whether the standard deviation rose or fell would depend on the composition of the students within the school. If only students at the top end of a school are affected by more time, as would be the case in low-performing schools, then the standard deviation is likely to rise as the distribution will be stretched out to the right. If students at the bottom end are affected, as would be the case in high-performing schools, the standard deviation is likely to fall as the distribution shifts to the right at the bottom. This is consistent with what we see in Panel B. The standard deviation on average falls for schools with medium and high performance with more time, but this is not the case for schools with low performance.

The patterns above suggest that time pressure might matter but differentially across students and call for a deeper analysis of the role of time in the decision-making process during multiple-choice exams. To obtain individual-level data on the role of time, we conducted a field experiment with twelfth-grade (senior) students at a high school in Turkey. The experiment replicated the natural experiment described above. We randomly assigned students to treatment (less time-constrained) and control (more time-constrained) groups. All students were given the same multiple-choice exam (comparable to the university entrance exam). To ensure that this was a relatively high-stakes exam for the subjects, we paid students according to their performance. The data from this experiment let us see how student performance responds to more test time and how heterogeneous these effects are by student ability, gender, and question difficulty. As there is a negative marking, we also look at the role of gender as time constraints are relaxed.

We use the data from this experiment in two ways. First, we look at the patterns directly. Then, we develop a structural model that captures the relevant forces at work. This model is specifically designed to shed light on student behavior in multiple-choice tests under time pressure. Building on previous work by Akyol et al. (2022), students decide on the answer to each multiple-choice question based on the signals they get. Students differ in terms of their ability and risk aversion. Student ability impacts the quality of the signal received. The higher the ability, the better the quality of the signal. The distribution of these signals is also influenced by two other factors: the available time and the difficulty of the question. The more time available, the better the quality of the signal, and the more difficult the question, the lower the quality of the signal. Higher risk aversion/ lower confidence requires a higher threshold of the signal needed for the student to choose to answer, which leads to increased question-skipping and, consequently, a reduction in expected scores.

Risk aversion is typically assessed and differentiated from confidence through questionnaire responses aimed at eliciting the extent of risk aversion. We take a different approach: As risk aversion is a primitive, there is no reason for it to change with the extent of time constraints. However, confidence could change with the extent of time constraints, and we expect more time to increase confidence. We use this to distinguish between the two. In other words, as a primitive, risk aversion does not change in the short term, so any observed changes in skipping behavior with changes in time pressure are attributed to changes in confidence levels.

We incorporate time constraints into the model as a component that alters the signal quality

of the correct answer in multiple-choice questions. Building on Akyol et al. (2022), this aspect of our model helps us to understand how signal quality, which is crucial for identifying the right answer, changes with the time allowed, the difficulty of the question, the ability of the student, and gender. Our approach allows us to explore counterfactual scenarios like the presence or absence of negative marking and the difficulty of the exam more easily and at a lower cost than typical decision-making experiments in a laboratory environment.

We find strong evidence that more time increases performance. Heterogeneity follows the pattern described in the natural experiment. Gains are higher for mid-difficulty questions than for low- or high-difficulty questions. Furthermore, where gains rise the most varies by student ability: lower-ability students gain the most in easier questions, while higher-ability students gain the most in more challenging ones. In other words, questions that are within reach for a student are where her performance improves the most when time constraints are relaxed.

The gains mentioned above appear in two forms: an increase in the correct fractions and a decrease in the skipping behavior, suggesting that students' information on the correct answer becomes more precise. We also document a surprising heterogeneity in the gains among men and women. We expected greater improvements in women's scores given the existing literature, which finds that women have lower confidence/higher risk aversion, so they skip when they should not if they were trying to maximize their score. This, it is argued, helps explain their lower performance in high-stakes exam settings (Niederle and Vesterlund, 2010; Arenas and Calsamiglia, 2022). Our prior, consequently, was that more time would raise the confidence of women more than that of men, so that women would skip questions less often, and their scores would thus rise more than men's. Instead, we observe a greater improvement for male students. The fraction correct rises for both men and women, but slightly more for men. The fraction skipped falls, but again falls more for men.

Our structural model helps explain what lies behind this pattern. Students receive a signal that varies in quality according to student ability, question difficulty, and time available. Given the signal they get, they attempt a question whenever their risk aversion and confidence warrant it. While more time does increase confidence, and a bit more for women, the main drivers of the differential response to time of men and women to having more time seems to come from time mattering less for women.

Using our estimated structural model, we examine how different decision-making components contribute to student performance through various counterfactual analyses. We first study how the ability of the exam to sort students (as measured by the rank correlation between the score and ability) changes with time pressure, student ability, and exam difficulty regimes. We find non-monotonic responses to sorting for all three of the above. This comes from the fact that sorting is highest when the exam is set to match the student's ability. Second, we study how alternative negative marking regimes (no penalty versus harsher penalty) affect students' responses in the presence of time pressure. Higher penalties improve sorting, and so do tighter time constraints. This makes sense because negative marking reduces guessing, and skipping is more likely to occur under more time pressure.

Our work also provides a novel rationale for ranking women by comparing them to other

women and men to men rather than together, as is almost universally done in practice on the grounds that doing so results in better sorting by ability. This makes sense as information processing (the parameters of the signal extraction function) differs by gender, so sorting on the basis of exam results is more accurate when done within gender.

1.1 Related Literature

Our paper contributes to several areas of the literature. First, we contribute to the literature by studying the impact of time pressure on risky choices, such as choosing among different gambles with varying expected utilities. Mostly using laboratory experiments, this literature provides evidence of the change in risk-taking with varying time constraints. It finds that subjects are less willing to take risks in high-time-pressure environments. This is in line with what we find: students skip more under greater time pressure (Ben Zur and Breznitz, 1981; Diederich et al., 2020; Hausfeld and Resnjanskij, 2018; Kocher et al., 2013; Nursimulu and Bossaerts, 2014). Hausfeld and Resnjanskij (2018) introduces exogenous opportunity costs of decision time, which can be thought of as analogous to stricter time constraints. They show that decision errors (mistakes) increase with high opportunity costs (greater time pressure), while risk aversion does not change with low or high opportunity costs. The first is in line with our results, and the second supports our assumption that risk aversion is primitive and, thus, would not change with time pressure.

Second, there is a growing literature on decision-making and response times. One strand of this literature examines the relationship between decision time, accuracy, and optimal timing of choices by incorporating cognitive models from psychology literature to decision-making in an economic context where the decision process is costly because of time constraints. (Chabris et al., 2009; Fudenberg et al., 2018; Kocher and Sutter, 2006; Sunde et al., 2022; Wang and Xu, 2015; Wilcox, 1993). Using data on online chess tournaments, Sunde et al. (2022) shows that faster decisions are correlated with higher performance, but the paper has no theoretical component. We provide a theoretical framework and conduct a field experiment to shed light on the issue and generate relevant data. This data is then used for both reduced form and structural estimation of the model. Another strand of the literature endogenizes the time spent on a question (see Fudenberg et al., 2018). They endogenize the time spent on an item (when the marginal cost per unit of time spent is given) as a function of the initial difference in the signals between choices. Their framework predicts that agents who receive stronger signals about the correct option decide faster, while those with weaker signals take longer. This selection effect leads to an observed pattern where faster decisions tend to be more accurate on average. However, this mechanism is not our focus. Instead, we assume a constant time allocation per question because the exam we analyze is paper-based, not computer-based, and thus, we could not measure the exact time each student spent on individual questions.

Third, we also contribute to the growing body of work on the gender gap in high-stakes environments. This includes theoretical and empirical models that analyze the role of several factors such as competition, stress, risk preferences, and social preferences on the gender gap in performance (Akyol et al., 2022; Baldiga, 2014; Cai et al., 2019; Arenas and Calsamiglia, 2022; Croson and Gneezy, 2009; Ellison and Swanson, 2021; Franco and Gomez-Ruiz, 2023; Hakimov et al.,

2023; Iriberri and Rey-Biel, 2021; Montolio and Taberner, 2021; Niederle and Vesterlund, 2010; Ors et al., 2013; Pekkarinen, 2015). A recent paper Galasso and Profeta (2024) investigates the effect of time pressure on the gender gap in math tests in an environment where there is no competition and no negative marking for incorrect answers. Their finding shows a reduction in the gender gap when time pressure is reduced or eliminated. Our paper contributes to this work by developing and estimating a rich model capable of explaining how various features of the exam affect exam performance when stakes are high and incorrect responses are penalized. We explore heterogeneity in responses by gender to see how time constraints impact the gender gap in performance. The lower performance of women is often attributed in the literature, see Baldiga (2014), to women being more risk-averse/less confident and so skipping too often. As more time would tend to increase confidence, the prior might be that more time would reduce the gender gap in performance. Our findings suggest the opposite: more time does improve confidence, but more so for men, so the gender gap increases.

Lastly, our paper adds to the existing body of research in psychometrics and education that looks at the factors affecting the power of exams to sort students by ability (see Bridges, 1985; Feinberg, 2004; Goldhammer, 2015; Jacob and Rothstein, 2016; Lu and Sireci, 2007; Onwuegbuzie and Seaman, 1995; Wild et al., 1982). Our paper builds on this work by introducing a theoretical framework and quantifying it using our experimental data to assess the impact of time pressure on multiple-choice exam test performance. Moreover, we extend our contribution to the education literature through policy simulations demonstrating the role of exam difficulty, time constraints, the number of questions, and negative marking in affecting the power of the exam in sorting students by ability. Our findings provide valuable insights for policymakers in exam design.

The remainder of the paper has the following structure. In Section 2, we describe the experiment and results. Section 3 outlines the theoretical framework that outlines the student's decision-making in tests, and Section 4 its estimation, identification, and validation. Section 5 illustrates how test time, exam difficulty, and negative marking can be combined for effective student sorting, and Section 6 concludes.

2 The Experiment

This section outlines the experiment conducted. We document and interpret the data patterns that need to be respected when constructing an estimable model.

2.1 Experimental Design

The experiment was conducted in the Fall of 2022 with senior high school students in a Turkish high school. They take weekly practice exams at their high school. These practice tests familiarize them with multiple-choice exams with negative markings. However, since the experiment was nine months before the exam, some of the material may not have been covered in their classes yet, potentially leaving them unprepared for certain topics.

The experiment was conducted in a real exam setting. Students were randomly assigned to

classrooms and seats, while proctoring was provided by the school's teachers.[8] Additionally, we provide payment for performance that ensures that participants take the experiment seriously.[9]

We enrolled 91 students from the grade 12 cohort to participate in the exam, of whom 83 ultimately took it.[10] The participants were randomly assigned to one of two groups: a control group, in which the total test time was shorter, and a treatment group, in which the total test time was longer. We refer to the control group as the time-constrained group and to the treatment group as the time-relaxed group. Both the time-constrained and time-relaxed groups had four absent students. 42% of the students who signed up were female, so we employed stratified sampling by gender to obtain our sample in each group. Consequently, 40% of the students in the time-constrained group and 43% of the students in the time-relaxed group were female. Table A1 in the appendix reports summary statistics and the observed balance between the two groups based on the survey we collected from the students before the exam. These results show that the background characteristics of the control and treatment groups do not differ significantly from each other.

The exam we administered consisted of multiple-choice questions, each with five possible choices, only one of which was correct. Correct answers were awarded one point, wrong answers were penalized with a deduction of 0.25 points, and skipping a question received zero points. Each student received the same booklet designed to assess their proficiency in Turkish, Mathematics, Science, and Social Science. The booklet comprises a total of 48 questions, with 12 questions from each of the four subjects.[11] Therefore, our unit of observation is student-by-question, resulting in a sample size of 3,984 (83 students × 48 questions). The fraction of questions in each section tracked the fractions in the real exams in both 2009 and 2010, though we reduced their number.[12] We chose to limit the number of questions so that the exams would not need too much time from the students.

The time-constrained group was required to complete the test in 40 minutes, whereas the time-relaxed group was given 75 minutes.[13] Though the exam structure was known to students before the exam, they were informed of the time available just before taking the exam.

To ensure that our results corresponded to a high-stakes exam, we rewarded students according to their rank in the exam results within their respective groups. Those ranked in the top three were given 30 USD.[14] The reward decreased by 3 USD for each subsequent set of three ranks.

[8] The random seat allocation is done by teachers as part of the instructions given to the school for administering the exam.

[9] It is well understood, see Akyol et al. (2021), that the results of low-stakes tests are likely to be downward biased and the rankings inaccurate. Students have no reason to even try to do well on low-stakes exams since, by definition, they do not matter. Our payments are large enough to make our experiment matter.

[10] Parental consent and self-consent received from 91 students.

[11] In our experiment, students were given two booklets sequentially. The first booklet assessed their general knowledge in Turkish, Mathematics, Science, and Social Science, while the second booklet measured advanced knowledge in the same subjects. Together, the booklets contained a total of 100 questions. After the time allocated for the general knowledge section ended, students received the second booklet. In this paper, we focus only on the general knowledge tests, as most students skipped the advanced questions due to the material not yet being covered in their classes.

[12] The number of questions per subject in the real exam was 30 in 2009 and 40 in 2010.

[13] This was also roughly the increase in time per question in the college entrance exam in 2010, the natural experiment we used to motivate our study.

[14] 30 USD corresponds to around 13 percent of the monthly net minimum wage in Turkiye during the relevant time

Specifically, the next three students (ranked 4-6) got 27 USD, the following three (ranked 7-9) got 24 USD, and so on. This design mimics the incentives of the actual exam, where higher scores provide more options for a student and are therefore highly preferred.

2.2 Experiment Results

2.2.1 Average Treatment Effects. We begin by analyzing the outcomes of the experiment to better understand the effect of extended time on students' exam performance. The rows give the fraction correct, wrong, conditional correct, and skipped answers, as well as the overall score. Each entry of Column 1 shows the average for the time-constrained group, while Column 2 shows the average for the time-relaxed group. The raw difference between the two groups is given in Column 3. Standard errors are presented in parentheses. In Column 4, we report the minimum detectable effect in absolute values using a one-sided t-test ($1.65\times$ standard errors presented in column 3). In Column 5, we report the p-values obtained non-parametrically with the Mann-Whitney U Test.

What might drive the performance of a student? It is reasonable to expect that the time available to reflect on the question and the student's ability are inputs. Thus, we would expect that a student with a given ability would be more likely to get a question correct, conditional on attempting it when more time is available. Furthermore, given negative markings, students would skip a question if they were unsure about their answer. As they are more likely to be sure of their answers with more time, we would expect a decrease in skipped questions as time constraints become less binding.

Our findings in Table 1 are in line with this simple intuition. The proportion of skipped questions decreases by nine percentage points, and this is significant at $p < 0.05$. The accuracy of the response, when the attempt is made, also rises slightly, although the increase is not significant.[15] Consequently, the total score rises with more time, and this difference is significant.[16] These facts are consistent with students having better information about the correct answer when they are given more time. If more time to think improves the accuracy of the signal, students will choose to answer more often, so the fraction that is skipped falls. The increase in the fraction conditionally correct could go either way as the additional questions answered when time pressure is relaxed tend to be those the student is less sure of.

What happens to the fraction wrong is also in line with such a story. The share of wrong answers falls slightly, but this change is not significant. This makes sense; on the one hand, the questions that would have been attempted even with less time are more likely to be correct, but on the other hand, the questions answered only because of having more time are the marginal ones and are less likely to be correct.

period.

[15] This is not unexpected: accuracy would be expected to increase for the questions which would have been attempted had the time given been less (which would raise accuracy), but the marginal questions which are attempted only because time is more would have lower accuracy than that of infra marginals ones. This would lower overall accuracy. The effect on overall accuracy could thus go either way, though the total score would be expected to increase.

[16] Note that with negative marking, the expected score is $N_c - N_w(-0.25)$ where N_c and N_w are the numbers of correct and wrong answers, respectively.

Table 1: Average Treatment Effects

	Control (1)	Treatment (2)	Difference (3)	Detectable Effect (4)	p-value (5)
Correct	0.67	0.76	0.09	0.03	0.00
	(0.11)	(0.10)	(0.02)		
Wrong	0.16	0.15	-0.01	0.03	0.55
	(0.07)	(0.07)	(0.02)		
Cond. Correct	0.81	0.83	0.02	0.03	0.12
	(0.08)	(0.08)	(0.02)		
Skip	0.17	0.08	-0.09	0.03	0.00
	(0.11)	(0.07)	(0.02)		
Score	0.63	0.72	0.09	0.03	0.00
	(0.11)	(0.11)	(0.02)		
Observations	42	41			

Notes: Standard errors are in shown in parentheses.

2.2.2 Heterogeneity by Gender. In this sub-section, we focus on differences in the effects of extra test time by gender, ability, and difficulty of the questions. This is important for two reasons. First, it provides more patterns that our model should be capable of reproducing. Second, it has implications for the efficacy of a test in sorting students correctly, something we explore later on. Note that while having a higher score is beneficial overall, what matters for placement is rank. If groups are different regarding their response to time pressure, sorting can worsen. Understanding what might lie behind such heterogeneity is crucial to understanding both who benefits and who loses from having extra time on the exam, as well as how differently constructed exams perform in terms of sorting students.

We begin by looking at the differential responses of men and women to having more time available per question. Table 2 shows the treatment effects on different outcomes for each gender. Both women and men see improvements with extra time - both groups' fraction of correct answers and overall test scores increase, while skipping decreases. However, the increase in the fraction correct and score is not significant at the 5 percent level for women. The reduction in skipping is larger for men (10.1 percentage points) than for women (7.5 percentage points). Note from Table 3 that men tend to skip less often than women to begin with. Despite this, with more time, their decrease in skipping is more than that of women. This is what drives the greater increase in score for men than for women, as skipping reduces the expected score since the penalty is actuarially fair. The fraction of wrong answers does not change significantly for either group. Table 3 looks at the difference in the four outcomes by gender in both the treatment and the control group. Note that the differences do exist: women do skip more than men, get a lower fraction wrong (consistent with their being less confident/more risk averse), and a lower fraction correct, which is to be expected as they skip more often. However, none of these differences are significant at the 5 percent level. While test scores are similar for men and women in the control group, the men fare better than women with more time, a consequence of the probability of skipping dropping more for men.

Table 2: Nonparametric Tests

	Men (1)	Women (2)
Correct	Δ = 0.116 (p = 0.000)	Δ = 0.062 (p = 0.121)
Wrong	Δ = -0.015 (p = 0.261)	Δ = 0.012 (p = 0.652)
Skip	Δ = -0.101 (p = 0.000)	Δ = -0.075 (p = 0.042)
Score	Δ = 0.119 (p = 0.000)	Δ = 0.059 (p = 0.202)

Notes: This table shows the treatment effects and nonparametric test results for each outcome and separately for male and female subgroups. The $p-$ values are obtained through nonparametric Mann-Whitney U tests.

Table 3: Across Gender Comparison

		Control (1)	Treatment (2)
Correct	**Men**	0.672	0.787
	Women	0.668	0.730
	Difference	Δ = 0.004	Δ = 0.057
		(p = 0.878)	(p = 0.050)
Wrong	**Men**	0.168	0.153
	Women	0.146	0.158
	Difference	Δ = 0.022	Δ = -0.005
		(p = 0.309)	(p = 0.739)
Skip	**Men**	0.161	0.060
	Women	0.186	0.112
	Difference	Δ = -0.025	Δ = -0.052
		(p = 0.504)	(p = 0.035)
Score	**Men**	0.630	0.749
	Women	0.631	0.691
	Difference	Δ = -0.001	Δ = 0.058
		(p = 0.778)	(p = 0.078)

Notes: This table shows the score outcomes for both genders in control and treatment groups as well as the score differences. The $p-$ values for the differences are obtained through nonparametric Mann-Whitney U tests.

Differences in cognitive processing between men and women can help explain the observed gender disparities in the effect of additional time on exam performance. In the literature, it is suggested that women tend to perform better on verbal tasks and exhibit stronger selective attention and rapid access to long-term memory, while men outperform in spatial reasoning, mental rotation, and problem-solving under time constraints (Halpern and Wai, 2019; Ramos-Loyo

et al., 2022). Studies using EEG data have shown that cognitive processing speed and neural efficiency differ between sexes, particularly in executive functions and complex problem-solving tasks. These differences may interact with time constraints in test settings: Women's advantages in structured, language-based tasks may make them less dependent on additional time, while men's strengths in abstract reasoning may benefit more from extended problem-solving periods. Additionally, research on cognitive endurance suggests that performance declines over time are more pronounced among disadvantaged groups, highlighting the role of sustained effort in test-taking (Brown et al., 2024). Furthermore, the literature indicates that men exhibit greater variance in cognitive performance, which means that some may experience sharper gains when given more time, whereas women's performance tends to be more stable (Arden and Plomin, 2006). These findings align with our results, suggesting that cognitive differences, rather than purely test-taking strategies, may partly explain why the additional time benefits men disproportionately.

2.2.3 Heterogeneity by Ability and Difficulty Next, we look at heterogeneity by question difficulty and student ability. To do so, we need to construct measures for them. The standard way of doing so is to use the Rasch item response model. This boils down to running the likelihood of the answer being correct depending on the question's fixed effect and the individual fixed effect. We do not do this because doing so would give us biased estimates of ability. If women, for example, underperform on multiple-choice exams and perform better on open-response questions, then using the Rasch approach to estimate ability in a multiple-choice exam will bias the estimate of women's ability downward. To account for such a possibility, we use all the information we have on performance. We estimate the ability of the student using information on past mock exams (eight of them) taken at the school, as well as five 11th-grade subject-specific GPAs. The subject-specific GPAs represent the weighted average of scores from at least two exams, including open-response questions. These scores should be minimally influenced by students' self-confidence or risk aversion due to the absence of negative marking in these exams. As there is a negative marking on the mock exams, the total score from these could be driven by both ability and risk aversion, complicating their use as an ability measure. Thus, we have 13 observations for each student.

We estimate ability as follows. To ensure comparability across exams, we first normalized the scores from each of the eight mock exams as well as the GPA for each subject. We use the min-max normalization to rescale these scores to a range between 1 and 2, ensuring consistency with the other model components. We then estimate the student-specific fixed effects using these normalized scores using the regression model $Y_{ie} = \beta_1 + \beta_2 D_e + \delta_i + \epsilon_{ie}$, where D_e is 1 if the type of the exam score is the GPA and 0 if not. The inclusion of the score-type dummy, D_e, allows for differences in performance in GPA and other scores.[17] δ_i serves as the proxy for student ability. In this way, our regression model controls for the type of exam. The individual fixed effects generated in this manner minimize the impact of risk aversion on the estimated abilities. The distribution of the ability measure used in our analyses is given in Figure A1.[18] The difficulty measure is

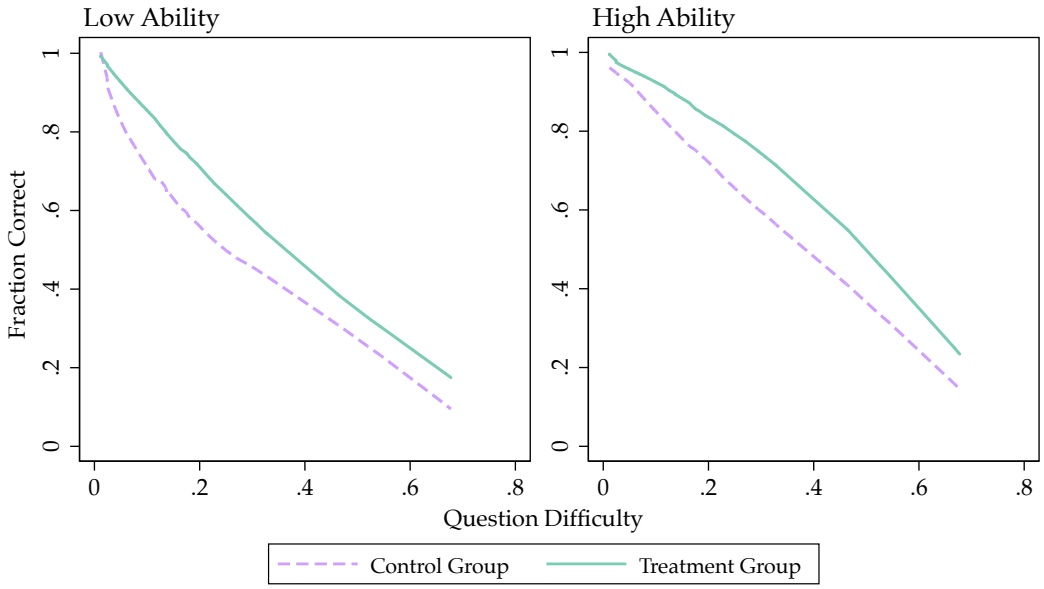
[17] Allowing for women to do differently in GPA scores than the exam scores, that is, having a Female dummy and its interaction with D_e in the regression, was not significant. For this reason, we only included the exam-type dummy.

[18] The correlation between this baseline ability measure and the score (both time constrained and not) in the experiment

estimated by question-fixed effects.

Once we have measured ability and difficulty, we can ask whether the gain from extra time varies by question difficulty and across different ability groups. Figure 2 presents the fraction of questions answered correctly for different difficulty levels of the questions. The left subfigure shows the results for the low-ability group, while the right subfigure does the same for the high-ability group. We define high and low ability as those above and below the median, respectively. High and low question difficulty are analogously defined. The fraction of people who get the question correct falls with the difficulty of the question for both the high and low-ability groups. In addition, for both figures, the difference in the fraction correct is largest for questions of intermediate difficulty. This observation makes sense as more time is unlikely to matter if the question is too easy or too difficult. Figure 3 depicts the treatment effects by gender. These are the analogs of the differences in the curves in Figure 2 when separated by gender.[19] Note that the treatment effect is the largest at a difficulty of around .15 for low-ability students and at a higher difficulty level of around .3 for high-ability students in Figure 2. This is so irrespective of gender, though the peak for males occurs at a higher level of question difficulty than for females in Figure 3. This observation suggests that time has the greatest impact when questions are difficult relative to student ability, but not out of reach.

Figure 2: Fraction Correct by Question Difficulty and Ability



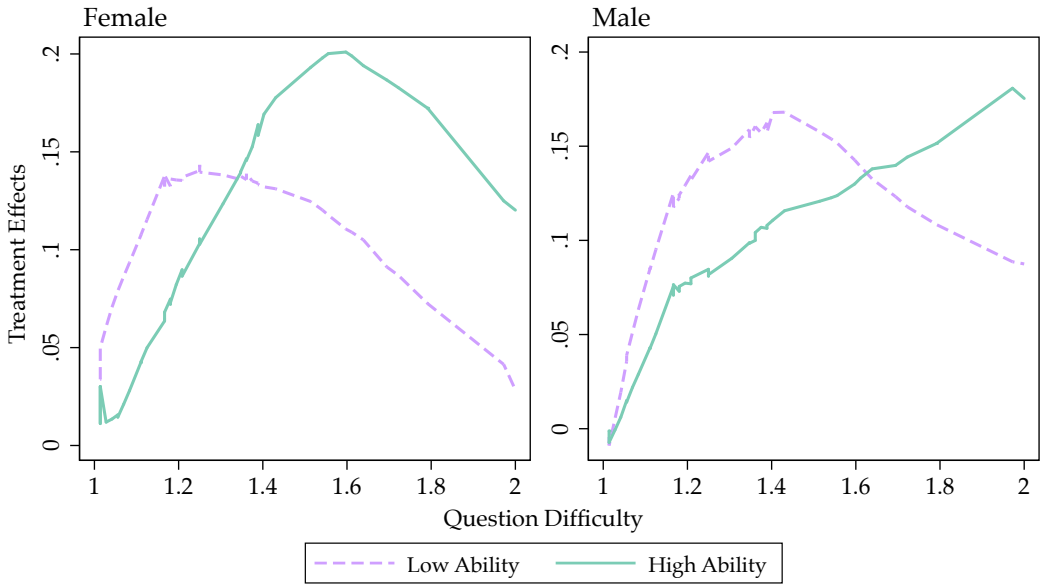
Note: This figure presents the fraction of the correct answers across control (purple curve) and treatment (green curve) groups for low (left panel) and high-ability (right panel) groups.

To summarize, the key patterns in the data are that more time improves the fraction correct and reduces skipping. However, there are differences by gender. Women skip more than men to begin with, which has been (perhaps wrongly) seen as coming from women being more risk-averse than men. Women also gain less from having more time than men, as reflected in men reducing skipping by more than women. This is consistent with time being more important as an input for

is 0.65 consistent with the need for a more careful approach to measuring ability than the standard Rasch one.

[19] We use a lowess smoother as a function of question difficulty.

Figure 3: Treatment Effects by Question Difficulty and Ability



Note: This figure presents the association between question difficulty and improvement with more time across different ability groups. The purple curve drawn nonparametrically shows this association for lower-ability students in the distribution, and the green curve shows the relationship for the higher-ability students. The x-axis displays the question difficulty, which is proxied by the fraction of incorrect answers. The y-axis represents the treatment effect, predicted non-parametrically. Both curves use a lowess smoother with a bandwidth of 0.8.

men. Time also has a greater effect on questions that are within reach for a student. These patterns will have implications for the efficacy of exams in sorting students by ability, which we analyze below.

3 The Model

In this section, we develop a model of the decision-making process in multiple-choice tests with time constraints. We begin by outlining the model. Our objective is to model the student's decision-making process with a view to estimating the structural parameters that drive students' choices. The model specifies the process underlying the student's choices. The estimation yields parameters, some of which can differ by gender and treatment. These parameter estimates are obtained by matching the heterogeneity in outcomes by gender, student ability, and question difficulty observed in the experimental data with those generated by the model through indirect inference.

3.1 The Environment

We present a model that describes how a rational decision-maker, a student in this case, solves a multiple-choice exam consisting of Q questions where each question has $\bar{t}$ units of time allocated to it. For each question, the student must choose whether to answer (A) or skip/not answer (NA), knowing that incorrect answers carry a penalty.

If a student chooses to answer the question, he chooses between K possible answers for the

given question. For each question, there is only one correct answer, denoted by c. A correct answer gives s points, and an incorrect one is penalized by $\frac{s}{K-1}$ points. Note that the penalty is actuarially fair as the expected value of just guessing is zero. Skipping a question yields zero points, making it a safe option.

Let π^k be the probability that the option k is chosen, and the probability of selecting the correct answer is denoted by π^c. We assume each option has a strictly positive probability of being chosen, conditional on spending t units of time. In other words, $\pi^k > 0 \ \ \forall k \in \mathcal{K} = \{1, 2, \ldots, K\}$.

3.2 The Decision Process

We assume the student spends an equal amount of time on each question.[20] The student approaching a question with K choices observes a $K-$dimensional signal (Z^k) for $k \in \mathcal{K}$. The higher the signal, the more likely it is that the choice is correct. Hence, the student chooses the choice with the highest signal. We assume that each Z^k follows a Pareto distribution F^k with support $[m_k, \infty]$ and shape parameter β^k. Hence, the density of the signal Z^k is $\frac{\beta^k m_k^{\beta^k}}{Z^{k\,\beta^k+1}}$.[21] As in Akyol et al. (2022), on which we base our model, we assume the lower bound of the signal support is common for all choices to prevent the possibility of a perfectly informative signal. If the lower bound for the correct answer differed from that for incorrect ones, then a signal between these bounds would be perfectly informative.

We assume that all the incorrect choices get a signal drawn from the same distribution with a different shape parameter. More precisely,

Assumption 1. *$\beta^k = \alpha$ for incorrect choices, while the distribution of the correct signal has $\beta^k = \beta^c$.*

Using Bayes' rule, the probability of choosing the correct answer given the signal vector $\mathbf{Z} = \{Z_1, Z_2, ..., Z_K\}$, can be written as:

$$\pi^c = \mathbb{P}\left(c \mid \mathbf{Z}\right) = \frac{\mathbb{P}(\mathbf{Z} \mid c).\mathbb{P}(c)}{\mathbb{P}(\mathbf{Z})} \tag{1}$$

Note that the unconditional probability that choice 1 (or any other choice) is correct is $1/K$, a constant, and can be ignored. The expression for the first term in the numerator when the first option is correct, i.e., for the case $c = 1$, is given by:

$$\mathbb{P}(\mathbf{Z} \mid c = 1) = \frac{\beta^c m^{\beta^c}}{Z_1^{\beta^c+1}} \frac{\alpha m^\alpha}{Z_2^{\alpha+1}} \cdots \frac{\alpha m^\alpha}{Z_K^{\alpha+1}}. \tag{2}$$

Substituting Equation (2) into Equation (1), we obtain:

$$\mathbb{P}\left(c \mid \mathbf{Z}\right) = \frac{\frac{1}{Z_c^{\beta^c+1}} \prod_{j \neq c} \frac{1}{Z_j^{\alpha+1}}}{\sum_k \left(\frac{1}{Z_k^{\beta^c+1}} \prod_{n \neq k} \frac{1}{Z_n^{\alpha+1}} \right)}. \tag{3}$$

[20] In future work, where we will have a computerized setup, not a paper one, as in this experiment, we hope to relax this assumption and focus on time allocation decisions.

[21] The mean signal is $\frac{\beta^k m_k}{\beta^k - 1}$ which is a decreasing function of β^k. In addition, the variance of the signal, $\frac{m_k^2 \beta^k}{(\beta^k-1)^2(\beta^k-2)}$, is also decreasing in β^k. Thus, the mean, relative to the variance, is increasing in β^k.

Note that m cancels out in the numerator and denominator so that it does not affect the probability of a correct answer. Finally, using the fact that $\frac{1}{Z_c^{\beta^c+1}} = \frac{1}{Z_c^{\alpha+1}} Z_c^{\alpha-\beta^c}$, we can further simplify the above expression as follows:

$$\pi^c = \mathbb{P}\left(c \mid \mathbf{Z}\right) = \frac{{Z_c}^{\alpha-\beta^c}}{\sum_k Z_k^{\alpha-\beta^c}}. \tag{4}$$

There is nothing special about the first answer being the correct one, so this expression also gives the probability of getting the correct answer. From Equation (4), we see that what matters for the informativeness of the signal is the difference between α and β^c. We assume that $\alpha \geq \beta^c$ so that c is chosen weakly more often because the mean signal for c will be weakly higher than the mean signal for the incorrect choices.[22] Thus, a given signal is more likely to be correct and more informative as $(\alpha - \beta^c)$ rises, while the decision maker is more clueless as these parameters come together.[23] It is not the mean levels but the difference in the mean levels of the signals for the correct and wrong answers that are key. For this reason, we normalize β^c to be 1.

Hence, the expected payoff from answering, given π^c:

$$\mathbb{E}\left[U^A\right] = \pi^c U(s) + (1-\pi^c) U\left(-\frac{s}{K-1}\right).$$

The student then chooses to answer or not, depending on which has the higher payoff. That is, the student answers if

$$\mathbb{E}\left[U^A\right] > U(0) \iff \pi^c > \frac{U(0) - U(-\frac{s}{K-1})}{U(s) - U(-\frac{s}{K-1})} = \bar{\tau}. \tag{5}$$

From Equation (5), it follows that the threshold parameter $\bar{\tau}$ depends on the concavity of the utility function, which represents confidence/risk aversion. Our data do not allow us to distinguish between confidence and risk aversion, as greater confidence and lower risk aversion both reduce the threshold $\bar{\tau}$.[24] We let this threshold change with the time allocated per question on the test. As risk aversion is a structural primitive and unlikely to change with time constraints, we argue that any observed change in the cutoff with varying time available per question comes from the change in confidence.

We call $\bar{\tau}$ the certainty cutoff. It is the probability threshold at which a student decides whether to answer a question. Specifically, a student answers if the probability of the best choice (i.e., the option with the highest signal) being correct exceeds this cutoff. Thus, $\bar{\tau}$ reflects both risk aversion and confidence. We allow this threshold to vary with gender and time constraints.

[22] Recall that the individual chooses the choice with the highest signal as this is most likely to be the correct one.

[23] Note that informativeness refers to the Blackwell partial order. Take an example with three choices. Under our assumptions, if choice $i = 1$ is the correct one, the probability 1 will be picked is $\pi^c > 1/3$. The probability that either of the other two (the wrong ones) is picked is (1-π^c)/2. Thus, if $i = 1$ is the correct one, an increase in π^c moves the point (π^c, (1-π^c)/2, (1-π^c)/2) towards the edge of the simplex or $(1, 0, 0)$. Similarly, for any other i being correct. As a result, the posterior with a higher π^c dominates (in the convex order) the posterior with a lower π^c. See Theorem 4.1 in Liang (2023). See Online Appendix A for additional details.

[24] One can use survey-based designs to elicit risk preferences to pin down risk aversion as in Baldiga (2014).

3.3 Modeling Time Constraints

In this section, we explain how we incorporated time constraints into our model. We allow α for each gender to depend on the time per question, t, the ability of the individual, θ, and the difficulty of the question, d, as follows.

$$\alpha = g(\theta, d, t) \tag{6}$$

In choosing our functional form for estimation in the following section, we ensure that α, the function representing the capacity to distinguish correct from incorrect answers, increases with the ability and the time spent on a question, while it decreases with the difficulty of a question. Critically, our specification allows for time to have a greater effect on the ability to distinguish between correct and wrong answers when the question is within reach of the student. This ensures that our functional form is flexible enough to capture the patterns in the data shown in Figure 3.

We parametrize Equation (6) as follows:

$$\alpha = \gamma_0 + \theta_i^{\gamma_1} d_q^{\gamma_2} \tilde{t}_i^{\gamma_3 \exp\{-\gamma_4|\theta_i - d_q|\}} \tag{7}$$

where $\tilde{t}_i = \gamma_5 + t_i$ to allow for enough flexibility in the form.[25] The term γ_0 just acts to scale the function. The second part, $\theta_i^{\gamma_1} d_q^{\gamma_2} \tilde{t}_i^{\gamma_3 \exp\{-\gamma_4|\theta_i - d_q|\}}$, can be thought of as a standard production function where the inputs are the ability of student i, the difficulty of question q, and time per question, t. We expect the coefficient on ability to be positive and on difficulty to be negative. The coefficient on time is more subtle. Its form allows it to be larger or smaller depending on the absolute value of the difference in ability and difficulty. This gives us the flexibility to capture the data pattern we see, namely that time seems to matter more for questions that are within reach of the student, the relationship depicted in Figure 3 in Section 2. The functional form in Equation (7) extends Akyol et al. (2022) by incorporating time constraints into their model.

To summarize, this specification has the capacity to replicate patterns seen in the experimental data. First, reducing time pressure increases both the likelihood of answering and, conditional on answering, the likelihood that the answer is correct. The effect of time on these outcomes varies with the difficulty of the question relative to the student's ability. On the basis of the data, we expect it to be the largest when difficulty and ability are closest. Second, more time improves the signal quality. This increases the likelihood of a correct response and diminishes the probabilities of skipping, conditional on the level of the certainty cutoff, τ. However, for the fraction wrong, there are two opposing effects at play with a change in the available time. As fewer questions are skipped, the fraction of incorrect answers rises. On the other hand, answered questions are more likely to be correct, which reduces the fraction of incorrect answers. Time constraints can affect the certainty cutoff, which is estimated from the data. A priori, we would expect it to increase confidence, thereby reducing the cutoff. This should result in a decrease in skipped responses and an increase in both the fraction correct and incorrect. The strength of these patterns in the data helps identify the parameters of the model.

[25] In our experiment t is either ~ 0.8 or ~ 1.6. If we did not incorporate a parameter that would scale t, we would have $\alpha(.)$ increasing in t if t is more than unity and decreasing in t if it is less than unity.

4 Identification and Estimation

We use indirect inference to estimate our structural model. We need to estimate the parameters governing signal production, the function $\alpha(.)$, as well as the certainty cutoffs. We allow the certainty cutoffs to change as time constraints change. We do so as the certainty cutoffs reflect both risk aversion (which does not change when time constraints become more or less binding) and confidence (which may change with time allotted). Estimation is performed separately by gender.

4.1 Ability-Difficulty Measures.

Ability and difficulty enter the signal production function, which governs the probability of a question being correct. We rely on the measures of student ability and question difficulty introduced in Section 2.2.3.

4.2 Signal Production and Certainty Cutoffs

We jointly identify the parameters of signal production and the certainty cutoffs. We allow all the parameters we estimate to differ by gender. We do so to allow for differences in the way that the inputs (time allowed, ability, and difficulty) enter into the production function rather than assuming they enter in the same manner. Note that in addition to the parameters in the production function, we also allow cutoffs to differ by gender. This captures differences in risk aversion/confidence by gender.

4.3 Estimation

To implement our identification approach, we need data on the empirical distribution of student responses and how they vary with time, question difficulty, and student ability. We use the data from our experiment to estimate the structural parameters of the model. In our experiment, for each individual, we observe whether the question was attempted or not and, if attempted, whether the answer was correct or not. The structural parameters we need to estimate are the parameters in $\alpha(.)$, and the cutoff parameters, $\bar{\tau}$, for each gender.[26] Recall that we parametrized α as given in Equation (7). We assume that students allocate their time equally across questions, i.e., $t_i = \bar{T}_i/Q$. For the time-relaxed group, the total time is 75 minutes, while it is 40 minutes for the control group. Thus, for the control group, $t = 40/48$, and for the treatment group, $t = 75/48$.

4.3.1 Estimation Strategy Overview. The estimation strategy uses indirect inference.[27] For each question and each individual, and for a given parameter vector, we obtain the values of the targeted moments in the simulated data. We compare them to those in the experimental data and

[26] Ideally, a nonparametric approach would be best for our setting due to the complex relationships between time, ability, and difficulty. However, data limitations prevent this.

[27] We could have used a Maximum Likelihood approach, but the small sample size results in large standard errors. This is why we chose to use indirect inference.

choose the parameters that best match the two.[28]

The parameters included are the parameters in the function $\alpha(.)$ and the certainty cutoffs under less and more time constraints. We allow all of these to differ by gender, and we allow the certainty cutoffs to differ by time as well. Our estimates of these parameters are in Table 4. The targeted moments are as follows. We divide students into two groups on the basis of their estimated abilities. We call these groups low ability if they are below the median ability and high ability otherwise. As the treatment effects for high and low-ability students, as depicted in Figure 3, are hump-shaped in question difficulty, we need at least a three-way classification for the difficulty to capture it. For this reason, we separate questions into quintiles in terms of difficulty. We then calculate the fraction correct and the fraction skipped (2) by difficulty quintile (5) for high-ability and low-ability students (2) and for the treatment and control groups (2). This gives us 5x8 moments for each gender.[29]

The parameters in $\alpha(.)$ are identified as follows. γ_0 is a scaling parameter and is identified through the baseline outcome. γ_1 is the elasticity of probability correct with respect to ability. It is identified by the difference in the probability correct in high and low-ability agents. Similarly, γ_2 is the elasticity of probability correct with respect to question difficulty. It is identified by the rate at which the fraction correct rises as questions become easier. γ_3 captures the baseline influence of time on performance and anchors the fraction correct. γ_4, the coefficient on the discrepancy in ability and difficulty, is captured by the differential impact of time across different ability and difficulty pairs. The cutoff parameters are identified through the fraction of skipped questions for given time constraints and gender.

Formally, the routine is as follows: for a given parameter vector $\mathbf{\Theta}$, we simulate the model, mainly the probability of getting correct and the probability of skipping at the question and individual level, compute a vector of moments $\mathbf{m}^{\text{model}}(\mathbf{\Theta})$ and compare them with the equivalent vector of moments in the data $\mathbf{m}^{\text{data}}$. We search for the parameter vector $\mathbf{\Theta}$ that minimizes the distance between the model generated and the empirical moments, obeying the loss function $\mathcal{L}(\mathbf{\Theta}) = \left(\mathbf{m}^{\text{model}}(\mathbf{\Theta}) - \mathbf{m}^{\text{data}}\right)' W \left(\mathbf{m}^{\text{model}}(\mathbf{\Theta}) - \mathbf{m}^{\text{data}}\right)$ where $\mathbf{W}$ is a positive definite weighting matrix. [30]

4.3.2 Parameter Estimates. Table 4 presents the parameter estimates from the indirect inference procedure and the bootstrapped standard errors. Panel A reports the estimates of the parameters defining the signal production function for both genders. Panel B reports the certainty cutoffs. The estimates show that all the estimated parameters, other than the discrepancy factor for women, are significantly different from zero. The first four parameters in panel A are significantly different by gender. The remaining two parameters, however, do not differ significantly by gender. The scaling factor for women is lower than that for men. This suggests that the signal quality for

[28] A detailed description of the estimation procedure can be found in Appendix B.

[29] Note that our estimation is not dependent on a particular functional form for utility. However, for our counterfactual analyses, we adopt a CARA utility function. This enables us to quantify the cutoff levels under various negative marking regimes.

[30] We search for the best-fitting parameter vector on a sequence of finer grids, followed by a local optimization procedure starting from a subset of best-fitting points from the narrower grid (obtained by a global Halton search procedure).

women is lower than that for men. This aligns with observed performance differences between women and men on these exams. The proficiency weight is higher for women, which means that higher ability matters more for women when determining signal quality. The difficulty weight is larger for women. As this is a negative number, women's signal quality is more adversely affected by difficulty. The time weight is lower for women, so more time improves women's signal quality by less.

Panel B reports the certainty cutoff estimates for men and women under both treatment and control. It suggests that the certainty cutoff of men is lower than that of women, but not significantly so. This is in line with the results in Akyol et al. (2022). The extra time allotted lowers the certainty cutoff significantly for both genders and more for women.[31] This is exactly what we had expected: confidence rises more for women than for men with more time. This fall enables students to attempt questions that might otherwise have been skipped. Despite this, the differences in the production function improve outcomes for men more than for women. The smaller decrease in skipping for women is due to the slightly lower value of the time weight for women. As a result, an increase in the time allocated for the exam does not reduce the gender gap. The estimated cutoff values are higher compared to the values in Akyol et al. (2022). This could be explained by the timing of the experiment. When we conducted this experiment, students were nine months away from the actual college entrance exam, which would explain the lower confidence of students.

[31] However, the difference in τ between the treatment and control across genders is small and not significant, using a two-sample t-test.

Table 4: Parameter Estimates

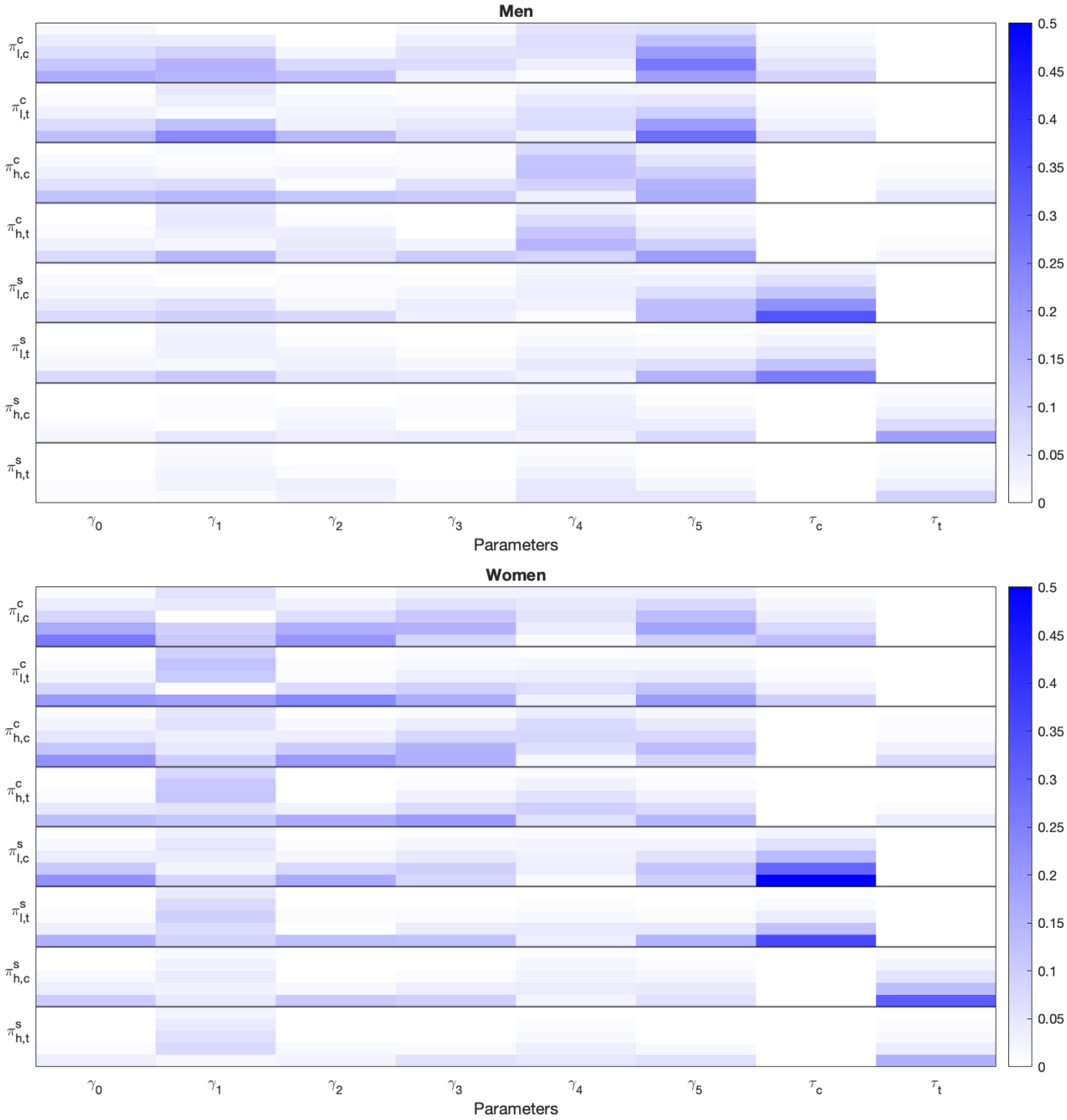
Parameter	Description	Men	Women	p_Δ
Panel A: Signal Production				
γ_0	Scaling	2.216	1.844	0.07
		(0.265)	(0.135)	
γ_1	Proficiency Weight	6.348	8.616	0.00
		(1.100)	(1.795)	
γ_2	Difficulty Weight	-10.486	-12.918	0.00
		(1.379)	(1.766)	
γ_3	Time Weight	3.631	2.613	0.02
		(0.759)	(1.028)	
γ_4	Discrepancy Factor	1.096	1.552	0.60
		(0.4915)	(1.832)	
γ_5	Time Scaling	0.603	0.827	0.49
		(0.190)	(0.352)	
Panel B: Certainty Cutoff				
$\bar{\tau}_c$	Cutoff: Control	0.335	0.345	0.61
		(0.009)	(0.016)	
$\bar{\tau}_t$	Cutoff: Treatment	0.308	0.310	0.91
		(0.016)	(0.013)	

Notes: This table shows the parameter estimates from the indirect inference procedure, explained in Section 4. Bootstrapped standard errors are computed over 1000 draws of moments after resampling data at the student-question level.

We estimated standard errors using a block-bootstrap procedure, which accounts for treatment and gender. Appendix B.2 details this bootstrap procedure. While all parameters are significant, the discrepancy factor is not. We believe the lack of significance for women's γ_4 is due to the small sample size rather than a model identification issue. To test this, we recomputed the standard errors using a simulation-based approach, which showed that when the sample size increased, the parameters became significant.[32]

[32] We explain the procedure for this simulation-based approach in more detail in Appendix B.2.1.

Figure 4: Sensitivity of Moments to Parameters



Notes: This figure presents the sensitivity of the moments used in the estimation to the parameters to be estimated. The y-axis represents moments, which includes probabilities of correct and skip responses under low/high proficiency, low/high difficulty, and control/treatment conditions. The x-axis represents the parameters in α and the certainty cutoff parameters, τ_c and τ_t.

4.3.3 Model identification and Performance Figure 4 depicts the elasticity of the moments with respect to each of the parameters. A parameter is identified if at least one moment exists that substantially affects it. Each row in the figure has five rows that correspond to each quintile of difficulty. As is evident, each parameter has at least one darker bar among the moments. This reassures us that the parameters are identified.

Next, we look at how well our model aligns with the data in some important ways. Figure 5

compares the targeted moments in the data and the model. The data and model moments always have the same patterns. The correct rate falls with difficulty while the skipped rate rises. The two curves that represent the data and the simulated data moments are close for the most part.

To perform our counterfactual experiments when allocated time changes, we need one more step, namely, to fill in the certainty cutoffs for all levels of time constraints. We have estimates of these certainty cutoffs for both men and women, but these estimates are limited to the two specific time constraint levels examined in our experiment.

Since the certainty cutoff changes with the time allowed in the exam, we need to specify how this cutoff changes for all values of time allowed, not just the two in the experiment. We need these interpolated values to conduct counterfactuals where we change the time allocated to the exam, since the estimated certainty cutoff changes with time. Our approach includes four points: two come from empirical estimates at $t = 0.8$ and $t = 1.6$, and two are theoretically posited. As t approaches 0, the function $\alpha(.)$ goes to γ_0. Since this is a small part of the value of $\alpha(.)$, we assume that all questions are skipped and that the certainty cutoff $\bar{\tau}$ is unity at $t = 0$.[33] Second, as t approaches infinity, there should be no lack of confidence, so the cutoff should asymptote to that corresponding to the individual's risk aversion. However, we do not observe risk aversion separately. We assume that agents are risk-neutral when they have infinite time to think, so the certainty cutoff is 0.2, or no risk aversion. The results of our imputations are depicted in Table 5 for men and women. Note that both cutoffs fall with the time allocated. Also, note that the certainty cutoff for women is not significantly different from that for men.

Table 5: A Sample of Extrapolated Certainty Cutoff Values

	$t = 0.5$	$t = 1$	$t = 1.5$	$t = 2$	$t = 2.5$	$t = 3$	$t = 3.5$	$t = 4$
Men, τ	0.466	0.325	0.311	0.301	0.290	0.280	0.271	0.262
Women, τ	0.477	0.331	0.312	0.301	0.290	0.279	0.269	0.259

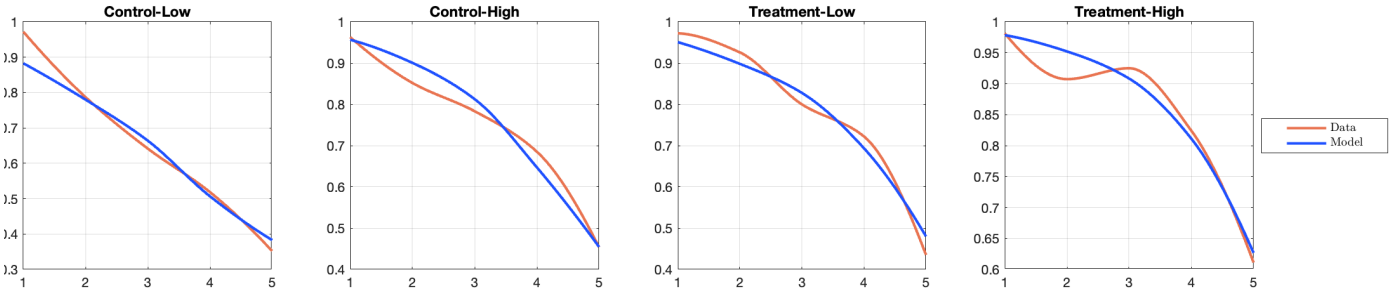
Now, we can conduct some counterfactuals to see how well the model matches the simulated data. Figure 6 shows how the fraction of correct answers, the overall scores, and the skipped questions change with additional time. We look at males and females separately. The fraction correct rises, and the fraction skipped falls as time increases, as expected, with diminishing returns to time. The fractions in the actual experiment for control ($t = 0.8$) and treatment groups ($t = 1.6$) are identified by cross markers. This clearly shows that the predictions and the data line up well.

Similarly, Figure 7 depicts the scores as time allocated increases, but now they are differentiated by ability and difficulty. As before, the data points are indicated by crosses. Note that these also match the simulations well. The gain from extra time for the low-ability group is mostly on low- and medium-difficulty questions, whereas the high-ability group benefits primarily on high-difficulty questions.

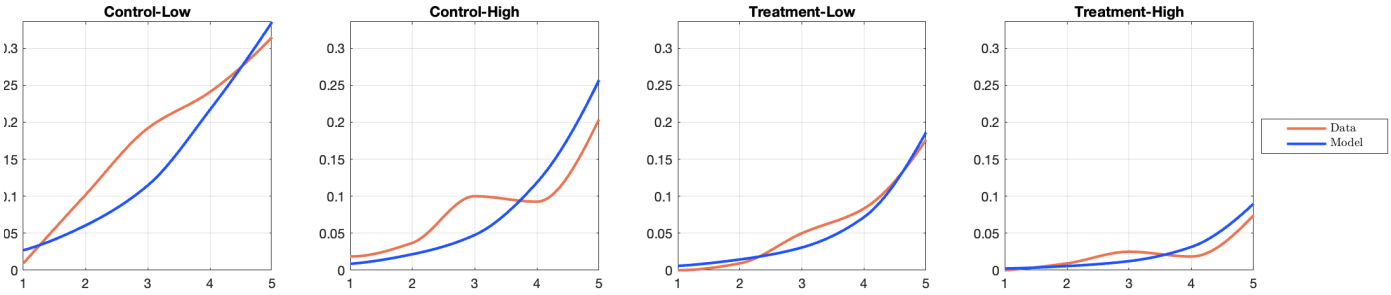
[33] We utilize the Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) to interpolate and extrapolate the certainty cutoff across different time values. We chose this because it ensures monotonicity, as suggested in our certainty cutoff estimates across treatment status. The PCHIP interpolation for a given time t is described by $\bar{\tau}(t) = a_i (t - t_i)^3 + b_i (t - t_i)^2 + c_i (t - t_i) + d_i$. The coefficients, $\{a_i, b_i, c_i, d_i\}$, are uniquely determined in each consecutive interval $[t_i, t_{i+1}]$, and they preserve the monotonicity of our empirical trend.

Figure 5: Model Fit

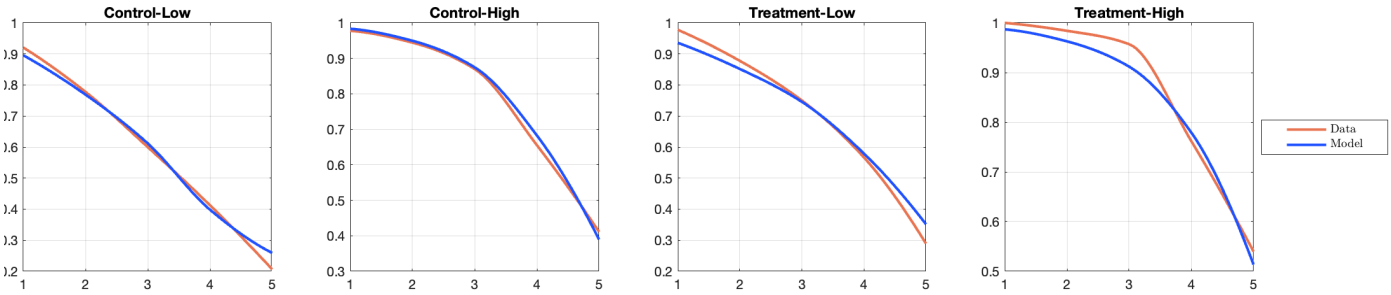
(a) Men - Correct Rate



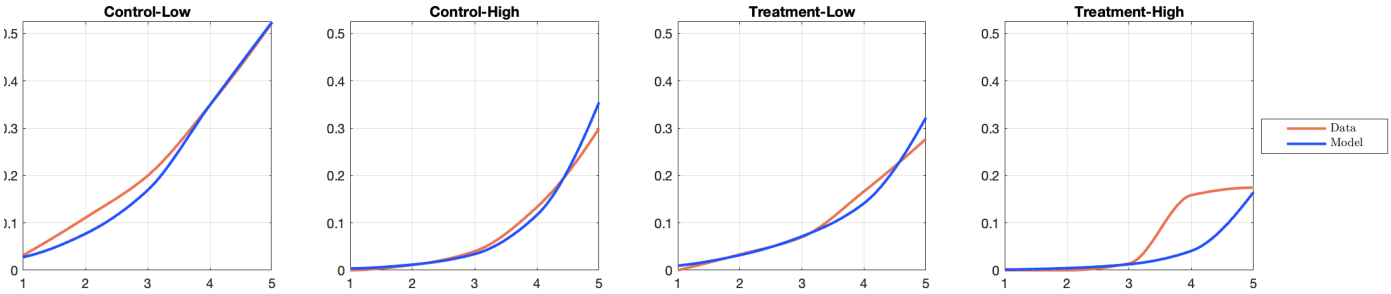
(b) Men - Skipping Rate



(c) Women - Correct Rate

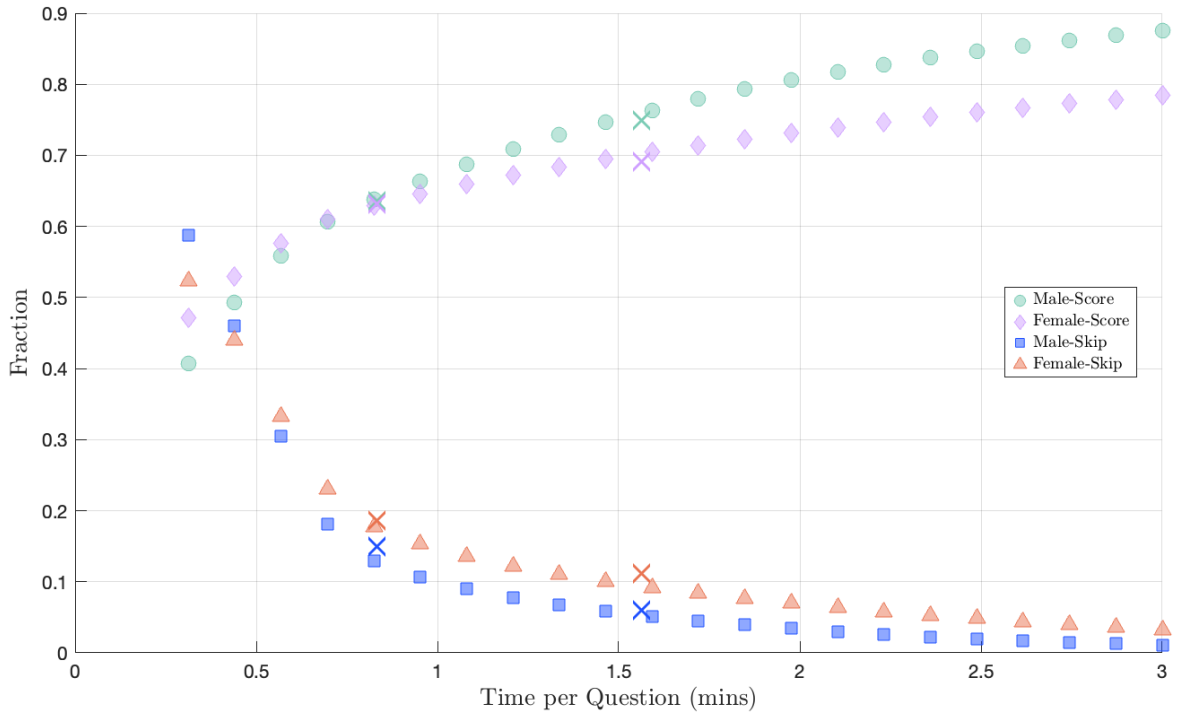


(d) Women - Skipping Rate



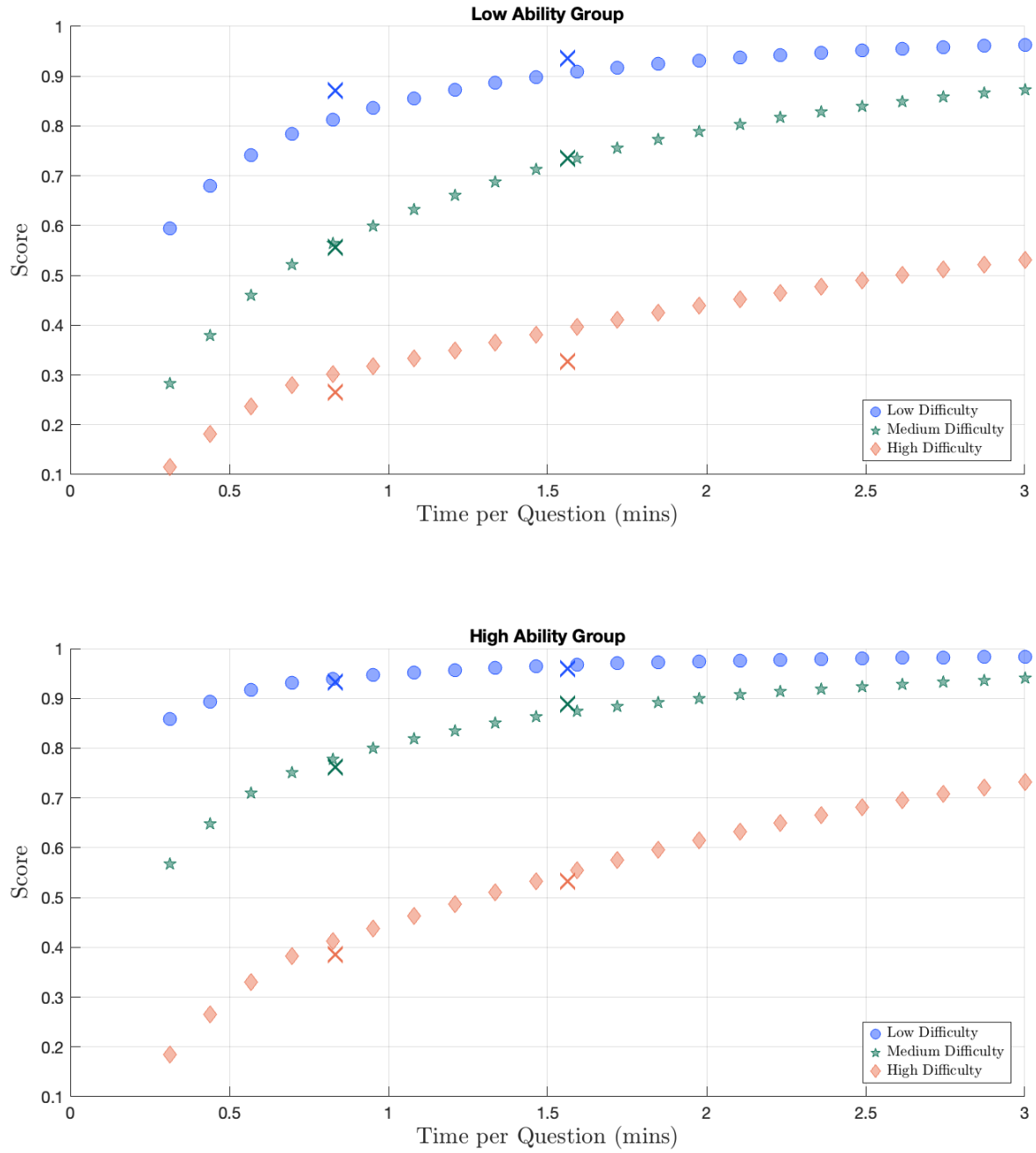
Notes: This figure illustrates the model fit by comparing key moments between the data and the model. The red lines represent the data, and the blue ones are the simulated ones. Panel (a) presents the correct response rates (on the y-axis) for each quintile of question difficulty (on the x-axis) for men, while panel (b) shows their skipping rates. Similarly, panel (c) displays the correct response rates for women, and panel (d) illustrates their skipping rates. The first column represents the model fit for the low-ability group in the control group, and the second column represents the high-ability group in the control group. The third column corresponds to the low-ability group in the treatment group, and the final column depicts the high-ability group in the treatment group. All curves are smoothed over the five points corresponding to difficulty bins.

Figure 6: Counterfactual Performance Under Alternative Testing Time – Male vs Female



Notes: This figure presents counterfactual performance under alternative test time availability for males and females. Counterfactual predictions are shown as fractions. Cross symbols at $t = 0.8$ and $t = 1.6$ refer to the actual fractions from the experiment for the relevant group.

Figure 7: Counterfactual Performance Under Alternative Testing Time – Ability-Difficulty Pairs



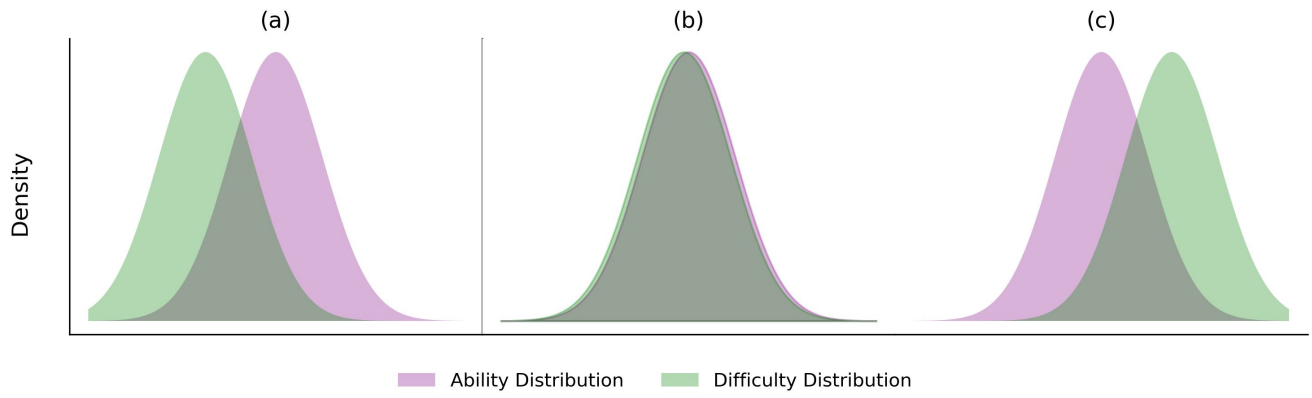
Notes: This figure presents counterfactual performance under alternative test time availability for ability-difficulty pairs. Counterfactual predictions are shown as fractions. Cross symbols at $t = 0.8$ and $t = 1.6$ refer to the actual fractions from the experiment for the relevant group.

5 Exam Design and Sorting

A key objective of exams is to sort the examinees. This section looks at how exams of differing difficulty will do in sorting student populations of differing abilities. We will use the rank correlation of the two as our measure of sorting, since sorting refers to ranking students by their ability, not the correlation between ability and score. Of course, exams differ in their format depending on who the exam is designed to sort between. An exam for choosing mathematically talented youth, like the American Mathematics Competition (AMC) administered to students in junior and senior high school in the US, by its nature, wants to sort among those at the top. As a result, the questions are very difficult, and the median score is zero. On the other hand, a college entrance exam like the SATs in the US needs to sort students with a much wider range of abilities and so has questions with a range of difficulty. However, it does poorly in sorting students at the top of the ability distribution, evidenced by a mass point at the perfect score.[34]

Using the estimated model, we first see how changing the number of questions in the exam affects the sorting. Further, how the question difficulty affects sorting at different numbers of questions under high and low-pressure environments. This helps provide some intuition for our next counterfactuals, where we examine in more detail how time pressure and the extent of negative marking affect the exam's capacity to sort students according to their ability.

Figure 8: Illustrative Diagram

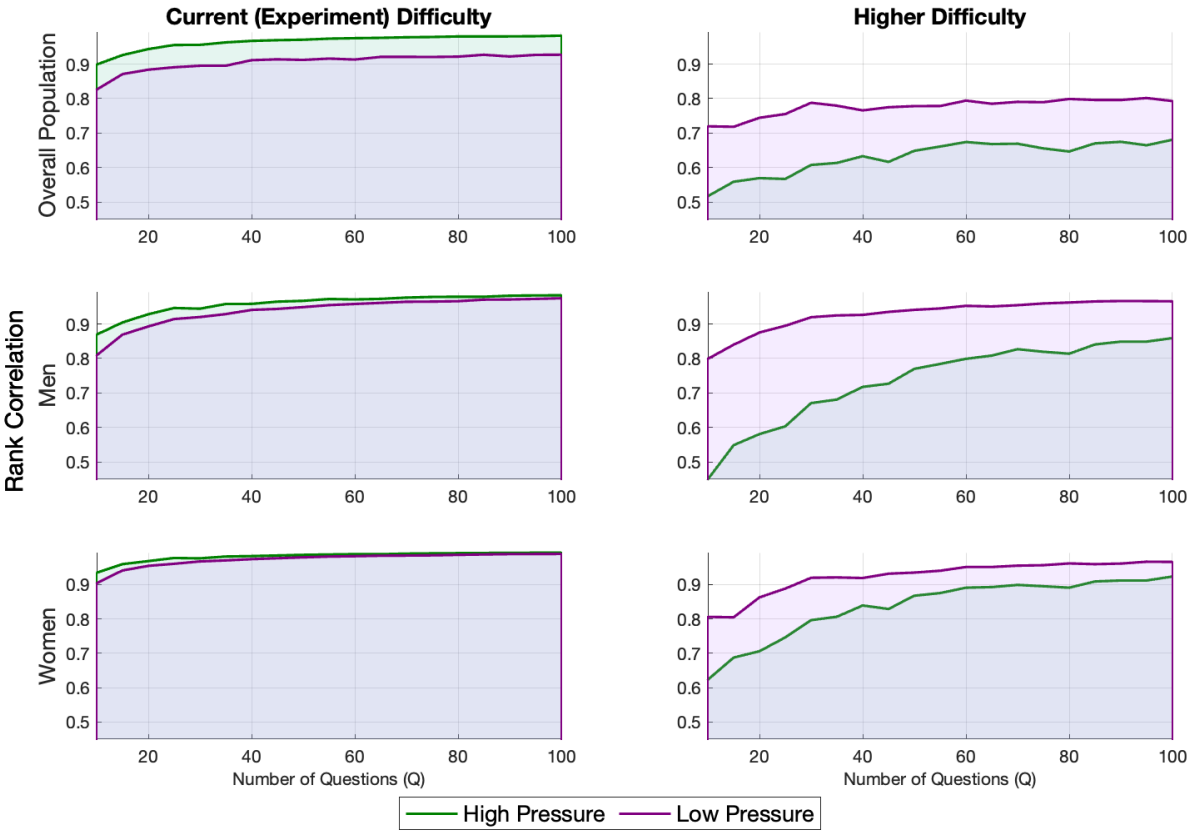


An illustrative figure is useful for understanding the mechanisms driving our results. In Figure 8, we present how the alignment between the ability distribution and the difficulty distribution affects the effectiveness of sorting in an exam. The x-axis can be thought of as the ability of the student and the difficulty of the questions. Suppose that the difficulty of the questions is in the interval marked in the graph. Then, if the ability of the students is roughly at the same interval (as drawn), the exam will be able to sort well. If the difficulty of the exam falls, sorting worsens, as the exam is able to sort only the students at the lower end of the ability distribution, while higher-ability students will all perform well, making it harder to distinguish among them. If the difficulty increases, lower-ability students will not be well sorted, as they will all perform poorly, making it harder to distinguish between them. The same arguments apply if ability distributions change:

[34] Not all exams wish to sort students by ability. Certification exams, for example, aim to ensure that a target level of achievement is attained without focusing on sorting students. These exams target the material covered and little else.

only when the ability distribution and the question difficulty are aligned will overall sorting be good. Since increasing time pressure is analogous to question difficulty rising or student ability falling, the same logic applies there.

Figure 9: Ability Score Rank Correlation - Number of Questions



Notes: This figure presents the relationship between exam design and sorting outcomes. The top row shows it for the overall population, the second for men, and the bottom for women. The left figure displays results for the difficulty distribution of the experiment, and the right figure for a higher difficulty exam.

5.1 Number of Questions

We begin by exploring whether the number of questions has any impact on the sorting quality. Figure 9 displays the ability-score rank correlation levels as a function of the number of questions across two different exam difficulties and under low and high time pressure. First, as expected, there is an increasing and concave relationship between the number of questions and the rank correlation, consistent with diminishing returns to the length of the exam. Second, the rank correlation (sorting) is higher for high-pressure environments than for low-pressure environments when the exam is easy, but this reverses when the exam is difficult. Moreover, sorting is worse when the exam is difficult. This is exactly what one would expect if the easy exam was too easy and the difficult one was too difficult to sort well. In the former case, a high-pressure environment would act like it was increasing the difficulty of the exam and would better align the difficulty of the exam with student ability. In the latter case, the difficult exam is made even more difficult by

increased time pressure, which further impairs its ability to sort effectively. Thus, time pressure improves the sorting with an easy exam and worsens it with a hard one.

The results for men and women follow the same general patterns but with some nuances. For both men and women, the rank correlation increases with the number of questions, though sorting quality is consistently lower under high-pressure conditions when the exam is difficult. However, the gap between high- and low-pressure environments appears more pronounced for men, suggesting that time matters more in higher-difficulty exams and more so for men.

5.2 Baseline Regime, Different Time Pressure

We consider three different time allotments per question: $t = 0.9$, $t = 1.7$, and $t = 3$.[35] We refer to them as high, medium, and low time-pressure environments. We take the empirical distribution of ability in our sample, depicted in Figure A1 in the Appendix. We create ten ability groups, with the first group formed by shifting the empirical distribution left by -0.5 and the tenth group shifted right by 0.5. Group 5 is the distribution in the data, i.e., the original one. An analogous procedure is followed for difficulty.[36] This gives us the ten proficiency and ten ability levels we see on each of the nine sub-figures' x and y axes. Note that ability and difficulty rise from groups 1 to 10. In Figure 10, we present the correlation between exam score ranks and ability ranks across various combinations of student proficiency levels and exam difficulty distributions under high, medium, and low time pressure using heatmaps. Darker colors represent higher rank correlations.

As discussed in Section 5, sorting is most effective when student ability aligns with question difficulty (as in Figure 8 (b)). If student ability is higher than question difficulty (as in Figure 8 (a)), only students at the lower end of the ability distribution will be well sorted. Higher-ability students will find the questions easy, and the exam will not sort them well. Similarly, if student ability is too low relative to question difficulty, only students at the top of the ability distribution will be well sorted, resulting in poor overall sorting. Time pressure, being analogous to increased difficulty, has the same effects.

Therefore, we would expect that as the ability increases or the time pressure falls, the questions must become more challenging to sort the students well. This is most evident in the low-pressure environment with all students, as shown in the top-right sub-figure, where sorting is high within an upward-sloping dark band. If the students have low ability, i.e., population proficiency, which is given on the horizontal axis, is low, a difficult exam (difficulty is on the vertical axis) will do little to sort them, as they are likely to answer most of the questions incorrectly. Similarly, if students have high ability, an easy exam will do little to sort them out, as they will all perform well. Only when student ability and question difficulty are close to each other does the exam sort well. A similar pattern is visible in the sub-figure just below, which also represents a low-pressure environment but focuses only on males. The pattern is less obvious elsewhere as the light region at the upper left corner has not been reached. If the questions had been even more difficult, sorting would have deteriorated further, as a greater number of students—who are, on average, of lower

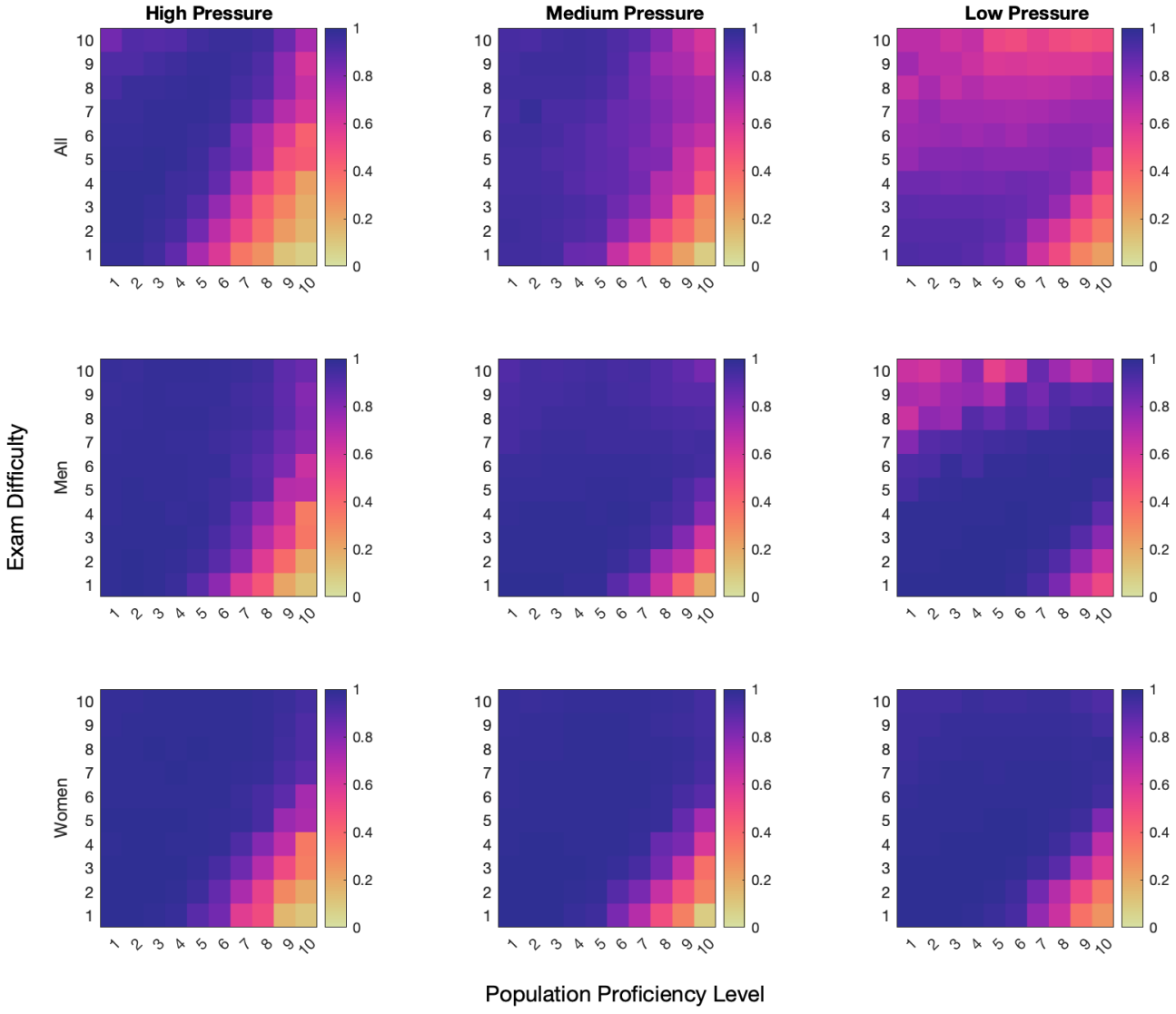
[35] The first two reflect the levels in our experiment.

[36] The first group is created by shifting -0.2 points to the left and the tenth group by 0.7 points to the right. This is because our actual difficulty distribution is inclined to the left, meaning it leans toward easier questions.

ability—would have answered them incorrectly.

What happens to sorting, i.e., how does the band in the extreme top right-hand corner move, as time pressure increases? Consider keeping the exam difficulty the same and increasing the time pressure. This is equivalent to lowering ability. In other words, it would move the band to the left. Analogously, given student ability, increasing time pressure is equivalent to making the exam more difficult, which results in the band moving up. Hence, as we move from the right to the left, the band moves up and to the left, as is evident.

Figure 10: Ability - Score Rank Correlations



Notes: This figure presents the relationship between exam design and sorting outcomes. The top figure shows it for all samples, the middle for men, and the bottom for women. The darker the color is, the higher the rank correlation is. Each sub-map presents a different time-constrained environment. The lower the pressure, the higher the test time available.

Next, consider how the ability to sort changes across rows, i.e., when comparing all students to gender-specific sorting. Note that the dark region, the sorting band (upward-sloping dark re-

Table 6: Counterfactual Negative Marking Regimes

	penalty = 0		penalty = 0.25		penalty = 0.5	
	Control	Treatment	Control	Treatment	Control	Treatment
Panel A: Men						
Correct	0.718	0.809	0.687	0.797	0.587	0.732
Skip	0.000	0.000	0.127	0.053	0.384	0.231
Score	0.648	0.762	0.640	0.760	0.580	0.723
τ	0.200	0.200	0.333	0.310	0.540	0.506
Panel B: Women						
Correct	0.724	0.775	0.682	0.755	0.593	0.681
Skip	0.000	0.000	0.166	0.087	0.388	0.287
Score	0.655	0.719	0.645	0.715	0.588	0.673
τ	0.200	0.200	0.342	0.311	0.553	0.508

Notes: This table presents the performance across genders (fraction correct, skipped, and overall score) under three negative marking regimes. The first column shows results for no penalty cases, the second for a 0.25 penalty, and the last for a 0.5 penalty.

gion), is larger and darker when students are sorted within gender. This is intuitive, as the signal production functions differ by gender, with time being more important for men. Hence, sorting the entire population will be harder than sorting within gender, as there is an extra dimension of heterogeneity in the former. In other words, ranking students within gender raises the correlation between ability and score substantially.[37] This suggests that affirmative action, by reserving a fraction of seats for women while allocating the remainder to men, might have the unexpected positive side effect of improving the sorting.

These counterfactuals also shed light on the attractiveness of adaptive testing. A traditional exam that includes the same questions for all examinees has a hard time sorting between students, especially if the student quality is very variable. An exam designed to distinguish top-performing students may fail to do so for those at the lower end of the distribution, and vice versa. In contrast, adaptive tests dynamically tailor question difficulty based on the examinee's performance, allowing for more precise sorting at all ability levels.

5.3 Alternative Negative Marking Regimes

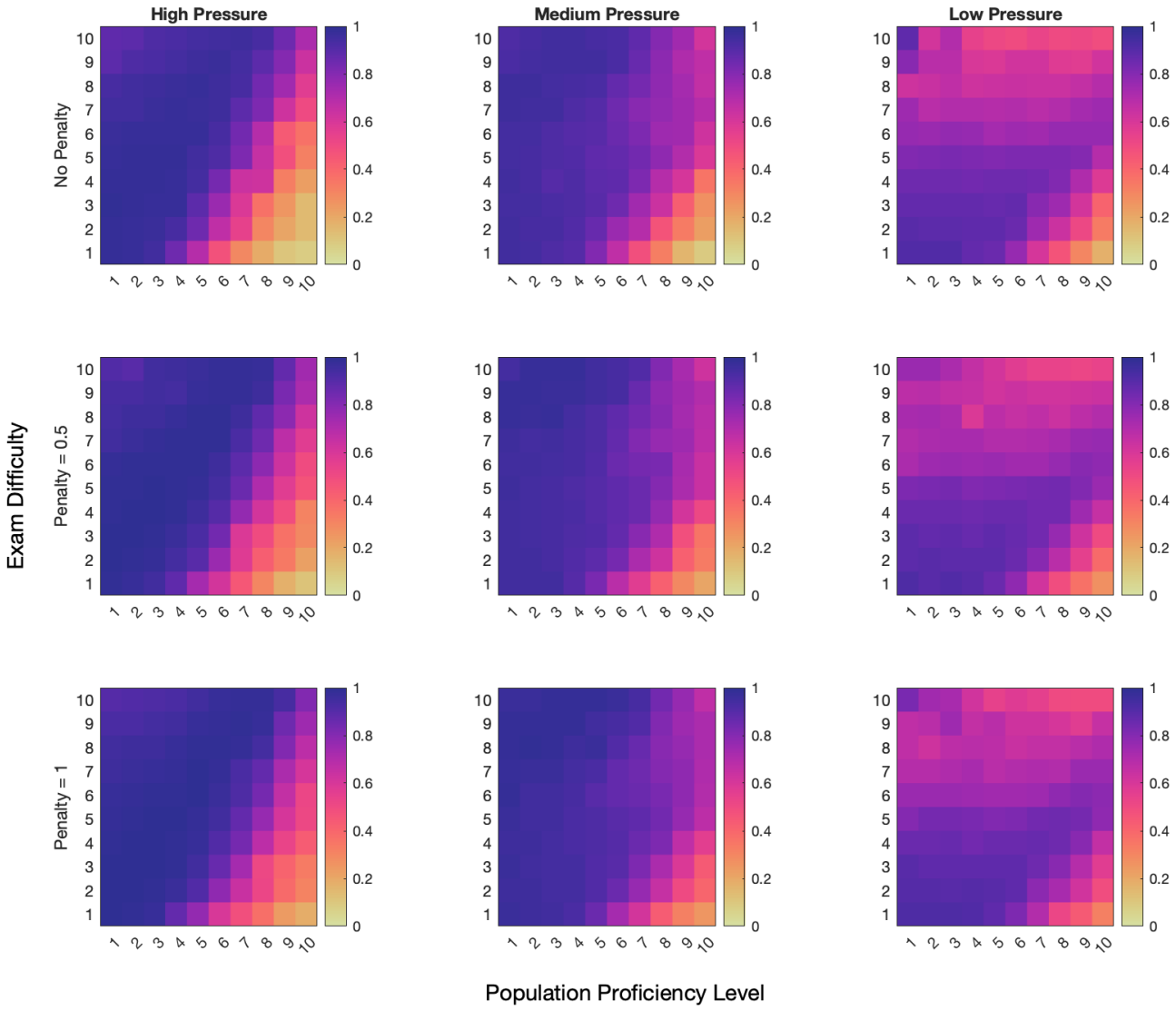
In this section, we examine the effects of different negative marking regimes using simulations. Table 6 reports the results of these simulations for the fraction of correct and skipped answers as well as the overall score under different negative marking regimes and treatment conditions. We specify the utility function to be a CARA utility function, namely, U(s)= $1 - e^{-\rho s}$. We back out the ρ that corresponds to the estimated cutoff level using Equation (5). With ρ obtained, we find the certainty cutoff under alternative penalty regimes.[38] Table 6 shows outcomes with no

[37] The rank correlation would increase mechanically as the sample size decreases. We compute the rank correlation for a random sample of the same size as our men's and women's samples, demonstrating that the patterns observed for men and women are not simply a result of smaller sample sizes.

[38] For example, take the female cutoff of 0.31. We use Equation (5) to find ρ in $1 - e^{-\rho y}$. Then, we find the cutoff for a 0.5 penalty, which is 0.506.

penalty, a 0.25 penalty, and a 0.5 penalty across control and treatment groups. We observe that as the penalty for incorrect answers increases, students get fewer questions correct, their scores go down, and they skip more questions. Since the cutoff for men is lower than for women, and they have a higher time weight (so that they gain more from increases in time), men gain more from having more time. They also lose slightly less as the penalty for a wrong answer increases.

Figure 11: Ability Score Rank Correlation - Penalty



Notes: This figure presents the relationship between exam design and sorting outcomes. The top figure shows it for no penalty exam, the middle for a .5 penalty, and the bottom for a penalty of 1. The darker the color is, the higher the rank correlation is. Each sub-map presents a different time-constraint environment. The lower the pressure, the higher the test time available.

Next, we examine how penalties for the wrong answer (negative marking) affect sorting and how this changes with time pressure. This is presented in Figure 11 for this relationship for three penalty scenarios: no penalty, a penalty of 0.5, and a penalty of 1. We focus on how patterns

change as the penalty for a wrong answer increases. Overall, a higher penalty reduces guessing. Since guessing tends to weaken sorting, this reduction in guessing leads to improved sorting, visually represented by an expansion of the darker region as the penalty rises. Additionally, sorting improves when time pressure increases, making this effect particularly pronounced under high time pressure.

6 Conclusion

This paper develops an empirical framework to analyze the relationship between time constraints and performance in multiple-choice exams. Using data from a field experiment motivated by a natural experiment, we adopt a structural approach to estimate the impact of time pressure on exam outcomes. Our approach exploits the insight that risk aversion—a primitive—remains constant with or without time pressure, whereas confidence may shift as time per question changes to estimate the increase in confidence coming from less time pressure. We show that additional time benefits test-takers by improving the precision of their signals, but these gains are not uniform across individuals.

Our findings reveal that the relationship between time constraints and performance is shaped by the interaction of student ability, question difficulty, and gender. The greatest improvements occur when question difficulty is well-matched to a student's ability, while very easy or very difficult questions show relatively smaller gains. Although both men and women perform better with additional time, the structural estimates suggest that men benefit more, primarily due to differences in how signal quality responds to time availability rather than due to differences in confidence or risk preferences.

In addition, we show that a pervasive belief in the education literature, namely that women are more risk averse/less confident than men, is not evident in our estimation. What is evident is that the production functions for signal quality differ significantly by gender. Once this is allowed, we find no significant differences in risk aversion by gender. Finally, our counterfactuals highlight the role of exam design in terms of its potential to sort students. They show that the ability of an exam to sort is non-monotonic in its difficulty and time pressure. In addition, negative marking improves sorting, especially in high-pressure environments. In this way, our study adds to the growing literature in education that combines structural models with experimental data to inform the design of better education policies.

We see this paper as a first step towards using models and structural estimation to help design better exams in various contexts. There is at least one obvious place for improvement in our paper. We assume that the students spend an equal amount of time on each question. Allowing time spent to be endogenous would result in students spending time on a question until the marginal benefit of time on the question equals the opportunity cost, namely the marginal benefit of spending that time on other questions. Increasing the overall time allowed would result in a greater increase in time spent on questions where the marginal benefit slowly falls. Such questions are likely to be those that are within the reach of the student. Allowing for this extension would enrich the predictions of the model. We do not explore this angle as we could not collect data on the

time spent on each question, which is necessary for such analysis, as our test was paper-based. We expect that endogenizing the time spent on each question will not qualitatively affect our results, though it would enrich the setting.

There are many potential extensions of our approach. It would be interesting to extend the model to low-stakes exams, where students often do not take the exam seriously. It would be interesting to try and identify students who have a high cost of effort and hence do not take the exam seriously and study how to improve this aspect.

Our approach also shows potential for application in other contexts of interest. For instance, the process of diagnosing a patient's condition can be viewed as doctors receiving signals that help them identify the correct condition from a set of possibilities. Factors such as the expertise of a doctor, the complexity of the diagnosis, time pressure, and potential biases (e.g., racial or gender biases) can be integrated into a model like ours. Such a model could be applied to the data and would help us to better understand how doctors make decisions under various constraints. It would shed light on the cost (in terms of obtaining the correct diagnosis) from the increasing time pressure being put on doctors by profit-maximizing healthcare systems. This is in contrast to the usual approach taken. See, for example, Philip and Ozkaya (2024).

Our approach might also be fruitful in studying insurance markets. Insurance firms classify potential buyers into different risk groups and adjust the price of the product based on these groups. Their ability to sort potential customers this way is analogous to that of a student's ability to find the correct answer. This approach differs from the standard industrial organization approach. See, for example, Cosconati et al. (2024).

References

Akyol, P., Key, J., & Krishna, K. (2022). Hit or Miss? Test Taking Behavior in Multiple Choice Exams. *Annals of Economics and Statistics*, (147), 3–50.

Akyol, P., Krishna, K., & Wang, J. (2021). Taking PISA Seriously: How Accurate are Low-Stakes Exams? *Journal of Labor Research*, *42*(2), 184–243.

Arden, R., & Plomin, R. (2006). Sex differences in variance of intelligence across childhood. *Personality and Individual Differences*, *41*(1), 39–48.

Arenas, A., & Calsamiglia, C. (2022). Gender Differences in High-Stakes Performance and College Admission Policies.

Baldiga, K. (2014). Gender Differences in Willingness to Guess. *Management Science*, *60*(2), 434–448.

Ben Zur, H., & Breznitz, S. J. (1981). The effect of time pressure on risky choice behavior. *Acta Psychologica*, *47*(2), 89–104.

Bridges, K. R. (1985). Test-Completion Speed: Its Relationship to Performance on Three Course-Based Objective Examinations. *Educational and Psychological Measurement*, *45*(1), 29–35.

Brown, C., Kaur, S., Kingdon, G., & Schofield, H. (2024). Cognitive endurance as human capital. *The Quarterly Journal of Economics*, qjae043.

Cai, X., Lu, Y., Pan, J., & Zhong, S. (2019). Gender Gap under Pressure: Evidence from China's National College Entrance Examination. *The Review of Economics and Statistics*, *101*(2), 249–263.

Chabris, C. F., Laibson, D., Morris, C. L., Schuldt, J. P., & Taubinsky, D. (2009). The allocation of time in decision-making. *Journal of the European Economic Association*, *7*(2), 628–637.

Cosconati, M., Wu, F., & Jin, Y. (2024). Competing under information heterogeneity: Evidence from auto insurance.

Croson, R., & Gneezy, U. (2009). Gender Differences in Preferences. *Journal of Economic Literature*, *47*(2), 448–474.

Diederich, A., Wyszynski, M., & Traub, S. (2020). Need, frames, and time constraints in risky decision-making. *Theory and Decision*, *89*(1), 1–37.

Ellison, G., & Swanson, A. (2021). Dynamics of the Gender Gap in High Math Achievement. *Journal of Human Resources*.

Feinberg, R. M. (2004). Does more time improve test scores in micro principles? *Applied Economics Letters*, *11*(14), 865–867.

Franco, C., & Gomez-Ruiz, M. (2023). *Bridging the Gender Gap in Access to STEM through In-Exam Stress Management*. Google Docs. Retrieved January 19, 2024, from https://drive.google.com/file/u/0/d/1exQpsAfrdWMk9C6BjwYGiS526mxWHzI8/view?pli=1&usp=embed_facebook

Fudenberg, D., Strack, P., & Strzalecki, T. (2018). Speed, Accuracy, and the Optimal Timing of Choices. *American Economic Review*, *108*(12), 3651–3684.

Galasso, V., & Profeta, P. (2024). Gender differences in math tests: The role of time pressure. *The Economic Journal*.

Goldhammer, F. (2015). Measuring Ability, Speed, or Both? Challenges, Psychometric Solutions, and What Can Be Gained From Experimental Control. *Measurement*, *13*(3-4), 133–164.

Hakimov, R., Schmacker, R., & Terrier, C. (2023). Confidence and College Applications: Evidence from a Randomized Intervention. *Rationality and Competition Discussion Paper Series*, (377).

Halpern, D. F., & Wai, J. (2019). Sex differences in intelligence. In R. J. Sternberg (Ed.), *The cambridge handbook of intelligence* (pp. 317–345). Cambridge University Press.

Hausfeld, J., & Resnjanskij, S. (2018). Risky Decisions and the Opportunity Cost of Time. *ifo Working Papers*, (269).

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

Jacob, B., & Rothstein, J. (2016). The Measurement of Student Ability in Modern Assessment Systems. *Journal of Economic Perspectives*, *30*(3), 85–108.

Kocher, M. G., Pahlke, J., & Trautmann, S. T. (2013). Tempus Fugit: Time Pressure in Risky Decisions. *Management Science*, *59*(10), 2380–2391.

Kocher, M. G., & Sutter, M. (2006). Time is money—Time pressure, incentives, and the quality of decision-making. *Journal of Economic Behavior & Organization*, *61*(3), 375–392.

Liang, A. (2023, May 18). *Information and Learning in Economic Theory*. arXiv: 2212.07521 [econ].

Lu, Y., & Sireci, S. G. (2007). Validity Issues in Test Speededness. *Educational Measurement*, *26*(4), 29–37.

Montolio, D., & Taberner, P. A. (2021). Gender differences under test pressure and their impact on academic performance: A quasi-experimental design. *Journal of Economic Behavior & Organization*, *191*, 1065–1090.

Niederle, M., & Vesterlund, L. (2010). Explaining the Gender Gap in Math Test Scores: The Role of Competition. *Journal of Economic Perspectives*, *24*(2), 129–144.

Nursimulu, A. D., & Bossaerts, P. (2014). Risk and Reward Preferences under Time Pressure. *Review of Finance*, *18*(3), 999–1022.

Onwuegbuzie, A. J., & Seaman, M. A. (1995). The Effect of Time Constraints and Statistics Test Anxiety on Test Performance in a Statistics Course. *The Journal of Experimental Education*, *63*(2), 115–124.

Ors, E., Palomino, F., & Peyrache, E. (2013). Performance Gender Gap: Does Competition Matter? *Journal of Labor Economics*, *31*(3), 443–499.

Pekkarinen, T. (2015). Gender differences in behaviour under competitive pressure: Evidence on omission patterns in university entrance examinations. *Journal of Economic Behavior & Organization*, *115*, 94–110.

Philip, M., & Ozkaya, Ö. (2024). Disparate treatment and outcomes in emergency departments: Evidence from florida [Accessed: 2024-12-21]. https://minu-philip.github.io/Files/PhilipMinu_JMPDraft.pdf

Ramos-Loyo, J., González-Garrido, A. A., Llamas-Alonso, L. A., & Sequeira, H. (2022). Sex differences in cognitive processing: An integrative review of electrophysiological findings. *Biological Psychology*, *172*, 108370.

Sunde, U., Zegners, D., & Strittmatter, A. (2022). *Speed, Quality, and the Optimal Timing of Complex Decisions: Field Evidence*. arXiv: 2201.10808.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456–477.

Wilcox, N. T. (1993). Lottery Choice: Incentives, Complexity and Decision Time. *The Economic Journal*, *103*(421), 1397–1417.

Wild, C. L., Durso, R., & Rubin, D. (1982). Effect of Increased Test-Taking Time on Test Scores By Ethnic Group, Years Out of School, and Sex. *Journal of Educational Measurement*, *19*(1), 19–28.

Appendices

A Additional Tables and Figures

A.1 Balance on Covariates

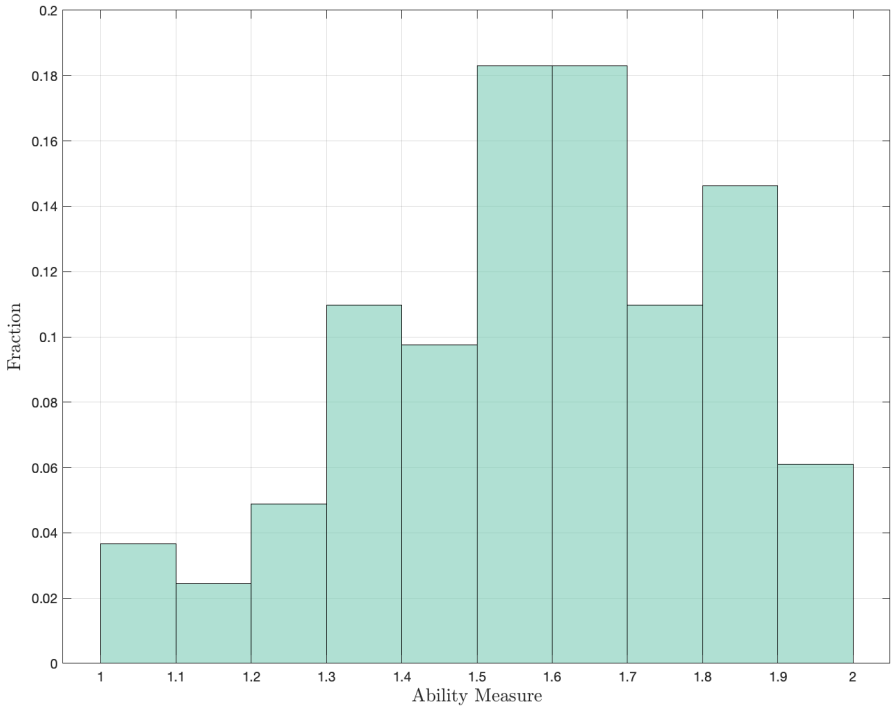
Table A1: Balance in Observable Characteristics Across Groups

	Time-constrained	Time-relaxed	Difference	p-value
	(1)	(2)	(3)	(4)
Mother Education	12.24	12.98	0.74	0.20
Father Education	14.43	14.70	0.27	0.46
Fam. Income Tercile	1.74	1.88	0.14	0.52
Tutoring	0.93	0.93	-0.00	0.98
11th Grade GPA	94.16	93.93	-0.22	0.36
Practice Exam Scores	0.03	-0.03	-0.06	0.59
Observations	42	41		
Joint F-Test				0.99

Notes: The data was collected through a survey prior to the experiment, and the statistical tests were performed using the Mann-Whitney U Test (also known as the Wilcoxon rank-sum test). Education variables capture the number of years of schooling completed. Family income is represented in terciles, dividing the sample into low, middle, and high-income groups based on the distribution of combined parental income. The 'tutoring' variable denotes the proportion of individuals attending a tutoring center. Additionally, the variable 'practice exam scores' refers to the average scores from weekly practice exams at this school, which have been normalized for consistency.

A.2 Measurement

Figure A1: Ability Measure Distribution



B Estimation

B.1 Detailed Estimation Procedure

1. For given parameter values of $\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5$, we calculate α_t given the (θ, d, t) tuple.

2. For a given cutoff value

$$\bar{\tau} := \frac{U(0) - U\left(-\frac{s}{C-1}\right)}{U(s) - U\left(-\frac{s}{C-1}\right)},$$

which represents the threshold required to justify answering a question, we find the probabilities of each outcome—correct, incorrect, or skip—through the following steps:

(i) **Simulate Signals for Incorrect Answers:**

- Simulate 20 draws for the incorrect answer that can be chosen with the second-highest probability. To ensure consistency in identifying the second-highest probability, define the order statistic for the highest probability. A grid of 20 values, uniformly distributed over the interval $[0, 1]$, represents these draws.
- Simulate five draws, each for the third- and fourth-highest probabilities. Since the second-highest probability has the most significant impact on decision-making, it receives more simulation draws. Reducing the number of draws for the third- and fourth-highest probabilities helps ease computational complexity.
- Pair each of the 20 draws for the second-highest probability with $5 \times 5 \times 5 = 125$ unique combinations of draws for the third- and fourth-highest probabilities, resulting in $20 \times 125 = 2,500$ combinations of incorrect answer signals.

(ii) **Simulate Signals for the Correct Answer:** For each of the $2,500$ combinations of incorrect answer signals, simulate the signal for the correct answer.

(iii) **Find Critical Signal Values for the Correct Answer:**

- Compute the minimum signal value for the correct answer required to justify attempting the question, i.e., when $\pi_t^x > \bar{\tau}$.
- Compute the maximum signal value for the correct answer required to justify selecting the incorrect answer with the second-highest probability.

(iv) **Calculate Outcome Probabilities:**

- Compute the probability of answering correctly, conditional on the simulated signals for the four incorrect answers.
- Determine the skipping probability as the probability that the correct answer signal is below its minimum threshold.
- Compute the probability of answering incorrectly as the residual:

$$P(\text{Incorrect}) = 1 - P(\text{Correct}) - P(\text{Skip}).$$

3. Using the probabilities of answering correctly and skipping, define the following loss function:

$$\mathcal{L}(\mathbf{\Theta}) = \left(\mathbf{m}^{\text{model}}(\mathbf{\Theta}) - \mathbf{m}^{\text{data}}\right)' W \left(\mathbf{m}^{\text{model}}(\mathbf{\Theta}) - \mathbf{m}^{\text{data}}\right), \tag{8}$$

where W is a weighting matrix.

4. Evaluate the loss function $\mathcal{L}(\mathbf{\Theta})$ on a grid of $20,000$ points for the parameters $\mathbf{\Theta}$. Use a Halton sequence to generate the grid points to ensure a well-distributed, low-discrepancy sequence over the parameter space.

5. Narrow the search to a finer grid around the points where the loss function is minimized. Select the best 10 points from the grid search as starting points for local optimization.

6. Use the interior-point method for local optimization. The parameter set with the best fit from this procedure is taken as the final estimate.

B.2 Standard Errors

In our student decision-making model, standard errors are estimated through a nonparametric bootstrap procedure. Specifically, the block-bootstrap procedure is designed to align with the randomized sampling process described in Section 2.1, balancing treatment status and gender. In addition, we balanced the proficiency level in our resampling to better represent different ability groups. The procedure is explained as follows:

- We start by grouping all experimental subjects into treatment-gender-proficiency bins, resulting in a total of 8 bins. Each bin contains $N_1, N_2, ..., N_8$ subjects, respectively.
- For each bin, subjects are randomly selected with replacement to generate bootstrap samples of sizes $N_1, N_2, ..., N_8$. This process is repeated for each bin until full samples of N_j are obtained.
- For each bootstrap sample, we calculate the corresponding data moments.
- Using these bootstrapped data moments, we re-estimate the model 1,000 times. The initial values for the estimation are set to the estimates presented in Table 4. The bootstrap estimation process is conducted in parallel on Penn State's Roar Collab cluster environment.
- Standard errors are computed as the standard deviation of the parameter estimates obtained from the 1,000 bootstrap iterations. In this calculation, we applied a log transformation for the parameters' distributions that do not follow a normal distribution.

B.2.1 Monte Carlo Based Bootstrap. To evaluate the potential low power of our discrepancy estimates due to the limited data size, we implemented a Monte Carlo-based bootstrap procedure, focusing on the female subset of our data, as the parameter of interest was not significant for this group. Specifically, we sampled the abilities of 1,000 students from the observed distribution of women's abilities. Following the procedure described in Section B.2, we resampled the simulated

data 300 times (to enhance computational efficiency) and conducted the indirect estimation procedure. From these 300 bootstrap samples, we calculated the bootstrapped standard errors. Our findings indicate that all parameters are statistically significant ($p = 0.05$). The mean and standard error of the parameters are provided in Table A2.

Table A2: Monte Carlo Based Bootstrap

Parameter	Description	Estimate	Std. Error
Panel A: Signal Production			
γ_0	Scaling	1.844	0.29
γ_1	Proficiency Weight	8.616	0.27
γ_2	Difficulty Weight	–12.918	0.26
γ_3	Time Weight	2.613	0.31
γ_4	Discrepancy Factor	1.552	0.79
γ_5	Time Scaling	0.827	0.07
Panel B: Certainty Cutoff			
$\bar{\tau}_c$	Cutoff: Control	0.345	0.003
$\bar{\tau}_t$	Cutoff: Treatment	0.310	0.001

Notes: This table shows the parameter estimates from the indirect inference procedure, explained in Section 4. Simulation-based bootstrapped standard errors were computed over 300 draws of moments after resampling the simulated data at the student level. See Appendix B.2 for details.

C Policy Simulations

To conduct our analysis on student sorting, we begin by simulating a student population divided into ten types. The ability distribution is built based on the mean and standard deviation derived from our experimental data. The simulated student population consists of 1,000 students, evenly split between males and females. For each gender, we adjust the ability mean by moving up to ± 0.5 points around the original while keeping the standard deviation constant. We obtain ten distinct ability means. Using the mean and standard deviation, we draw 500 student abilities for each gender from a normal distribution.

A similar approach is applied to question difficulty. We create ten difficulty means within the range of 1 to 2 and use these means, along with the standard deviation of question difficulties from our data, to draw 100 questions per exam difficulty set. These exam difficulty draws remain constant across genders.

Under each time regime, the cutoff is adjusted by using PCHIP interpolation separately for each gender. For every combination of student population, exam, and time regime, we calculate students' scores and compute the correlation with their abilities. Each cell in the resulting heatmap displays the correlation between student abilities and their scores.

Online Appendix for "Do Time Constraints Matter? How, Why, and for Whom?"

A Signal Informativeness

In this section, we analyze the information content of two matrices, X and Y, which represent the belief structures of students answering a multiple-choice question with five options to choose from, only one being correct. The matrix Y captures prior beliefs before any signal is received, while X represents updated beliefs after receiving a signal.

Each matrix is structured such that columns correspond to the correct answer, and rows represent the students' subjective beliefs about which option is correct, conditional on that correct answer. We aim to show that the signal structure represented by X is more informative than Y under Blackwell ordering. Furthermore, we show that as π^c increases, X becomes strictly more informative.

The belief matrix Y, representing initial beliefs before any signal is observed, is given by:

$$Y = \frac{1}{5}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}. \tag{9}$$

Each entry Y_{ij} represents the probability that students believe that the answer i is correct, conditional on the answer j being the true correct answer. Since no signal has been received, the beliefs are uniform.

After receiving a signal, students update their beliefs, leading to the matrix X:

$$X = \begin{bmatrix} \pi^c & \frac{1-\pi^c}{4} & \frac{1-\pi^c}{4} & \frac{1-\pi^c}{4} & \frac{1-\pi^c}{4} \\ \frac{1-\pi^c}{4} & \pi^c & \frac{1-\pi^c}{4} & \frac{1-\pi^c}{4} & \frac{1-\pi^c}{4} \\ \frac{1-\pi^c}{4} & \frac{1-\pi^c}{4} & \pi^c & \frac{1-\pi^c}{4} & \frac{1-\pi^c}{4} \\ \frac{1-\pi^c}{4} & \frac{1-\pi^c}{4} & \frac{1-\pi^c}{4} & \pi^c & \frac{1-\pi^c}{4} \\ \frac{1-\pi^c}{4} & \frac{1-\pi^c}{4} & \frac{1-\pi^c}{4} & \frac{1-\pi^c}{4} & \pi^c \end{bmatrix}. \tag{10}$$

The parameter π^c represents the probability that a student correctly identifies the answer after receiving the signal conditional on this answer being correct, as in equation 4. We assume:

$$0.2 \leq \pi^c < 1. \tag{11}$$

Thus, X encodes a more structured belief updating process, where students are more likely to believe the correct answer when receiving the signal.

Establishing Blackwell Informativeness

A signal structure X is said to be more informative than Y in the Blackwell sense if there exists

a stochastic matrix G such that:

$$Y = GX. \tag{12}$$

That is, Y is a garbling of X, meaning Y can be obtained from X by applying some probabilistic transformation.

To construct such a matrix G, we solve:

$$G = YX^{-1}. \tag{13}$$

Given the structure of X, we find that:

$$G = \begin{bmatrix} 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \end{bmatrix}.$$

This matrix satisfies the properties of a stochastic matrix: all elements are non-negative, and each row sums to 1.

Thus, we have successfully shown that Y can be obtained from X via the garbling matrix G, confirming that X is more informative than Y in the Blackwell sense.

A.1 Effect of Increasing π^c on Informativeness

To show that X becomes strictly more informative as π^c increases, we analyze how the off-diagonal entries of X behave. The informativeness of a signal structure increases when posterior beliefs are more dispersed. An example is a mean-preserving spread.

Consider two key properties:

- Higher Correct Answer Probability: As π^c increases, the diagonal elements of X increase, which means that students become more confident in the correct answer.
- Lower Confusion Probability: The off-diagonal terms $(1 - \pi^c)/4$ decrease, which means that students are less likely to assign probability incorrectly to the wrong answers.

Since higher dispersion in the posterior beliefs indicates a more informative signal structure, increasing π^c ensures that the beliefs become less uniform and more concentrated on the correct answer. This leads to a higher mean-preserving spread in posteriors, which is a well-known measure of increased informativeness under Blackwell's theorem.

We have shown that the signal structure X is more informative than Y under the Blackwell information ordering by constructing a stochastic garbling matrix G. Additionally, we have shown that as π^c increases, X becomes strictly more informative, as it results in a greater concentration of probability on the correct answer, reducing the ambiguity in student beliefs.