Innocuous Exam Features? The Impact of Answer Placement on High-Stakes Test Performance and College Admissions*

Catalina Franco† Erika Povea‡

June 25, 2024

Abstract

We exploit randomness in college entrance exams in Colombia to study how the placement of answers impacts multiple choice test results and access to college. Using administrative data, we find that: first, applicants are 5% less likely to answer correctly when the correct answer is the last in the choice set (option D). And, second, that one SD higher share of correct answers in D in the math section reduces applicants' overall performance and their preferred major admission rate by 3%. Considering lifelong college access implications, we show how seemingly innocuous exam features disproportionately affect unlucky test takers.

JEL CODES: C93, D83, I21, I23, I24
KEYWORDS: Multiple choice tests, answer placement, performance, admissions

*We thank Universidad Nacional de Colombia for generously sharing their administrative data with us. We obtained valuable feedback from Martin Brun, Aline Bütikofer, Valentina Duque, Leonardo Garzón, Marcela Gomez, Brian Jacob, Katrine Løken, Robert Metcalf, Katherine Michelmore, Germán Reyes, Tanya Rosenblat, Frank Schilbach, Erik Sørensen, Oda Sund, Heidi Thysen, Alexander Willén, and participants at the U. Michigan Causal Inference in Education Research Seminar (CIERS), Bergen-Berlin Behavioral Economics Workshop, UiB-NHH PhD workshop, the SNF Brown Bag Lunch, the Midway PhD seminar at NHH, and the 2nd Workshop on Field Experiments in Economics and Business in Düsseldorf.

†Center for Applied Research (SNF) and FAIR at NHH, Helleveien 30, 5045 Bergen, Norway. Email: catalina.franco@snf.no, corresponding author.

‡Economics Department and FAIR at NHH Norwegian School of Economics, Helleveien 30, 5045 Bergen, Norway. Email: erika.povea@nhh.no.

1 Introduction

Performing well in multiple-choice standardized tests is highly consequential as they can give access to educational opportunities that affect future economic outcomes. Despite the recent and growing debate in countries like the US to eliminate standardized testing from college applications,[1] standardized tests are still very popular because they are thought to be designed objectively, and to provide a fair, efficient and cost-effective mechanism to assess performance and allocate educational slots. One concern with standardized tests, especially in high-stakes settings, is that performance differences may reflect inequities in the testing process rather than underlying ability (Duquennois, 2022). In fact, it has been shown that seemingly minor and random factors —unrelated to academic ability— can affect standardized test performance. For example, environmental aspects such as the level of pollen (Bensnes, 2016), pollution (Ebenstein, Lavy, & Roth, 2016) or temperature (Park, 2022; Chang & Kajackaite, 2019) negatively affect performance. Features of the test itself such as the format (multiple-choice vs. open ended or written vs. oral) (Griselda, 2022; Andresen & Løkken, 2020), monetary themes in questions (Duquennois, 2022), gendered language (Cohen et al., 2023), and aspects related to cognitive fatigue (Sievertsen, Gino, & Piovesan, 2016; Brown, Kaur, Kingdon, & Schofield, 2022; Reyes, 2023) can also negatively affect the performance of certain groups of students such as female and low-income students.

This paper exploits a source of random variation resulting from the placement of the correct answer on multiple-choice questions in a high-stakes college entrance exam. While existing research provides causal evidence on how more noticeable sources of randomness impact outcomes beyond performance (e.g., randomization of exam subject as in Landaud, Maurin, Willage, & Willén, 2024),[2] we focus on a source of variation that most would consider to be completely innocuous and show that it can have important implications for inequality in performance and college admissions.

We leverage the random allocation of exam booklets across students and of correct answer options across exam booklets in the college entrance exam (CEE) to Universidad Nacional, Colombia's top public university. To prevent cheating in the exam, the university preserves the order of the

[1] Nearly 2,000 colleges across the US removed test score requirements due to equity and diversity concerns as well as the pandemic. In April 2024 the New York Times reports that many selective universities are reinstating test score requirements: https://www.nytimes.com/2024/04/11/us/harvard-test-scores-admissions.html.

[2] Landaud et al. (2024) find that students who are lucky to be randomized into exam subjects they excel in have higher high school graduation rates and earn higher wages.

questions in all booklets, but randomly assigns whether the correct answer to every question is in location A, B, C or D. All applicants assigned to the different booklets undergo identical exam and testing conditions, except for the placement of the correct answer. The exam takers have 3.5 hours to answer 120 questions on paper in five subjects presented in this sequence: reading, math, science, social science, and Raven's test-style questions. All components are weighted equally in the overall score and regardless of the college major that the applicants intend. The stakes for most students are typically high since tuition in public universities in Colombia is highly subsidized and there are few scholarship or loan programs for students (Londoño-Vélez, Rodríguez, & Sánchez, 2020).[3]

To conduct the analysis, we obtained administrative data from the university detailing question-by-question performance for every applicant taking the CEE for admission to the first semester of 2020 and 2023.[4] Our study includes a sample size of 95,873 applicants who were assigned to 14 booklets and applied for admission to one of the university's five main campuses across the country. In addition to the question-by-question data, we have access to data on the random allocation to booklets, a broad set of demographic and schooling characteristics of the applicants and their parents, applicants' scores in each subject and overall score in the CEE, and whether they were admitted to their first choice in the admission cycle to which they are applying. In addition, for the 2020 cohort, we obtained matched data with the national administrative registry of higher education attendees. Hence, we can observe which institution applicants eventually enroll in, if at all, up to the first semester of 2021.

Several interesting findings emerge from our analysis. First, we examine how the placement of the correct answer impacts the likelihood of answering a question correctly. By leveraging random variation at the question level, we estimate a two-way fixed-effect model with student and question fixed effects. Comparing the probability of answering correctly when the correct answer is in A (approximately 44% across the whole exam), we note a slight decrease in the likelihood of answering correctly when the correct answer is in options B and C, and a substantial decrease when it is in

[3]Due to private universities outnumbering public universities, every year, nearly 100,000 applicants, or about 20% of the high school graduating class in Colombia take the CEE for admission to U. Nacional. The admission rate across all majors is about 10%. Admissions to private universities require another standardized test (SABER 11) that all high school graduates take in their last year of high school.

[4]These exams correspond to the last CEE before and first after Covid. The university did not administer a CEE to determine admissions between these two years.

option D. Specifically, students encountering questions with the correct answer in location D are 2.3 percentage points (pp) (or 5.3%) less likely to answer correctly compared to when it is in A. We term this phenomenon the "D effect" henceforth.

In descriptive analyses of the distribution of student answers in locations A, B, C and D, it is noticeable that students tend to choose D less frequently than the other answer options and than what the true distribution of correct answers suggests they should. We conjecture that students may experience a "primacy effect," which would suggest that items at the beginning of a list receive more attention than items at the end (Asch, 1946). For example, students may miss or ignore answers towards the end of the choice set because they find a potential candidate solution among the first options or do not take enough time to carefully assess all answer options (Feenberg et al., 2017). To provide external validity of this pattern, we analyze the pattern of student answers from over 15 million test takers using data from the 2018 PISA test and the 2016-2019 ENEM high school assessment exam for admissions to Brazilian colleges in Figure 3.[5] We observe that in most countries in the PISA test, with the exception of the two top scorers (China and South Korea), students are less likely to select option D. Similarly, in Brazil, where options go from A to E, students are substantially less likely to choose option E.

Who can we expect to be most affected? In our conceptual framework we propose effects for two types of students: high (H) and low (L) types. H types have higher ability and are more focused, hence have a higher likelihood of knowing or finding the correct answer. Low types, on the other hand, have the opposite characteristics, making them less likely to know or find the correct answer. We expect both types to exhibit primacy effects due to the challenging nature of the exam. However, we anticipate larger effects among low types, as high types are likely to find the correct answer and to be more meticulous when reviewing the answer options. We find that the D effect is more than twice as large for low than for high types (3.4 pp vs. 1.4 pp, respectively).[6]

Next, we study the implications of having a lower probability of answering correctly when the

[5] We thank Germán Reyes for sharing the Brazilian data with us. As far as we know, these tests do not randomize the placement of the correct answer. However, this information is something that test administrators usually do not provide perhaps because it is considered to be innocuous.

[6] We define high and low types to be students in the top and bottom quintiles of the CEE performance distribution. We anticipate that there may be some endogeneity at the limits of the quintiles since some students who would be otherwise classified in a quintile are observed in a different one due to the share of correct answers in D that they face. We note that this will not affect our main outcomes related to admissions since less than half of the students in the top quintile are admitted, so those students are far from the limit between the top and the second best quintile.

correct answer is in D. We use the random allocation of booklets, which, based on their construction, offer different shares of questions with correct answers in location D.[7] We first compute the standardized share of answers in D for each of the five exam subjects and estimate the effect on the standardized score of each subject. Being randomly assigned to a booklet with a share of D that is one SD above the mean in math (6.7 pp over a mean of 0.2536) reduces the score in that subject by 0.025 SD. No other subject-specific score is significantly affected by its own share of D. Hence, our subsequent discussion of the overall exam performance and college admissions outcomes focuses on the effect of the share of D in math since we find that the D effect mainly affects math performance and math is typically considered as a subject that is highly predictive of future academic and labor market success (e.g., Falch, Nyhus, & Strøm, 2014; Joensen & Nielsen, 2016).

Our final and most important result relates to the effects of having a larger share of questions with correct answer in D on academic outcomes with future economic consequences: Overall performance and admission to applicants' first choice major at U. Nacional.[8] A share of D one SD higher than the mean in math decreases students' overall performance in the exam by 0.01 SD and reduces the likelihood of being admitted to their first choice major by 3%. The share of D in other exam subjects does not have any effect on performance or admissions outcomes. In addition, we find no differential effects on the performance and admission outcomes for women relative to men or for low SES relative to medium SES students.[9] However, using matched data with the universe of higher education attendees, we find long-lasting consequences for male applicants with a high share of D in math. Within 2 years after taking the exam, these applicants are 0.5 pp less likely to attend any college than those who did not face a high share of D in math in the CEE to U. Nacional.[10]

The substantial impact of the share of D in math on overall performance raises questions about whether math is inherently challenging or if other factors contribute to underperformance in this

[7] The university does not purposely balance the shares of correct answers in different locations across subjects or booklets.

[8] Franco and Hawkins (2022) provide a thorough explanation of how applicants select major choices and how college slots are assigned at this university during the 2020 admission cycle. There were a few changes in the admissions process in 2023; however, these changes do not impact our analyses.

[9] Very few high SES students apply for admission to this university since it is more likely that they attend elite private universities.

[10] Being a public university, U. Nacional charges tuition fees according to the applicants' family income. Applicants who cannot get access to this university have limited options as most possibilities involve private universities which charge high tuition fees.

subject. According to the university's Rasch model used for exam scoring, math is identified as the most difficult exam subject based on the assessment of individual item difficulty. This difficulty alone can contribute to the D effect, as students facing harder questions may resort to "satisfice" by selecting the first seemingly correct option down the list, and to "skim" due to limited time, prioritizing questions in other subjects where they are more likely to know the answers. However, the time constraint can confound the effect of math being a hard subject. To assess whether this is the case, we estimate separate effects for two instances: in 2020 the exam questionnaire was removed after the math section, creating a hard time cap after math, and in 2023 a single questionnaire was used for the whole exam. We find that the effect of the share of D on the math score is similar in size in both years, suggesting that math is indeed a hard subject. We note that the D effect gets monotonically larger as the math section progresses in 2020 when the questionnaire was removed, which we take as evidence that when the time constraint is binding, students may "satifice" and "skim" even more.

Finally, we analyze alternative mechanisms to satisficing and skimming including cognitive fatigue, non-responses, guessing, and questions with answer options of the type "none or all of the above," which usually appear in location D. We are able to rule most of these mechanisms, and provide evidence that guessing may be part of the explanation but cannot explain the full effect. Ultimately, disentangling satisficing (e.g., selecting C because that is the option that seems correct) from guessing (e.g., selecting C because the heuristic says to choose C when guessing) would be difficult even in experimental settings.

We build on the growing literature exploring the effects of exam features on academic performance (Andresen & Løkken, 2020; Duquennois, 2022; Griselda, 2022; Cohen et al., 2023). This literature has focused on uncovering effects of exam features that disproportionately affect subgroups of students such as women or low-income students. Our study contributes with a previously-unexplored exam feature, namely the randomness of the correct answer location. By demonstrating that students facing a higher proportion of questions with correct answers located in option D in mathematics are less likely to secure admission to their preferred major, and highlighting the heterogeneity effect in overall college admissions for male applicants, our study stands among the first in uncovering the broader implications—beyond exam performance—of random exam features. The paper by Landaud et al. (2024) uses randomness in the exam subject that students in Norway have

to take at the end of high school. Given the advantages of the Norwegian register data, they can document long-term effects of being tested in unlucky exam subjects on academic and economic outcomes.[11] Relative to that study, where the source of variation is arguably heavy handed and probably not found in many other settings across the world, we show the detrimental consequences of random exam features using a source of variation that most would consider completely innocuous.

We also contribute to the literature on factors affecting exam performance due to random variation in environmental factors. The impact on performance we find is of similar size relative to previous work studying the effects of a one SD increase in air pollutants leading to performance declines of 0.025 SD (Ebenstein et al., 2016; Bensnes, 2016), a 10 pp increase in monetary-themed math questions leading to a performance reduction of 0.026 SD (Duquennois, 2022), and taking a test every hour later in the day reducing performance by 0.01 SD (Sievertsen et al., 2016). We consider our results to be meaningful given that they emerge from a measure to prevent cheating that is not supposed to affect students' health, psychological well-being or cognitive abilities in any way.

Finally, we contribute to the literature in educational psychology examining whether the position of response options matters. Hohensinn and Baghaei (2017) posit that answer options towards the end of the choice set make the question slightly more difficult. Attali and Bar-Hillel (2003) find that questions with correct answers in the middle are easier and less discriminating due to "edge aversion" and "middle bias" according to which test constructors hide the correct response in the middle and test takers correctly anticipate this. Our paper is the first to quantify the effect of the correct answer location on the likelihood of answering correctly and to investigate the effects of facing different shares of questions with correct answers in "bad" locations on important academic outcomes. In addition, our analysis of patterns of student answers in the CEE, the PISA test and the Brazilian CEE indicates that answer options toward the end of the choice set do not make questions harder, but that students tend to not peruse the full choice set.

Overall, our findings show that exam factors that appear to be harmless or insignificant can generate inequalities which can in turn contribute to unequal outcomes or advantages for certain individuals. Our paper provides invaluable insights for designing or grading exams in a way that

[11] The study by Falch et al. (2014) is more limited in scope but also uses the randomization of the exam subject at the end of middle school in Norway together with the announcement of the specific subject the students will be evaluated on a few days before the exam to see effects on performance.

enables students to fully realize their academic potential. An actionable policy implication for educational institutions interested in maintaining the randomization of the correct answer location is to assess the magnitude of the resulting disadvantage and incorporate it into the weighting of the final score. By considering and adjusting for this factor, institutions can strive for a fairer evaluation of student performance and enhance the overall integrity of the assessment process.

2 Institutional Background

The Colombian higher education system comprises both public and private schools. Admissions to universities of both types typically occur twice a year. Private universities assess applicants through interviews and scores on the national standardized exam, while many public universities have their own college entrance exams (CEE) and base admissions solely on individual performance in that exam. In Colombia, high schools do not segregate or track students. Public universities subsidize tuition fees based on the household income of the student. Consequently, numerous Colombian students, particularly those from low and middle socioeconomic status (SES) families, aspire to secure a place at a public university due to their affordability and prestige.

For our research, we use question-level data from the CEE to Universidad Nacional de Colombia, the country's largest and most prestigious public university. The CEE score is the sole requirement for admission to this university. Due to the limited availability of affordable and high-quality higher education institutions in Colombia, there is intense competition for the limited number of slots, which significantly increases the importance of the CEE for students. Universidad Nacional de Colombia offers approximately 5,000 slots per semester nationwide, attracting over 50,000 applicants each semester. Students can take the CEE as many times as they wish, as long as they pay a small registration fee (around $25 USD) and are not currently enrolled at the university.

2.1 The College Entrance Exam

The university conducts the CEE twice a year, in March for admission in August and in September for admission in January. All applicants show up at a specific date and time to take the CEE in person using a paper-based format at their individually assigned location. The exam is the same for all applicants regardless of their intended major. Furthermore, the CEE score is exclusive to

this university and is not utilized by any other institution, nor do other universities have access to the exam questions or scores. The exam questions are designed by a committee of professors within the university. The Admissions Office, which is in charge of scoring the exam and assigning slots, and hence us, cannot have access to the questionnaires to increase the level of independence in the scoring process. In addition, the weight given to scores in individual exam subjects is the same regardless of which major the applicant intends to enroll in.

The duration of the CEE is 3.5 hours, during which candidates answer a total of 120 multiple-choice questions. The flow of the exam is as follows: The first block of questions (1-39) are part of a general subject that is mostly reading comprehension questions,[12] the second block contains math questions (40-60), the third contains natural science questions (61-80), the fourth contains social science questions (81-100), and the last subject is a Raven's-type test (101-120). During the scoring process, the Admissions Office usually drops a small number of questions that they identify as being problematic; hence, the actual number of questions we use in the analysis is below 120 (115 in the 2020 data and 116 in the 2023 data). It is important to note that there are no penalties for incorrect answers. The exam is graded using a scantron machine, and once the scores are generated, they cannot be disputed or revised.

2.2 Correct Answer Placement

All questions in the exam have four answer options: A, B, C, and D, listed vertically on the paper questionnaire. To avoid cheating, since applicants are sitting relatively close to each other, the university randomizes where the correct answer is placed in every question. To implement the randomization in practice, the university creates different booklets to limit the number of versions to be printed. The students are aware that not everybody is taking exactly the same exam, but are not necessarily informed that the source of the difference is the location of the correct answer. Other than the random order of the correct answer, all applicants face the same questions, the same order of exam subjects and the same order of questions within subject. Although students are not required to follow the order in which the questions are presented, there is a strong incentive to do so given how the paper questionnaire is folded.

Between the two datasets we employ, the questionnaires differ in one aspect. In the 2020 dataset,

[12] A few of these questions are in math, science or social science but always related to the text that students read.

the university employed two different questionnaires, whereas in the 2023 data, a single questionnaire was employed, encompassing all 120 questions. In 2020, the first questionnaire containing reading and math was collected after two hours and replaced with the second questionnaire that included the three remaining exam subjects to be solved in 1.5 hours. Students could not get the first questionnaire back in case they finished early with the second questionnaire.

3 Conceptual Framework

Consider two types of students: high and low. High (H) types are students of higher ability who typically can attain a high level of focus and hence their likelihood of knowing or being able to find the correct answer in any exam question is high. Conversely, low (L) types have lower ability and do not remain focused for long periods of time, so their likelihood of knowing or finding the correct answer in any question is low.

Since most students will be strapped for time when taking the exam since it is a long, difficult and highly academically discriminating exam, all students will be likely to experience primacy effects and cognitive fatigue.[13] However, given the characteristics of the types described above and recent evidence that students with certain features (e.g., richer students) may have higher cognitive endurance (Brown et al., 2022), we hypothesize that any effect of the correct answer placement will disproportionately affect low types relative to high types. For example, high types may be more careful going through all the answer options and may manage time better, both of which might help them be less affected by any primacy effect or cognitive fatigue.

We base our conceptual framework on the behavioral theories developed by Feenberg et al. (2017) to rationalize why individuals receiving the NBER's weekly working papers announcement are more likely to click on, download and cite the first paper listed in the email. Three main theories are insightful in our setting.[14]

The first model based on Feenberg et al. (2017) suggests that students "satisfice," that is, students start perusing the answer options down the list and stop once they find an option that

[13] There is no official ranking of college entrance exams in Latin America or Colombia according to their level of difficulty. Some resources speaking of the difficulty of the exam are (in Spanish): https://www.eltiempo.com/vida/educacion/que-tan-dificil-es-el-examen-para-ingresar-a-la-universidad-nacional-414132

[14] We discuss the role of guessing in Section 7. We do not consider guessing as being based in a theory, but rather a heuristic that students apply when they do not know the answer to a question. We note that disentangling guessing from satisficing may be implausible.

seems correct. If the students know what the correct answer is and they can spot it easily, we expect no satisficing. However, if this is not the case, students will evaluate options one by one and select one that seems reasonable without going all the way down the list given the time constraint. Note that, the other answer options, usually called the "distractors," appear before the correct answer when it is in D. Distractor options are plausible but incorrect answers based on students' common errors or misperceptions. So, in some cases, students may know the answer or have a good idea of what the answer could be, but they instead mistakenly choose one of the distractors since it is a plausible answer and it appears higher up in the list.[15]

The second model adapted from Feenberg et al. (2017) involves "skimming" of answer options under tight time constraints. Consider a student who realizes that she will not have enough time to complete all questions in the exam. Her marginal cost of assessing all answer options for every question increases because the opportunity cost of time is increasing. Hence, the student will not carefully consider all answer options as she would under no time constraints. We then expect that skimming will be more common when the time constraint becomes salient, for example, after finishing a exam subject and realizing that one does not have enough time for the following subject.

In a third model we posit that "cognitive fatigue" prevents students from carefully evaluating all answer options because the marginal cost of searching for answers increases since students become increasingly tired. Consequently, we would anticipate observing smaller effects of the location of the correct answer option at the beginning of the exam and larger effects as the exam progresses. Alternatively, the location of the correct answer itself may generate or increase cognitive fatigue. For instance, if multiple questions have the correct answer in option D, it involves more extensive "searching" for the answer. This increased cognitive effort may progressively fatigue students as they need to assess a greater number of answer options compared to those whose correct answers appear higher up in the list.

Overall, these conceptual frameworks could explain why different types of students, of high and low ability, may be differentially affected by the placement of correct answers in a multiple-choice exam. We hypothesize that low-ability students would be disproportionately affected by answer placement compared to high-ability students, based on theories of satisficing, skimming under time

[15] If the correct answer is in A, for example, students do not have the chance to fall for the distractors since they see the an answer that seems reasonable, which happens to be the correct answer, first.

constraints, and cognitive fatigue. These mechanisms suggest that low-ability students may be more prone to fall for distractors if correct answers appear later in the list, may not carefully consider all options and could experience increased cognitive fatigue as the exam progresses, impacting their performance more significantly.

4 Data, Sources of Variation and Descriptive Statistics

4.1 Sources of Data

Our main data source is the CEE to U. Nacional. There are two waves of data collection in which 50,608 applicants were examined for admission in the first semester of the 2020 cycle, and 45,265 undergo the examination for admission in the first semester of 2023. In the sample encompassing both exams and after dropping data from students who take a special version of the exam (e.g., disabled students) we consider a total of 95,873 applicants. While we mainly use the individual-level dataset, we also analyze the choices made by every applicant at every question, which leads to over 11 million applicant-question observations.

Our final dataset is a combination of two sources: a survey completed by each applicant during the exam registration and the exam-related question-by-question choices and their performance. For the 2020 cohort only, we linked our data to the administrative registry from the National Higher Education Information System (abbreviated SNIES in Spanish) available from the second semester of 2019 (when our students take the exam) up to the second semester of 2021 (two years later).

The survey provides insights into the applicants' socio-demographic characteristics and their intended major. We have extensive self-reported information about the student and their family, including details about parental education, number of siblings, socioeconomic status, among others. The "SNIES Data" contains individuals' attendance histories at all higher education institutions in Colombia. For every attendee, every semester, the SNIES data provides information on the attendees' major name, institution name and institution type (academic or technical). Importantly, the SNIES allows us to follow applicants to Universidad Nacional who do not enroll there.

To assess the external validity of the patterns of student answers we find, we analyze student responses from other multiple-choice exams worldwide. The High School Assessment Exam (ENEM) dataset in Brazil comprises information from approximately 15 million students who took the exam

between 2009 and 2016 (Reyes, 2023). Admission to highly selective federal universities depends on ENEM scores. This exam evaluates four subjects (language arts, math, natural sciences, and social sciences) along with an essay, encompassing 180 multiple-choice questions. Similar to the CEE, examiners implement random allocation of test booklets. Although the sequence of subjects and the question sets remain consistent across booklets, the order of questions within each subject is randomized. In this case, the multiple-choice options range from A to E. We use this dataset to see whether students are less inclined to choose option E, the last option in that exam, and to assess whether D is equally chosen relative to A, B and C when alternative E is present. Additionally, we analyze student responses from the PISA exam conducted in 2018 to determine whether the phenomenon of avoiding the last option in multiple-choice questions persists. PISA randomizes the order of the questions (Borghans & Schils, 2015; Borgonovi & Biecek, 2016; Balart et al., 2018; Akyol et al., 2021). PISA 2018 assessed approximately 600,000 students from 79 participating countries and economies, focusing on mathematics, science, and global competence. Based on our online search, we could not find information on whether PISA or ENEM randomize the order of the answer options.

4.2 Sources of Random Variation

We exploit two sources of random variation. First, we have variation at the applicant-question level, that is, every question a given applicant faces has a correct answer in a location that has been randomly assigned by the Admissions Office to be in A, B, C or D. Figure 1, Panel (a) plots this source of random variation. Second, we use the random assignment of applicants to exam booklets. There are 14 types of booklets, eight in the 2020 exam and six in 2023. By random chance, some booklets have a higher fraction of correct answers in certain locations. Figure 2 shows the variation we exploit in each of the five exam subjects. Each subplot shows the frequency of students facing the shares of answers in location D specified in the horizontal axis. In our subsequent analysis, we will refer to this as "share of D", representing the proportion of correct answers in D relative to the total number of questions.

4.3 Outcomes of Interest

Using the source of random variation at the applicant-question level, our main outcome of interest is whether a given question is answered correctly when the correct answer is in B, C or D relative to when it is in A. We have 11,070,660 applicant-question observations to examine this.

Using the variation from the share of D in each exam subject and overall in the exam, we assess how a large share of correct answers in D impacts three main outcomes. First, we study the impact on same subject performance and overall performance in the exam. For the former we use the share of D in each exam subject and see how the standardized score in the same subject is affected. For the latter, we see how the share of D in each subject and the overall share of D in the whole exam affect the total standardized exam score used for university admissions. Second, we assess the impact on whether the applicant is admitted to his or her first choice major in the admission cycle in which they take the CEE.[16] Third, using the linked dataset to the universe of higher education attendees, we examine whether applicants taking the CEE ever attend a higher education institution within two years of the CEE (only available for the 2020 cohort). Furthermore, we split this variable into ever attending university, ever attending a university other than U. Nacional, ever attending an elite university[17] and ever attending a technical/vocational higher education institution.[18]

4.4 Descriptive Statistics

The applicant-question variation is summarized in Figure 1, Panel (a), where we see that across the exam and all booklets, 25.5% of correct answers are in A, 24.1% in B, 25.7% in C and 24.7% in D. If we compare the fraction of correct answers in location D to the fraction of student answers selecting D (Panel (b)), students systematically choose D less that they should (compare Panel (b) to (a) in Figure 1). This holds true across all exam subjects (Panel (b), Figures A1-A5). Importantly, applicants avoiding the last alternative in the choice set is not exclusive to the CEE in Colombia.

[16] If not admitted, applicants can retake the CEE as many times as they want in the following admission cycles provided they pay a small fee.

[17] "Elite" is defined as the top ten Colombian universities in the Quacquarelli Symonds (QS) 2023 world rankings: Universidad de Los Andes (private), Universidad Nacional (public), Universidad Javeriana (private), Universidad Pontificia Bolivariana (private), Universidad Externado (private), Universidad Icesi (private), Universidad de Antioquia (public), Universidad de la Sabana (private), Universidad del Rosario (private), Universidad Eafit (private).

[18] The programs at technical schools are similar to associate degrees in the US since they are shorter and focus on a professional skill such as computer technician or accounting assistant. They are often considered an inferior option to attending university.

In Figures 3 and A6 we provide descriptive evidence that students taking other standardized tests worldwide from the sources described above are less likely to choose the last option in the choice set.

The share of D plotted in Figure 2 for each exam subject shows a wide range of variation in most subjects from 0.1 up to 0.4 of correct answers in location D. The bottom panel of Table 3 shows the mean and SD of the share of D variable for each exam subject. The means are between 0.2401 and 0.2536, and the SD across all subjects is close to 0.07. We use this standardized variable to estimate the effect of a higher share of D on our main outcomes relating to exam performance and access to higher education.

5 Empirical strategy

5.1 Effects on Question Performance

We use the applicant-by-question data to assess whether the location of the correct answer influences the likelihood of obtaining a correct response. We employ a two-way fixed-effect model described in Equation 1 in which the outcome variable is an indicator for correct answer for applicant i in question q. The variable of interest x_{iq} indicates the location of the correct answer in one of the alternatives A, B, C or D. We include fixed effects for student γ_i, question λ_q, and year of testing ω_t. The standard errors are clustered at the booklet level in all specifications.

$$y_{iq} = \delta_l x_{iq} + \gamma_i + \lambda_q + \omega_t + u_{iq} \tag{1}$$

The student fixed effects capture variation across exam questions within applicant, and the question fixed effects capture factors that do not vary within question but vary across students. Since we have two years of data we also include indicators for year. Controlling for these fixed effects, the coefficient δ_l gives the causal effect of being assigned a question with correct answer in location l (B, C or D) on the likelihood of answering the question correctly relative to when the correct answer is in A.

To study the outcomes of different types of applicants proxied by quintiles of exam performance, we run Equation 1 separately for each subgroup.

5.2 Effects on Academic Outcomes

To study the effects of the share of D on academic outcomes we use the random allocation of booklets to applicants. We use this variation in the following specification for each exam subject separately and for the overall share of D across all exam subjects:

$$y_i = \alpha + \beta \text{ Share of } D_i + \varepsilon_i \tag{2}$$

Where the variable Share of D_i is the standardized version of the fraction of questions with correct answers in location D for applicant i. The fraction of correct answers in D is randomly assigned as the booklets are constructed based on the random allocation of correct answer options to each question. Since the share of D is standardized with mean zero and SD equal to one, the coefficient β provides the causal effect of a one SD higher share of D on the academic outcomes defined in Section 4.3. We include campus and year fixed effects in the results tables to account for potential differences among applicants applying to the different university campuses and year in which they apply. Standard errors are clustered at the booklet level.

6 Results

6.1 Effects on Question Performance

Table 1 presents the results of the two-way fixed-effects model, where the dependent variable is an indicator of answering a question correctly. The omitted category represents the correct answer located in alternative A. Overall, when the correct answer is in location A, the probability of answering the question correctly is approximately 44% (Column 1).

Columns (2)-(4) account for student fixed effects, Columns (3)-(4) include question fixed effects, and Column (4) incorporates both types of fixed effects along with an indicator for the year of testing. Both the last two specifications are considered equivalent, indicating that the impact of the answer location remains consistent across different years of testing, including periods before and after the Covid pandemic. Including all fixed effects in Column (4) we observe that the probability of answering a question correctly decreases by 0.29 pp if the correct answer is in location B and 0.37 pp if it is in option C. Both are small effects, even if the coefficient for location B is significant,

especially when compared to the reduction in the probability of responding correctly when the correct option is D equal to 2.3 pp. This point estimate implies that applicants facing a question with correct answer in D are 5.3% less likely to answer it correctly. The visual representation of the D effect by question deciles is in Figure 4. At the start of the exam the effect is already about 2 pp, becomes larger in question deciles corresponding to the math and science subjects, to go back to levels similar to the beginning of the exam in social science and Raven's.

In Table 2 we split the sample by quintiles of exam performance. The first quintile contains the 20% of students with the lowest overall scores, while the fifth quintile contains the applicants with the top 20% scores in the exam. By doing so, we can observe the students' types characterized in Section 3 where we can think of the bottom quintile as type L applicants and the top quintile as type H applicants. We reproduce Equation 1 for these sub-samples and observe that, in line with Table 1, the likelihood of answering correctly decreases substantially across quintiles for correct answers located in D. The effect decreases monotonically across quintiles, being largest at 3.3 pp in the bottom quintile and smallest at 1.4 pp in the top quintile. These values correspond to a drop of 11% and 2%, respectively, in the probability of answering correctly. The effect on correct answers located in B or C is only significant in the bottom quintile, where we also observe the probability of answering a question correctly declining monotonically by 0.3 pp if the correct answer is in option B to 0.6 pp if it is in option C, and then jumping to 3.3 pp when it is in alternative D. The dynamics of the D effect for H and L types is in Figure 4. Overall, these results suggest that L-types are much more sensitive to question order effects than more skilled students.

6.2 Effects on Academic Outcomes

We initially examine the impact of a share of D that is one standard deviation larger than the average on performance within the same subject, as presented in Table 3. At the bottom of the table we report the mean and the SD of the share of D for every subject. All means are around 0.25 and the SDs are between 0.067 and 0.079. That is, in general, our point estimates reflect the effect of facing an exam subject with a share of D of 0.32 relative to the mean of 0.25. The effect in the reading component (Column 1) —the first subject in the exam— is close to zero and not statistically significant. The effect for the remaining subjects is negative, but only significant for math (Column 2), in which having a share of D one SD above the mean reduces math performance

by 0.025 SDs.

Next, we turn to estimate the effects of a high share of D on the overall exam performance and admission to applicants' first-choice major at U. Nacional. The results are presented in Table 4, Columns (1) and (2), respectively. The estimates are obtained using variation in the share of D from the whole exam (Panel A) and from each exam subject (Panel B). Panel A shows that there is no effect of the share of D from the whole exam on overall performance or admission to U. Nacional. However, this share of D may not be particularly important given that there seems to be large heterogeneity of the impact of the share of D across the different subjects as shown in Table 3. Hence a large share of D in the whole exam may correspond with high share of D in some subjects but not in others, or in some specific questions types and not in others, which in the end compensate each other and do not generate any measurable effect.

Panel B of Table 4 shows that a one-SD larger than the mean share of D in math leads to a reduction of 0.01 SD in overall performance, which in turn translates into a 3% reduction in the likelihood of being admitted to the applicants' first choice major at U. Nacional. The overall admission rate in our sample is 9.76%. The point estimates of the share of D for the other subjects is close to zero and not statistically significant, which suggests that math may be special or different relative to other subjects (see the discussion in Section 6.4).

While these effects may initially seem small, the effect on math performance is similar in magnitude as the effect of a one-SD increase of ambient pollen (Bensnes, 2016) or air pollutants (Ebenstein et al., 2016), which decrease performance via health deterioration. In our setting, where the exogenous variation is minimal and its potential effects on students' health or cognitive abilities are not obvious, it is remarkable that the effects are similar in size to those observed in prior studies. Additionally, our estimates exceed those found by Duquennois (2022), who finds that a 10 pp increase in the fraction of monetary-themed questions reduce math performance by 0.026 SD among poorer students.[19] In her case, the most likely mechanism is attention capture since monetary-themed questions make scarcity salient, which leads to stress and inattention. The random order of answer options cannot in any foreseeable way hinder cognitive abilities since it is not supposed to affect students' health or psychological well-being.

[19]Note that the increment in the right-hand-side variable in her case (10 pp) surpasses ours (6.7 pp). Using the size of the effect she finds, we would expect a 0.017 SD decrease in math performance, which is smaller than our observed effect. Note also that she finds this effect on poorer students, while our effect is the average for everyone.

Relating to the broader economics of education literature, we compare our effect to the literature evaluating the effect of class size on test scores (e.g., Angrist & Lavy, 1999; Krueger, 1999; Cho, Glewwe, & Whitler, 2012; Angrist, Lavy, Leder-Luis, & Shany, 2019). A recent working paper finds that the implied effect size is zero across most of the 62 studies they analyze. After correcting for publication bias, the effect size in papers published in the top-5 journals is an increase of 0.035 SDs when reducing the class size by 10 students (Opatrny, Havranek, Irsova, & Scasny, 2023). Hence, our effect size would be equivalent to reducing the class size by 2.9 students.

Our last set of findings utilizes the SNIES dataset, which contains information about individuals attending higher education institutions.[20] In Panel A of Table 5, we present the impacts of a share of D one SD higher than the mean in math on various outcomes. These outcomes include whether applicants ever attended a university (Column 1), ever attended a university other than U. Nacional (Column 2), ever attended an elite university (Column 3), and ever attended a technical/vocational higher education institution (Column 4). The coefficients for ever attending a university and ever attending an elite university are negative, although significance is only observed at the 10% level. Nevertheless, these negative signs suggest potential adverse effects of a high share of D in maths on the broader higher education prospects of students. Further exploration into the heterogeneity of these effects follows.

6.3 Heterogeneity Effects

In various contexts, it is anticipated that the impact of random factors on exam performance might vary among specific student subgroups. For instance, allergens like pollen and pollution can disproportionately affect students prone to allergies, particularly boys and those from lower-income backgrounds, as these groups often experience higher rates of asthma and other air-quality-related conditions (Ebenstein et al., 2016). In our scenario, it remains unclear whether specific subgroups would be more susceptible to the influence of a larger share of correct answers in D. As mentioned earlier, we anticipate this source of randomness to be unrelated to physical or psychological well-being. However, if cognitive fatigue is the primary factor influencing our results, we might expect

[20] Overall, approximately 66% of our 2020 cohort sample is identifiable in the SNIES dataset. While there may be matching problems due to errors in ID numbers, we think the main reason for non-matches is that applicants do not attend a higher education institution in the follow-up period. Among students in the top quintile of exam performance, who we expect to attend higher education, the match rate rises substantially to 89%.

more pronounced effects on students from low socioeconomic status (SES), aligning with findings by Brown et al. (2022) indicating that poorer students exhibit a steeper cognitive decline curve during exams. Additionally, if a higher proportion of correct answers in D induces stress among students, women might be disproportionately affected, as previous research suggests that their performance is more affected by high stakes than the performance of men due to choking under pressure (Cai et al., 2019).

Table 6 illustrates the heterogeneity effects of a larger share of D in math on overall performance and admission to applicants' first choice major by gender (Panel A) and SES (Panel B). Following Equation 2, the analysis of heterogeneity involves incorporating a female/low-SES indicator and an interaction between the share of D and the female/low-SES indicator. We do not observe variations in overall performance or first-choice admission across these subgroups. This finding is reassuring, indicating that while some inequality arises from differing shares of correct answers in D in math, it does not exacerbate existing disadvantages for women and low-SES students in terms of college admission. This is noteworthy given that women and low-SES students typically perform less favorably than men and high-SES students in mathematics (OECD, 2015).

In Table 5, Panels B and C present similar heterogeneity effects by gender and SES, respectively, but focus on outcomes reflecting broader access to education, such as ever attending different types of higher education institutions. The only outcome with heterogeneous effects is ever attending university, where substantial gender differences are evident. Compared to male applicants with the mean share of D in math, men with one standard deviation larger of share of D are 0.54 pp less likely to enroll at any university within two years of taking the U. Nacional CEE. Conversely, the coefficient is reversed for female applicants, suggesting that this negative effect on university attendance is concentrated among male applicants. Overall, this heterogeneity suggests that male applicants may face long-term consequences from being randomly assigned a math exam with a high share of correct answers in D.

6.4 Why is Math Important?

The results above suggest that math may have particular implications for students' success in this high-stakes exam. We propose two explanations for why this may be the case. First, math is usually a subject where substantial gaps have been documented, especially in terms of gender

(Ellison & Swanson, 2010; Fryer Jr & Levitt, 2010). Many students suffer from math anxiety, which is correlated with lower academic performance (Carey, Devine, Hill, & Szűcs, 2017). The PISA 2018 results indicate that students tend to achieve fewer correct answers in math compared to other test subjects.[21] Given that math often produces lower scores and induces more significant variations in performance and anxiety globally, it is not surprising that we find large and significant effects of the share of D in math. Additionally, we use the item difficulty analysis performed by U. Nacional to check which exam subjects are more difficult. Math is the hardest subject in the exam, with several of the questions being correctly answered only by students whose overall performance is higher than 2 SD above the mean.

Math being the exam subject where applicants typically struggle the most has implications to better understand our results. If students were proficient in setting up and solving the math problems without difficulty, we would not expect to see a D effect, as students could arrive at the correct answer through systematic problem-solving. However, when questions are intricate, students may need to rely on eliminating incorrect answers, a process that, under time constraints, could lead to our observed results, especially when the correct answer is the last of the choice set.

A second explanation for observing effects primarily in math could be that, from the students' perspective, math may be perceived similarly to other exam subjects, but a binding time constraint artificially amplifies the detrimental impact of the share of D in math. To explore this, we compare the level and slope of the D effect in Figure 4 between the 2020 and 2023 exams. In 2020, when the questionnaire was removed after the math section, applicants needed to complete all math questions before the exam proctors took away the questionnaire corresponding to the first two exam subjects (reading and math). This potential time pressure might prompt students to rush through some questions, particularly in the second subject. In 2023 applicants could progress at their own pace through the math questions as they had a single questionnaire throughout the entire exam.

Figure A7 plots the D effect across exam question deciles separately for 2020 and 2023. The slope is decreasing in both cases, which is consistent with the D effect becoming larger in math, but it is much larger in 2020 when the questionnaire was removed after math. We interpret the evidence from 2020 as suggestive of applicants following their own pace to answer the questions in the reading section, and after finishing this section, realizing that they do not have enough time

[21] The OECD average performance in math in the PISA exam is 487, while it is 489 in reading and 489 in science.

to answer the math section. The likelihood of answering correctly in 2020 is about -2.2 pp at the start of the math section and goes down to -3.5 pp at the end of the math section, which is the moment in which the questionnaire is taken away. The decline in 2023 is not so sharp probably because the questionnaire was not removed and students answered at their own pace. Further, we find similar effects of the share of D in math on math performance for both 2020 and 2023 equal to -0.0251 SDs (0.0017) and -0.0216 SDs (0.0022), respectively. For this reason, we do not think that the time constraint is what only matters for explaining why we see effects of the share of D in math on performance and admissions.

Overall, we argue that the reasons why we find that the share of D in math affects outcomes are that math is an important and difficult exam subject, and that the addition of a tight time constraint increases how demanding this exam subject can feel to students. However, time constraints are not the only explanation for observing the effect of the share of D in math and not in other subjects.

7 Alternative Mechanisms

In our conceptual framework, we outlined various potential factors influencing our outcomes, with a focus on satisficing and skimming when faced with tight time constraints. This section presents findings from econometric analyses aimed at ruling out alternative mechanisms.

We first assess the evidence in support of cognitive fatigue as a mechanism. As explained earlier, it is possible to understand fatigue as a cause for the D effect if students are too tired to carefully consider all answer options in the choice set. This interpretation is rooted in the idea that the marginal cost of searching for answers increases as the exam progresses, particularly for fatigued individuals. Cognitive fatigue would imply that the D effect would be null at the beginning of the exam and would increase as the exam progresses, but Figure 4 illustrates that the effect starts at a level of -2 pp and has a similar level in the last exam subject. In fact, the effect becomes monotonically larger as the exam progresses and goes back to its initial level toward the end of the exam. If cognitive fatigue were the main mechanism, this pattern would suggest that students become less fatigued toward the end of the exam, which is unlikely since the exam is long and challenging.

Building on the existing literature highlighting that individuals from low socioeconomic status

(SES) backgrounds experience higher levels of cognitive fatigue (Brown et al., 2022), we leverage our heterogeneity analysis by SES in the preceding section, which revealed no divergent effects among low-SES students. If cognitive fatigue were the driving force, we would anticipate more pronounced effects of the share of D in math for low-SES students.[22] Hence, we do not think that cognitive fatigue is generating the D effect.

Another way to understand cognitive fatigue is that correct answers in D promote higher levels of fatigue among students because they need to work harder to find the correct answer when they have a larger share of correct answers in D. To explore this idea, we undertake an exercise examining the specific arrangement of correct answers within the math subject. If correct answers in option D disproportionately deplete students' cognitive resources, we would expect the probability of answering correctly to decrease after a sequence of Ds compared to a similar sequence involving other letters. In the 2020 math exam section, we identify instances of consecutive correct answers within the same letter (e.g., C C and D D). We assess whether the likelihood of answering the next question correctly is affected by whether the preceding sequence is D D relative to C C, which would align with an immediate effect of cognitive fatigue. The findings from these mini-natural experiments, detailed in Table A1, do not provide evidence supporting that a sequence of Ds leads to a lower likelihood of answering the next question correctly.

Another potential explanatory mechanism for our results is non-responses. While we have been treating non-responses as incorrect answers thus far, it is plausible that correct answers in option D might lead to more non-responses rather than incorrect answers. In this scenario, the underlying mechanism would resemble satisficing/skimming, as students unable to find the answer within the options listed higher up would opt to leave the question unanswered. To investigate this possibility, we replicate the results of the two-way fixed effects model (Equation 1) using non-responses as an outcome variable, as outlined in Table A2. The point estimates consistently hover around zero, with at least three zeroes after the decimal point, and are statistically insignificant. Thus, we conclude that non-responses are unlikely to account for our observed results.[23]

[22] One caveat is that we do not observe high SES students in our sample since they typically apply to elite private universities. The literature documenting SES gaps in cognitive fatigue compare low- and high-SES students, so our null differences in this dimension could be because low- and medium-SES students in our sample may not differ substantially in their cognitive fatigue.

[23] The presence of the answer options "All of the above" or "None of the above" as the last item in the choice set is common in exams. Our effect could be explained if students avoid selecting these options because they are more involved than simpler answer choices. However, with only one out of 120 questions in the 2018 exam being of this

A related alternative mechanism is guessing. When not being sure of the correct answer, students may use heuristics to avoid leaving a question unanswered. For example, students are often advised in test preparation courses or in online resources to always select the same answer option when unsure, and to avoid extreme options (A and D in our case) under the argument that test makers assign answers to the mid-range more often.[24] Then it follows that guessing may explain our results if students choose B or C when they do not know the answer, rather than D because that is the advice the received. One reason why we believe guessing is not behind our results is that the estimates in Table 1 compare the likelihood of answering correctly when the answer is in B, C or D relative to A. Hence, the D effect we document compares D vs. A, so if the heuristic tells students to avoid extremes, we should not see a difference between when the answer is in D relative to when it is in A.

Another reason why we believe that guessing is not the main explanation behind our results comes from comparing the results of the two-way fixed effect model for students in the top quintile of performance who intend and do not intend a STEM major.[25] The basis for this exercise is that students who intend STEM may be less likely to guess, especially in math, since they must have a good domain of this subject given that they want to pursue a major in this field. We use only students in the top quintile of performance to avoid compositional differences of willingness to major in STEM across the exam performance distribution affecting our estimates. We replicate the results of Table 1 with this sample restriction and adding interactions with an indicator for whether the student intends a STEM major in Table A3. The D effect is smaller in magnitude for students intending STEM, but by no means the interaction coefficient reverses the main effect. Guessing, may thus be part of the underlying mechanism of the D effect, but it does not fully explain it. We note that separating guessing (choosing C because that is the heuristic) from satisficing (choosing C because that is the option that seems correct) is probably unfeasible.[26]

type, we do not believe it significantly contributes to the mechanism in our setting. We do not have access to the 2020 or 2023 questionnaires as the Admissions Office does not know the content of the questions.

[24] See for example the advice on this online post with over one million views: https://www.quora.com/When-you-guess-a-letter-on-a-multiple-choice-test-which-one-is-it.

[25] Quintiles are obtained based on overall performance in the same exam.

[26] Eye-tracking software in a controlled experiment may be helpful to separate these two to some extent, although it is unclear that it can detect small distance between answer options what would accurately mimic the layout of options in a real exam.

8 Robustness Checks

In this section, we assess the robustness of our results using two specifications. The first specification controls for the shares of answers in options B and C, rather than solely focusing on the share of D, as outlined in Equation 2.[27] This adjustment aims to eliminate the possibility that high shares of other answer options may be simultaneously or independently influencing the share of D result. For example, if there is a positive correlation between the share of C and the share of D, and it is the share of C driving the result, the expectation is that the effect of the share of D would change when incorporating the shares of correct answers in B and C.

The results of the specification adding the shares B, C and D are in Tables A4 and A5. Table A4 shows that, even after controlling for the other shares in answer options, math is the only subject where a higher share of D negatively impacts same-subject performance. The share of D coefficient is of similar magnitude as the one in Table 3.[28] In Table A5 we present the results for our three main outcomes: overall exam performance, admission to the first-choice major, and enrollment at any university within two years of the exam (only for the 2020 sample). The table indicates that the impact of the share of B and share of C in math is not statistically significant. Nevertheless, the share of D effect persists, aligning with our findings in the main tables regarding enrollment in the first choice major and university enrollment. Despite a slight increase in the standard error upon controlling for the share of B and C, the negative impact of share of D in math on the total exam score remains negative and of similar magnitude.

The second robustness check controls for the share of D in other subjects that may affect how the share of D in math operates. For example, a high share of D in math may occur simultaneously with a high share of D in reading and the reason why we see an effect of the share of D in math is partially because of the high share of D in reading. We account for the role of the composition of answer options in other subjects by adding the share of D in other subjects to Equation 2. The results from the specification adding the share of D in other subjects is in Table A6. The share of D in other subjects is not statistically significant and does not explain away the effect of the share of D in math on our main outcomes. If anything, the result on ever attending university becomes

[27] We do not include the share of A to avoid collinearity.

[28] The only coefficient that changes is the share of D in reading on reading performance, which goes from 0.0021 to -0.0213 and is marginally significant. The share of C in reading has a strong effect on performance of similar magnitude as the share of D in math.

larges and statistically significant at the 1% level.

9 Discussion and Conclusion

Our findings shed light on the impact of test design and correct answer placement on exam performance and access to higher education. Using random variation given by the location of the correct answer in a multiple-choice college entrance exam, we observe that test-takers are less likely to answer questions correctly when the correct answer is located in D compared to A. Moreover, a higher share of questions in the math section located in D has a negative effect on exam performance, admission to applicants' first-choice major, and on the likelihood of ever attending university. These results highlight the significance of seemingly innocuous features of a test in influencing educational outcomes.

Our results are likely applicable to various test settings, as we provide evidence that students in other tests worldwide, such as the PISA test and the college entrance exam for Brazilian universities, tend to choose the last option in the choice set less frequently in multiple-choice questions. Therefore, when the correct answer is in the last position, students will be less likely to answer the question correctly. We provide suggestive evidence that this behavior is driven by satisficing/skimming which potentially arises because students find a candidate solution before reaching the end of the choice set without carefully examining all the solutions. This would be more important when students are unsure of the correct answer, such as in difficult questions and exam subjects such as math, and under tight time constraints.

Regarding the policy implications of our research, one potential approach is to balance the answer key across exam subjects so that everyone faces the same share of correct answers in D in every exam subject. However, previous research has shown that savvy test takers can improve their SAT scores when they are aware of a balanced distribution of correct answers (Attali & Bar-Hillel, 2003). Another strategy could involve making adjustments to the Rasch model during the grading process to compensate individuals who face a higher share of questions in D, thereby leveling the playing field behind the scenes. This is the strategy that seems more reasonable and fair to implement if institutions still want to keep randomizing the order of answer options.

In summary, our research highlights the importance of considering the design and distribution

of test questions to ensure fair and equitable access to education. By understanding the impact of these factors, policymakers and educational institutions can take steps to mitigate biases and promote equal opportunities for all students given that performance is exams similar to the one we analyze opens the door to many educational and labor market opportunities for students.

References

Akyol, P., Krishna, K., & Wang, J. (2021). Taking PISA seriously: How accurate are low-stakes exams? *Journal of Labor Research*, *42*, 184–243.

Andresen, M. E., & Løkken, S. A. (2020). The final straw: High school dropout for marginal students. *Available at SSRN 3753354*.

Angrist, J. D., & Lavy, V. (1999). Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, *114*(2), 533–575.

Angrist, J. D., Lavy, V., Leder-Luis, J., & Shany, A. (2019). Maimonides rule redux. *American Economic Review: Insights*, *1*(3), 309–324.

Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, *41*(3), 258.

Attali, Y., & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement*, *40*(2), 109–128.

Balart, P., Oosterveen, M., & Webbink, D. (2018). Test scores, noncognitive skills and economic growth. *Economics of Education Review*, *63*, 134–153.

Bensnes, S. S. (2016). You sneeze, you lose:: The impact of pollen exposure on cognitive performance during high-stakes high school exams. *Journal of Health Economics*, *49*, 1–13.

Borghans, L., & Schils, T. (2015). *The leaning tower of PISA* (Tech. Rep.). Working Paper. Accessed February 24. http://www. sole-jole. org/13260. pdf.

Borgonovi, F., & Biecek, P. (2016). An international comparison of students' ability to endure fatigue and maintain motivation during a low-stakes test. *Learning and Individual Differences*, *49*, 128–137.

Brown, C. L., Kaur, S., Kingdon, G., & Schofield, H. (2022). *Cognitive endurance as human capital* (Tech. Rep.). National Bureau of Economic Research.

Cai, X., Lu, Y., Pan, J., & Zhong, S. (2019). Gender gap under pressure: evidence from China's National College entrance examination. *Review of Economics and Statistics*, *101*(2), 249–263.

Carey, E., Devine, A., Hill, F., & Szűcs, D. (2017). Differentiating anxiety forms and their role in academic performance from primary to secondary school. *PloS one*, *12*(3), e0174418.

Chang, T. Y., & Kajackaite, A. (2019). Battle for the thermostat: Gender and the effect of temperature on cognitive performance. *PloS one*, *14*(5), e0216362.

Cho, H., Glewwe, P., & Whitler, M. (2012). Do reductions in class size raise students' test scores? evidence from population variation in minnesota's elementary schools. *Economics of Education Review*, *31*(3), 77–95.

Cohen, A., Kricheli-Katz, T., Regev, T., Karelitz, T., & Pumpian, S. (2023). Gender-neutral language and gender disparities. *Available at SSRN*.

Duquennois, C. (2022). Fictional money, real costs: Impacts of financial salience on disadvantaged students. *American Economic Review*, *112*(3), 798–826.

Ebenstein, A., Lavy, V., & Roth, S. (2016). The long-run economic consequences of high-stakes examinations: Evidence from transitory variation in pollution. *American Economic Journal: Applied Economics*, *8*(4), 36–65.

Ellison, G., & Swanson, A. (2010). The gender gap in secondary school mathematics at high achievement levels: Evidence from the American Mathematics Competitions. *Journal of Economic Perspectives*, *24*(2), 109–128.

Falch, T., Nyhus, O. H., & Strøm, B. (2014). Causal effects of mathematics. *Labour Economics*, *31*, 174–187.

Feenberg, D., Ganguli, I., Gaule, P., & Gruber, J. (2017). It's good to be first: Order bias in reading and citing NBER working papers. *Review of Economics and Statistics*, *99*(1), 32–39.

Franco, C., & Hawkins, A. (2022). Strategic Decisions have "Major" Consequences: Gender Differences in College Major Choices.

Fryer Jr, R. G., & Levitt, S. D. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics*, *2*(2), 210–240.

Griselda, S. (2022). The Gender Gap in Math: What are We Measuring? *Available at SSRN 4022082*.

Hohensinn, C., & Baghaei, P. (2017). Does the position of response options in multiple-choice tests matter? *Psicológica*, *38*(1), 93–109.

Joensen, J. S., & Nielsen, H. S. (2016). Mathematics and gender: Heterogeneity in causes and consequences. *The Economic Journal*, *126*(593), 1129–1163.

Krueger, A. B. (1999). Experimental estimates of education production functions. *The Quarterly Journal of Economics*, *114*(2), 497–532.

Landaud, F., Maurin, É., Willage, B., & Willén, A. (2024). The Value of a High School GPA. *Review of Economics and Statistics*, 1–24.

Londoño-Vélez, J., Rodríguez, C., & Sánchez, F. (2020). Upstream and downstream impacts of college merit-based financial aid for low-income students: Ser Pilo Paga in Colombia. *American Economic Journal: Economic Policy*, *12*(2), 193–227.

OECD. (2015). *The ABC of Gender Equality in Education: Aptitude, Behaviour, Confidence.* http://dx.doi.org/10.1787/9789264229945-en.

Opatrny, M., Havranek, T., Irsova, Z., & Scasny, M. (2023). Class size and student achievement: A modern meta-analysis.

Park, R. J. (2022). Hot temperature and high-stakes performance. *Journal of Human Resources*, *57*(2), 400–434.

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

Sievertsen, H. H., Gino, F., & Piovesan, M. (2016). Cognitive fatigue influences students' performance on standardized tests. *Proceedings of the National Academy of Sciences*, *113*(10), 2621–2624.

10 Figures

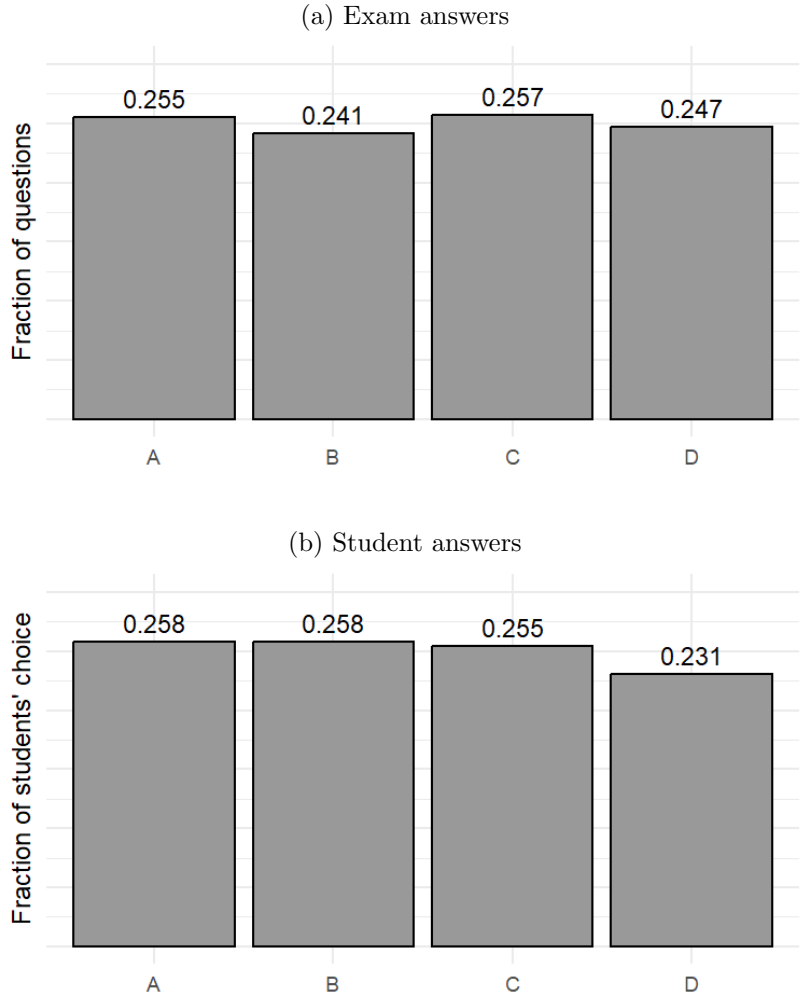


Figure 1: Fraction of exam answers and student answers in each option

Notes: Figure (a) describes the true fraction of questions with correct answers in each of the answer options: A, B, C and D using the randomization performed by the university. Figure (b) shows how often students chose A, B, C and D as their answer. The plots correspond to questions across the whole exam.

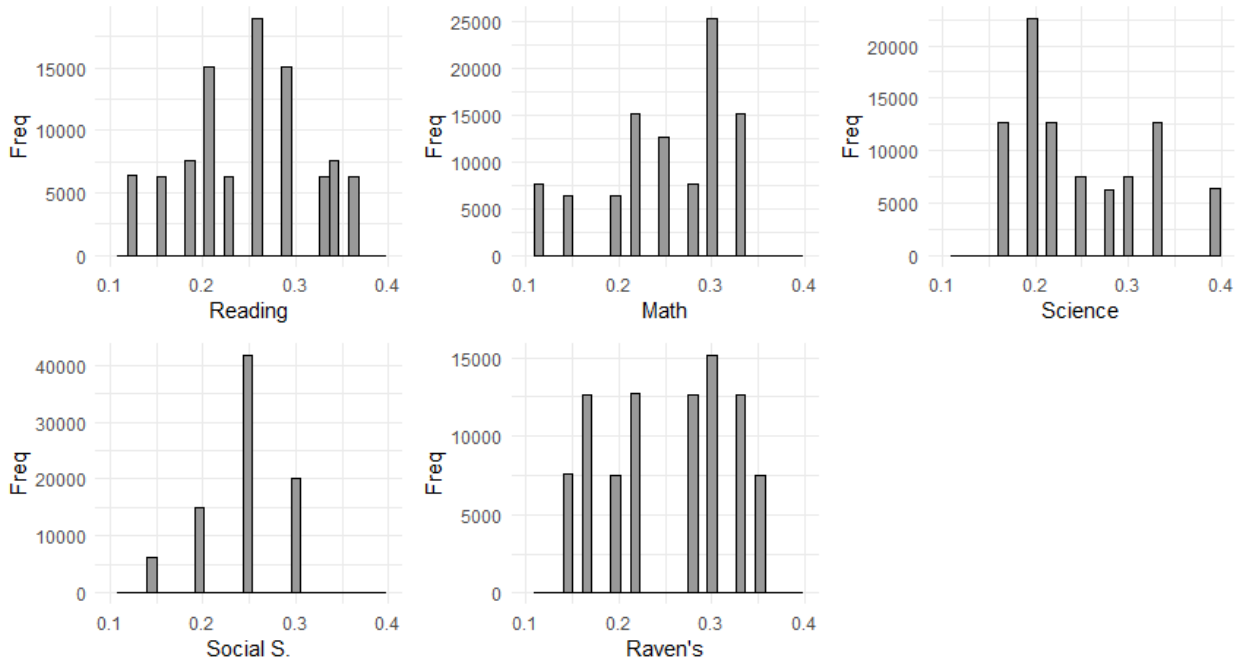
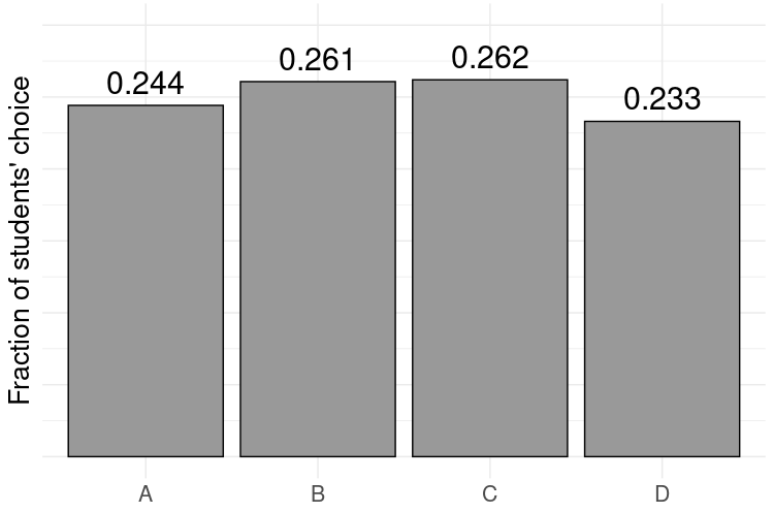


Figure 2: Fraction of correct answers in location D relative to total of questions in each subject

Notes: Each applicant is randomly assigned to a booklet (14 types) that contains the same structure of correct answer locations. By random chance, the fraction correct answers in location D varies by booklet. The figure shows the frequencies of students whose booklet has the proportion of correct answers in option D in the x axis. For example, in Reading, 15,000 students had a share of D equal to 0.2. The higher frequencies mean that students with different booklets have the same share of correct answers in D.

(a) Student answers at PISA



(b) Student answers at ENEM

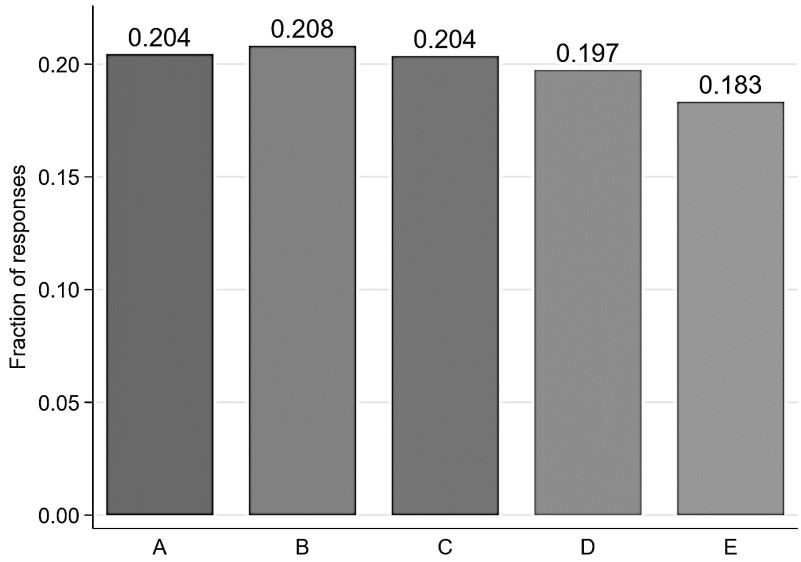


Figure 3: Fraction of student answers in each option

Notes: These figures show the frequency in which students selected options A, B, C, and D as their answers. Figure (a) encompasses responses from 606,627 students from 79 different countries who participated in the 2018 PISA exam. In Figure (b), the results are presented for Brazil's national college admission exam, known as ENEM. This examination significantly influences the academic choices of millions of high school students annually.

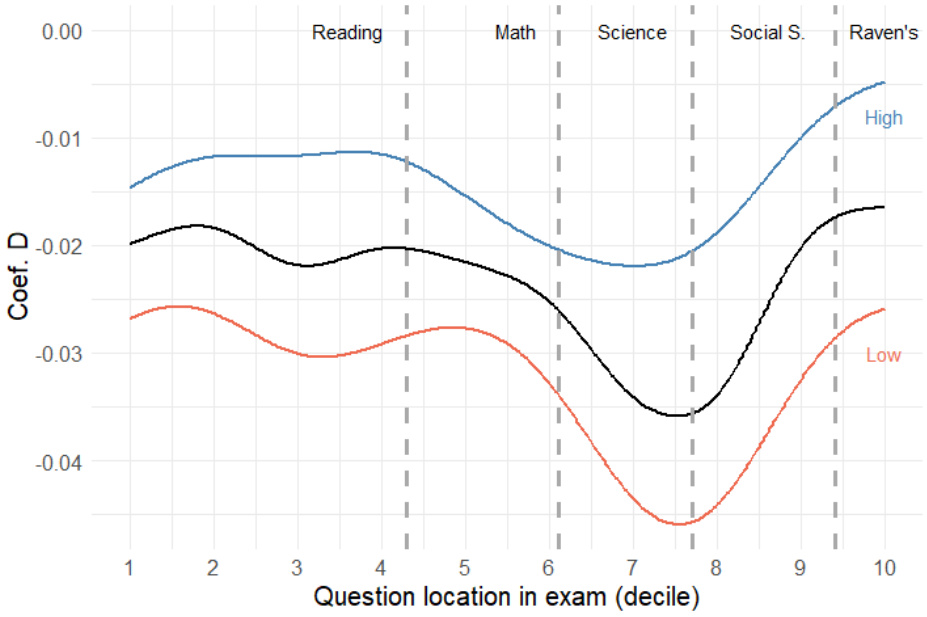


Figure 4: Share of D coefficient as exam progresses

Notes: The figure shows the effect of facing the correct answer in D (relative to A) on the likelihood of answering correctly obtained from Equation 1. Observations are at the student-question level. Instead of looking at the performance of the exam as a whole, we split the questions into question deciles to explore the effect of the answer location at different stages of the exam. The horizontal axis exhibits the question deciles (where 1 is the first decile of the test and 10 is the last decile). The vertical dashed lines help to visually identify every subject in the exam. The plots display the smoothed values of a kernel-weighted local polynomial regression, with a bandwidth of 0.6. The regression was run separately for the top and lowest-scoring students in the exam. Quantile 1 (lowest) appears in red, and quintile 5 (highest) is colored in blue. The black middle line includes the entire student population.

11 Tables

Table 1: Likelihood of correct answer

Dependent Variable:	Correct answer (=1)			
	(1)	(2)	(3)	(4)
Constant	0.4394***			
	(0.0111)			
B	0.0138	0.0139	-0.0029**	-0.0029**
	(0.0162)	(0.0159)	(0.0013)	(0.0013)
C	−0.0001	0.0002	-0.0037	-0.0037
	(0.0178)	(0.0178)	(0.0023)	(0.0023)
D	-0.0250*	-0.0253*	-0.0234***	-0.0234***
	(0.0126)	(0.0125)	(0.0014)	(0.0014)
Fixed-effects				
Student		✓	✓	✓
Question			✓	✓
Year				✓
N	11,070,660	11,070,660	11,070,660	11,070,660

Notes: Question order fixed across students, and the location of the correct answer randomized within the alternatives A, B, C and D in every question. The omitted category in the regressions corresponds to option A. Each applicant is randomly assigned to a booklet (14 types) that shares the structure of correct answer locations. The regression shows the probability of answering correctly depending on the location of the correct answer. Col (2) includes applicant fixed effects, Col (3) adds question fixed effects, and Col (4) include all fixed effects. Standard errors are clustered by booklet. $*p < 0.10, **p < 0.05, ***p < 0.01$.

Table 2: Likelihood of correct answer by quintile of performance

Dependent Variable:	Correct answer (=1)				
	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
	(1)	(2)	(3)	(4)	(5)
B	-0.0031**	-0.0027*	-0.0042*	-0.0026	-0.0020
	(0.0013)	(0.0013)	(0.0020)	(0.0015)	(0.0012)
C	-0.0060**	-0.0036	-0.0028	-0.0027	-0.0024
	(0.0024)	(0.0027)	(0.0024)	(0.0024)	(0.0021)
D	-0.0328***	-0.0260***	-0.0241***	-0.0193***	-0.0142***
	(0.0018)	(0.0017)	(0.0016)	(0.0017)	(0.0016)
Fixed-effects					
Student	✓	✓	✓	✓	✓
Question	✓	✓	✓	✓	✓
Year	✓	✓	✓	✓	✓
N	2,214,271	2,214,181	2,214,404	2,214,101	2,213,703

Notes: Question order fixed across students, and the location of the correct answer randomized within the alternatives A, B, C and D in every question. The omitted category in the regressions corresponds to option A. Each applicant is randomly assigned to a booklet (14 types) that shares the structure of correct answer locations. The regression shows the probability of answering correctly depending on the location of the correct answer. Cols (1)-(5) split students by quintile of their overall exam performance. Standard errors are clustered by booklet. $*p < 0.10, **p < 0.05, ***p < 0.01$.

Table 3: Effect of share D on subject performance

	(A) Reading	(B) Math	(C) Science	(D) Social	(E) Raven's
	(1)	(2)	(3)	(4)	(5)
share D (std)	0.0021	−0.0249***	−0.0097	−0.0086	−0.0047
	(0.0046)	(0.0041)	(0.0111)	(0.0063)	(0.0047)
Share D Mean	0.2500	0.2536	0.2401	0.2464	0.2429
Share D Sd	0.0693	0.0667	0.0787	0.0720	0.0785
Campus FEs	✓	✓	✓	✓	✓
Year FEs	✓	✓	✓	✓	✓
N	95,873	95,873	95,873	95,873	95,873

Notes: Observations at the student level. Question order fixed across students, and the location of the correct answer randomized within the alternatives A, B, C and D in every question. Each applicant is randomly assigned to a booklet (14 types) that shares the structure of correct answer locations. By random chance, some booklets have a higher fraction of correct answers in location D. "Share D" refers to the standardized proportion of correct answers in option D relative to the total number of questions in each subject. The regression estimates the impact of a higher share of D on the same subject's performance. In each column, the dependent variable is the exam subject standardized score. The mean and standard deviation used for standardizing the share D variable reported at the bottom of the table. All regressions include fixed effects for campus where applicants sought admission and year of testing. Standard errors are clustered by booklet. $*p < 0.10, **p < 0.05, ***p < 0.01$.

Table 4: Effect of share of D on total performance and likelihood of admission

	Total Score	1st Choice Admission
	(1)	(2)
Panel A: General		
share D (std)	−0.0039	−0.0012
	(0.0061)	(0.0014)
Panel B: By subject		
share D Reading (std)	0.0017	−0.0002
	(0.0050)	(0.0014)
share D Math (std)	−0.0100**	−0.0026**
	(0.0041)	(0.0011)
share D Science (std)	−0.0002	0.0008
	(0.0051)	(0.0013)
share D Social (std)	0.0019	−0.0002
	(0.0061)	(0.0013)
share D Raven's (std)	−0.0065	−0.0012
	(0.0049)	(0.0010)
Outcome mean	0.0000	0.0976
Campus, Year FEs	✓	✓
N	95,873	95,873

Notes: Observations at the student level. Question order fixed across students, and the location of the correct answer randomized within the alternatives A, B, C and D in every question. Each applicant is randomly assigned to a booklet (14 types) that shares the structure of correct answer locations. By random chance, some booklets have a higher fraction of correct answers in location D. "Share D" refers to the standardized proportion of correct answers in option D relative to the total number of questions in each subject. In Panel A the regression estimates the impact of a higher share of D computed by taking all questions in the exam. Each row in Panel B is a regression that estimates the share of D by exam subject. The dependent variable in Col (1) is the standardized measure of overall exam performance with mean 0 and SD 1. Col (2) is a binary indicator that equals one if the student was admitted to his/her first-choice major. All regressions include fixed effects for campus where applicants sought admission and year of testing. Standard errors are clustered by booklet. $*p < 0.10, **p < 0.05, ***p < 0.01$.

Table 5: Effect of share D in math on overall access to higher education institutions in Colombia

Dependent Variables:	University	Other University	Elite University	Technical School
	(1)	(2)	(3)	(4)
Panel A: Overall				
Share D Math (std)	-0.0018*	0.0003	-0.0029*	0.0002
	(0.0009)	(0.0011)	(0.0014)	(0.0012)
Panel B: Gender				
Share D Math (std)	-0.0054***	-0.0008	-0.0043	0.0014
	(0.0011)	(0.0015)	(0.0023)	(0.0013)
Share D Math (std) × Female	0.0071***	0.0017	0.0031	-0.0022
	(0.0011)	(0.0023)	(0.0021)	(0.0013)
Panel C: Low vs. Medium SES				
Share D Math (std)	-0.0027	-0.0012	-0.0016	0.0015
	(0.0024)	(0.0013)	(0.0021)	(0.0012)
Share D Math (std) × Low SES	0.0021	0.0029	-0.0019	-0.0026*
	(0.0038)	(0.0018)	(0.0026)	(0.0012)
Campus FEs	✓	✓	✓	✓
N	50,608	50,608	50,608	50,608

Notes: Observations at the student level. Question order fixed across students, and the location of the correct answer randomized within the alternatives A, B, C and D in every question. Each applicant is randomly assigned to a booklet (14 types) that shares the structure of correct answer locations. By random chance, some booklets have a higher fraction of correct answers in location D. "Share D" refers to the proportion of correct answers in option D relative to the total of questions in math. The regressions estimate the impact of a higher share D in math on the outcomes in the column headers using linked data to the universe of higher education attendees in Colombia up to two years after the 2020 exam. Col (1) measures whether the applicant is ever observed attending university, Col (2) whether the applicant ever attends a university other than U. Nacional, Col (3) whether the applicant ever attends one of the top-10 universities in Colombia, and Col (4) whether the applicant ever attends a technical/vocational higher education institution. All regressions include fixed effects for campus where applicants sought admission and year of testing. Standard errors are clustered by booklet. $*p < 0.10, * * p < 0.05, * * *p < 0.01$.

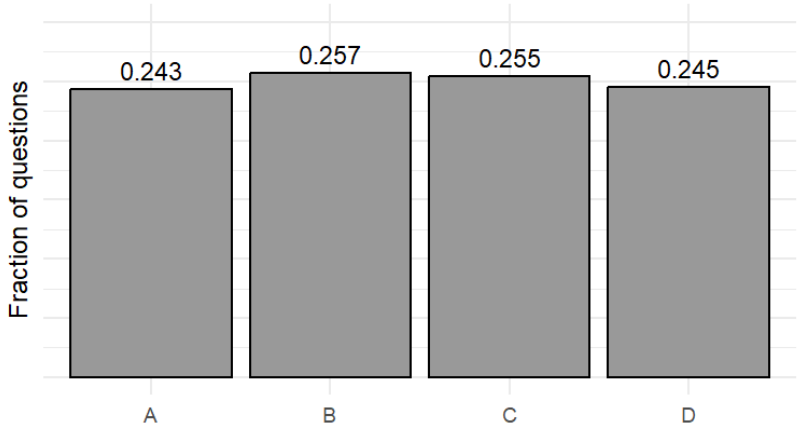
Table 6: Effect of share D in math on the total performance and likelihood of admission, heterogenenity effects of gender and socioeconomic status

	Total Score	1st Choice Admission
	(1)	(2)
Panel A: Gender		
share D Math (std)	−0.0125	−0.0014
	(0.0085)	(0.0017)
share D Math (std) × female	0.0052	−0.0024
	(0.0144)	(0.0020)
Panel B: Low vs. Medium SES		
share D Math (std)	−0.0076	−0.0029
	(0.0114)	(0.0018)
share D Math (std) × Low SES	−0.0025	0.0007
	(0.0180)	(0.0023)
Campus FEs	✓	✓
Year FEs	✓	✓
N	95,873	95,873

Notes: Observations at the student level. Question order fixed across students, and the location of the correct answer randomized within the alternatives A, B, C and D in every question. Each applicant is randomly assigned to a booklet (14 types) that shares the structure of correct answer locations. By random chance, some booklets have a higher fraction of correct answers in location D. "Share D" refers to the proportion of correct answers in option D relative to the total of questions in math. Panel A includes the interaction between the standardized share D of math and gender (female=1). Panel B includes the interaction between the standardized share D in math and socio-economic status (low SES=1). All regressions include fixed effects for campus where applicants sought admission and year of testing. Standard errors are clustered by booklet. $*p < 0.10, * * p < 0.05, * * *p < 0.01$.

A Appendix Figures

(a) Exam answers



(b) Student answers

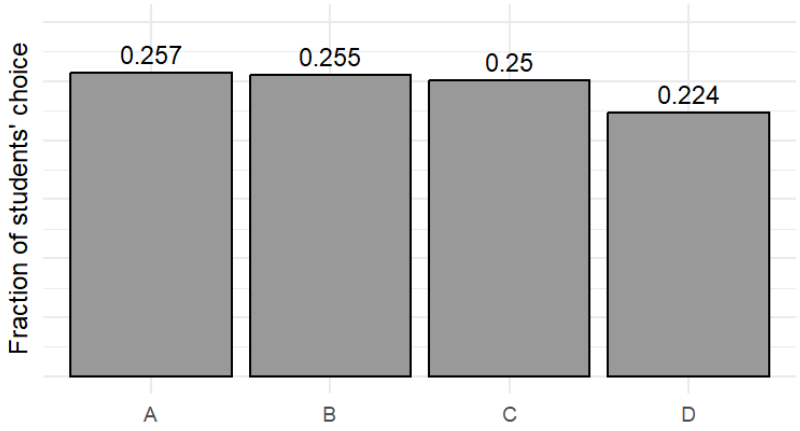
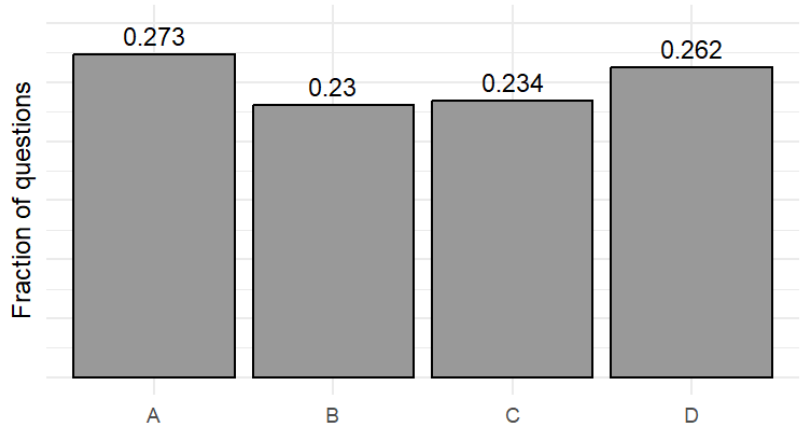


Figure A1: Fraction of exam answers and student answers in each option (Reading)

Notes: Figure (a) describes the true fraction of questions with correct answers in each of the answer options: A, B, C and D in Reading. Figure (b) shows how often students chose A, B, C and D as their answer. That is, each of the 95,873 students made 39 choices throughout this subject.

(a) Exam answers



(b) Student answers

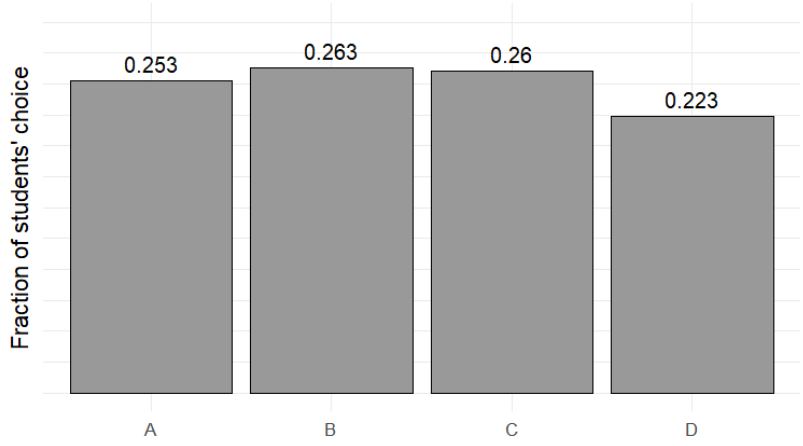
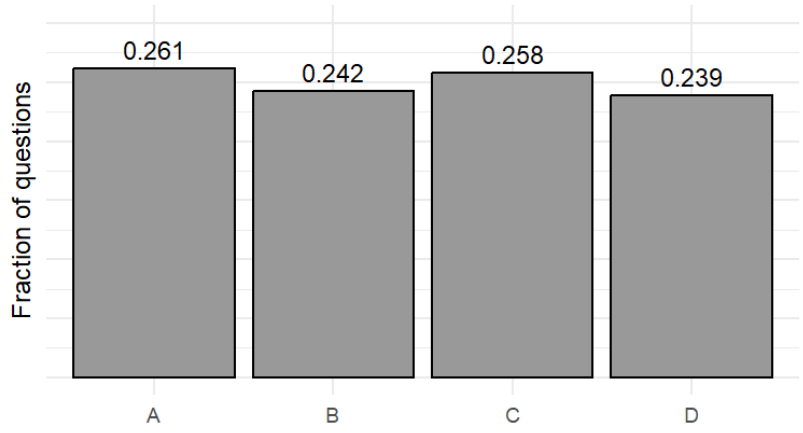


Figure A2: Fraction of exam answers and student answers in each option (Math)

Notes: Figure (a) describes the true fraction of questions with correct answers in each of the answer options: A, B, C and D in Math. Figure (b) shows how often students chose A, B, C and D as their answer. That is, each of the 95,873 students made 21 choices throughout this subject.

(a) Exam answers



(b) Student answers

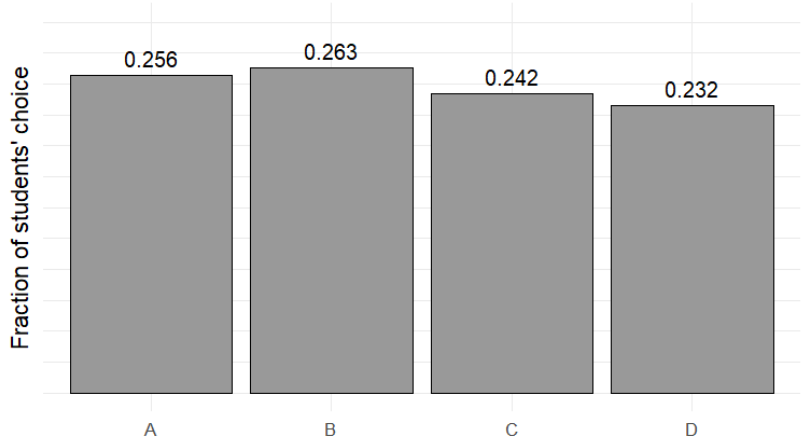
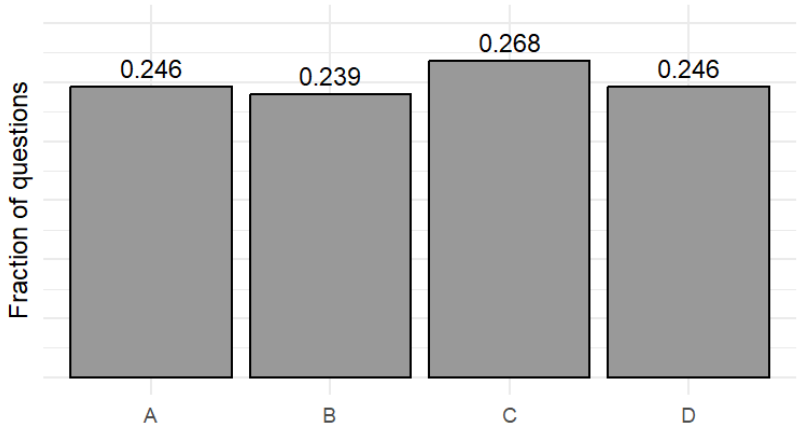


Figure A3: Fraction of exam answers and student answers in each option (Science)

Notes: Figure (a) describes the true fraction of questions with correct answers in each of the answer options: A, B, C and D in Science. Figure (b) shows how often students chose A, B, C and D as their answer. That is, each of the 95,873 students made 21 choices throughout this subject.

(a) Exam answers



(b) Student answers

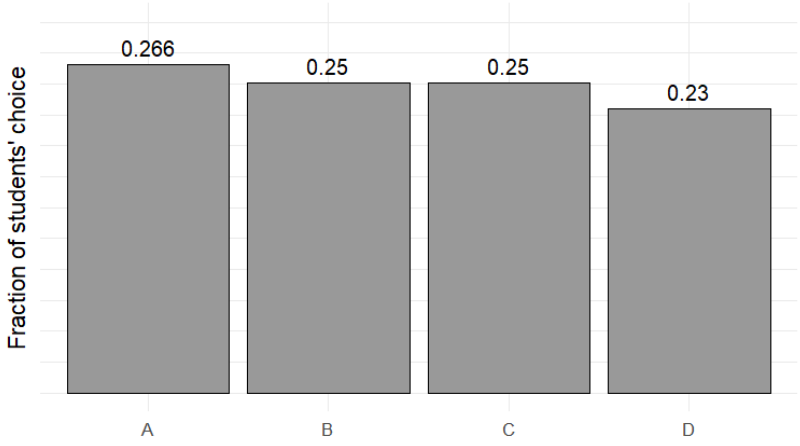
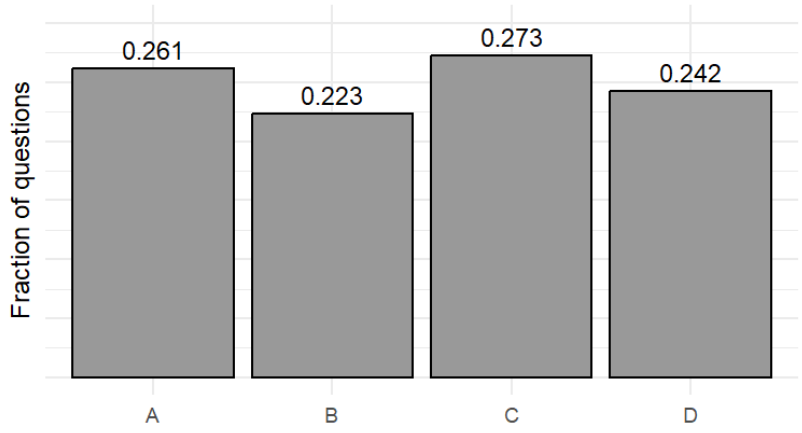


Figure A4: Fraction of exam answers and student answers in each option (Social Science)

Notes: Figure (a) describes the true fraction of questions with correct answers in each of the answer options: A, B, C and D in Social Science. Figure (b) shows how often students chose A, B, C and D as their answer. That is, each of the 95,873 students made 20 choices throughout this subject.

(a) Exam answers



(b) Student answers

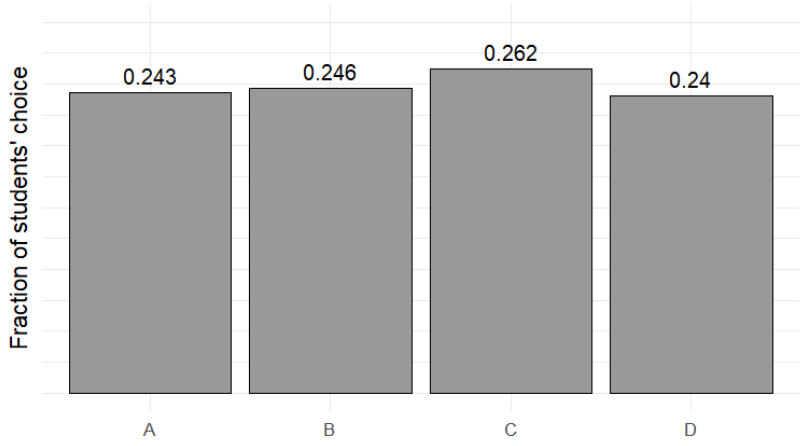


Figure A5: Fraction of exam answers and student answers in each option (Raven's)

Notes: Figure (a) describes the true fraction of questions with correct answers in each of the answer options: A, B, C and D in the Raven's test-type questions. Figure (b) shows how often students chose A, B, C and D as their answer. That is, each of the 95,873 students made 20 choices throughout this subject.

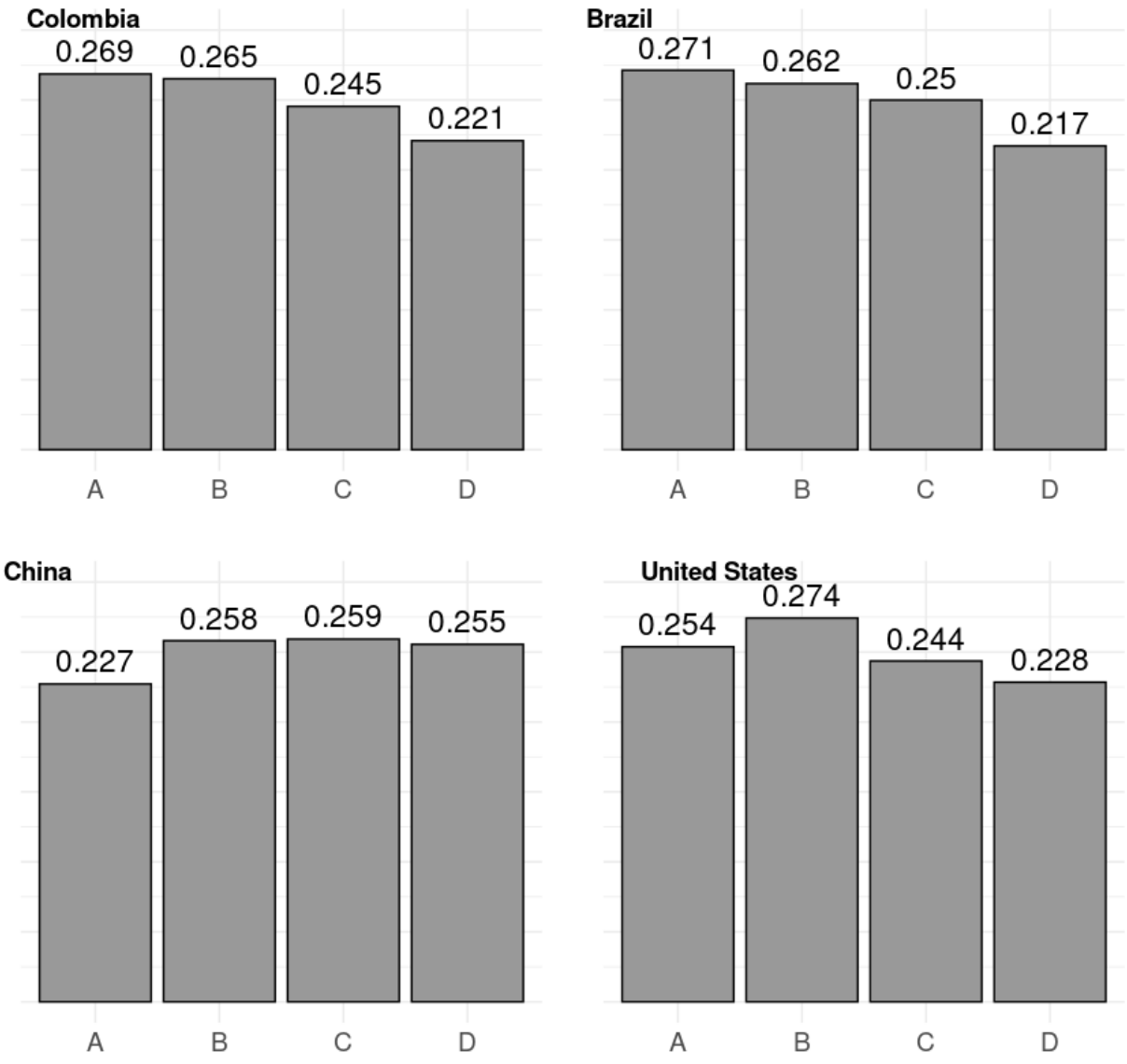
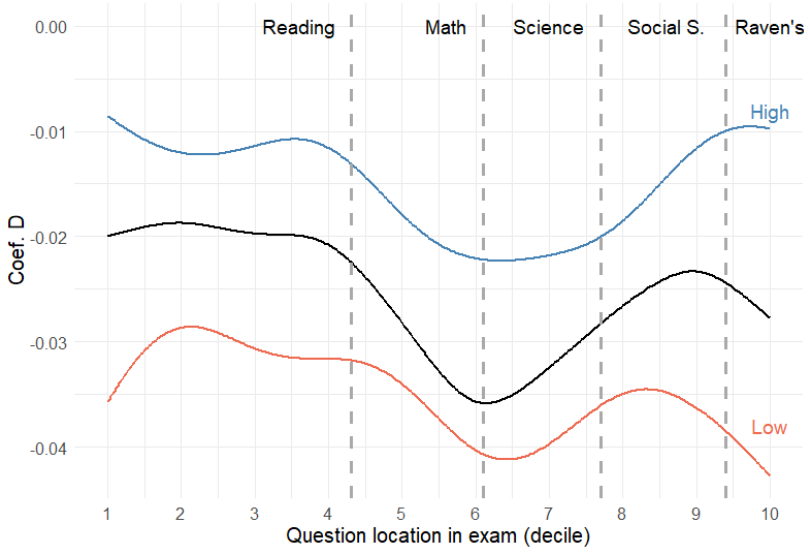


Figure A6: Fraction of student answers in each option (PISA by country)

Notes: The four panels present the fraction of times students in the PISA test chose A, B, C and D in four different countries. The sample of countries ranges from low scorers (Colombia and Brazil), medium scorers (The United States) an high scorers (China) based on the performance in the 2018 PISA exam.

(a) D effect 2020



(b) D effect 2023

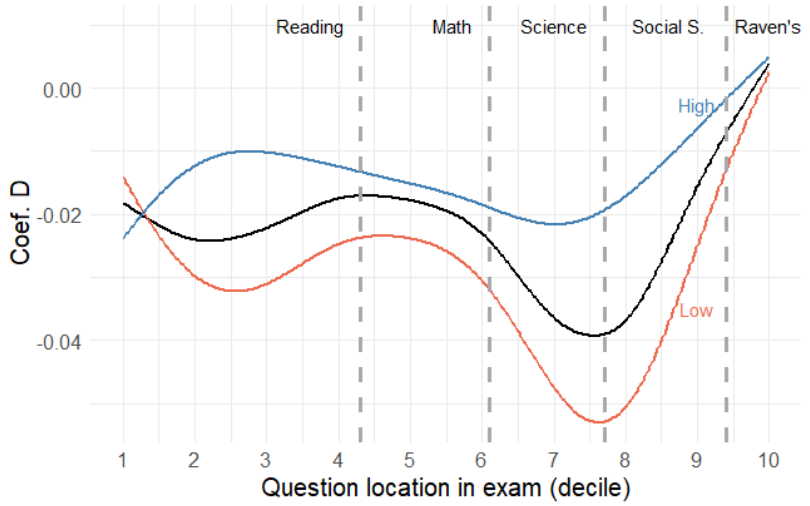


Figure A7: Share D coefficient as exam progresses

Notes: The figure shows the effect of facing the correct answer in D (relative to A) on the likelihood of answering correctly obtained from Equation 1 for students in the 2020 dataset (Panel (a)) and in the 2023 dataset (Panel (b)). Observations are at the student-question level. Instead of looking at the performance of the exam as a whole, we split the questions into deciles to explore the effect of the answer location at different stages of the exam. The horizontal axis exhibits the question deciles (where 1 is the first decile of the test and 10 is the last decile). The vertical dashed lines help to visually identify every subject in the exam. The plots display the smoothed values of a kernel-weighted local polynomial regression, with a bandwidth of 0.6. The regression was run separately for the top and lowest-scoring students in the exam. Quantile 1 (lowest) appears in red, and quintile 5 (highest) is colored in blue. The black middle line includes the entire student population.

B Appendix Tables

Table A1: Likelihood of correct answer in question following a D D vs. a C C sequence

	(1)	(2)
D D vs. C C	0.027	0.027
	(0.154)	(0.143)
Answer to next question in A		-0.063
		(0.124)
Constant	0.304**	0.332**
	(0.085)	(0.097)
N	59703	59703

Notes: The outcome is the probability that the next question after a sequence of DD vs. CC is correct. Because the correct answers are well mixed, there are no sequences longer than two equal letters in a row, and the only correct answers after CC or DD are either A or B. Column one shows the point estimates without controlling for the letter of the correct answer option after the sequence, while column 2 adds an indicator for whether the letter is A. If a sequence of correct answers in D would generate cognitive fatigue immediately, we would expect to see the D D vs. C C coefficient to be negative and significant. $*p < 0.10, * * p < 0.05, * * *p < 0.01$.

Table A2: Likelihood of not responding

Dependent Variable:	Non-response (=1)			
	(1)	(2)	(3)	(4)
Constant	0.0035***			
	(0.00004)			
B	-0.0003***	-0.0003***	-0.00004	-0.00003
	(0.00005)	(0.00004)	(0.00005)	(0.00005)
C	0.0002***	0.0001***	-0.0001	-0.0001
	(0.00005)	(0.00005)	(0.00005)	(0.00005)
D	-0.0002***	-0.0002***	-0.00003	-0.00003
	(0.00005)	(0.00004)	(0.00005)	(0.00005)
Fixed-effects				
Student		✓		✓
Question			✓	✓
N	11,070,660	11,070,660	11,070,660	11,070,660

Notes: Question order fixed across students, and the location of the correct answer randomized within the alternatives A, B, C and D in the multiple-choice set. The omitted category corresponds to letter A. Each applicant is randomly assigned to a booklet (14 types) that shares the structure of correct answer locations. The regression shows the probability of leaving a question empty (no response) depending on the location of the correct answer. See the bottom panel to see whether student or question fixed effects are added. Standard errors are clustered by booklet. $*p < 0.10, * * p < 0.05, * * *p < 0.01$.

Table A3: Likelihood of correct answer, STEM applicants

Dependent Variable:	Correct answer (=1)	
	All exam	Math
B	−0.0058***	0.0241***
	(0.0011)	(0.0031)
C	−0.0065***	0.0194 ***
	(0.0017)	(0.0049)
D	−0.0268***	−0.0291***
	(0.0021)	(0.0031)
Applied to Stem major × B	0.0017	−0.0015
	(0.0016)	(0.0069)
Applied to Stem major × C	−0.0025	0.0030
	(0.0021)	(0.0090)
Applied to Stem major × D	0.0046***	0.0144***
	(0.0012)	(0.0037)
Fixed-effects		
Student	✓	✓
Question	✓	✓
N	1,012,160	1,012,160

Notes: Question order fixed across students, and the location of the correct answer randomized within the alternatives A, B, C and D in the multiple-choice set. The omitted category corresponds to letter A. Each applicant is randomly assigned to a booklet (14 types) that shares the structure of correct answer locations. The regression shows the probability of answering correctly depending on the location of the correct asnwer. The interaction term assesses whether applicants intending STEM majors are impacted differently by the answer location. All columns include student and question fixed effects and the standard errors are clustered by booklet. $*p < 0.10, **p < 0.05, ***p < 0.01$.

Table A4: Effect of share of choice alternatives on subject performance

	(A) Reading	(B) Math	(C) Science	(D) Social	(E) Raven's
	(1)	(2)	(3)	(4)	(5)
share B (std)	−0.0071	−0.0024	−0.0059	−0.0028	0.0040
	(0.0054)	(0.0094)	(0.0080)	(0.0107)	(0.0077)
share C (std)	−0.0244**	0.0063	0.0181*	−0.0016	0.0147*
	(0.0090)	(0.0095)	(0.0089)	(0.0060)	(0.0075)
share D (std)	−0.0213*	−0.0232***	−0.0021	−0.0105	0.0002
	(0.0112)	(0.0076)	(0.0074)	(0.0084)	(0.0081)
Share B Mean	0.2576	0.2266	0.2429	0.2393	0.2234
Share B Sd	0.0477	0.0838	0.0925	0.0764	0.1224
Share C Mean	0.2499	0.2429	0.2575	0.2679	0.2698
Share C Sd	0.0627	0.0726	0.0930	0.0846	0.0936
Share D Mean	0.2500	0.2536	0.2401	0.2464	0.2429
Share D Sd	0.0693	0.0667	0.0787	0.0720	0.0785
Campus FEs	✓	✓	✓	✓	✓
Year FEs	✓	✓	✓	✓	✓
N	95,873	95,873	95,873	95,873	95,873

Notes: Observations at the student level. Question order fixed across students, and the location of the correct answer randomized within the alternatives A, B, C and D in every question. Each applicant is randomly assigned to a booklet (14 types) that shares the structure of correct answer locations. By random chance, some booklets have a higher fraction of correct answers in some locations, for instance in alternative B. "Share B" refers to the proportion of correct answers in option B relative to the total of questions in each subject. The mean and standard deviation used for standardizing the share of B, C and D variables are reported at the bottom of the table. All regressions include fixed effects for campus where applicants sought admission and year of testing. Standard errors are clustered by booklet. $*p < 0.10, * * p < 0.05, * * *p < 0.01$.

Table A5: Effect of shares of multiple-choice options on total performance and likelihood of admission

	Total Score	1st Choice Admission	University
	(1)	(2)	(3)
share B Math (std)	0.0007	−0.0024	−0.0001
	(0.0083)	(0.0017)	(0.0019)
share C Math (std)	0.0039	−0.0015	0.0004
	(0.0070)	(0.0014)	(0.0020)
share D Math (std)	−0.0082	−0.0040**	−0.0021*
	(0.0080)	(0.0014)	(0.0011)
Campus FEs	✓	✓	✓
Year FEs	✓	✓	
N	95,873	95,873	50,608

Notes: Observations at the student level. Question order fixed across students, and the location of the correct answer randomized within the alternatives A, B, C and D in every question. Each applicant is randomly assigned to a booklet (14 types) that shares the structure of correct answer locations. By random chance, some booklets have a higher fraction of correct answers in some locations. "Share i" refers to the proportion of correct answers in option i relative to the total of questions in subject i. The mean and standard deviation used for standardizing the share of B, C and D variables are reported at the bottom of the table. The dependent variable in Col (1) is the standardized measure of overall exam performance with mean 0 and SD 1. Col (2) is a binary indicator that equals one if the student was admitted to his/her first-choice major. Col (3) is a binary indicator that equals one if within two years of taking the CEE, the student attends any higher education intuition. All regressions include fixed effects for campus where applicants sought admission and year of testing, except for Col (3) in which we consider the 2020 cohort only. Standard errors are clustered by booklet. $*p < 0.10, **p < 0.05, ***p < 0.01$.

Table A6: Effect of shares D of all subjects on total performance and likelihood of admission

	Total Score	1st Choice Admission	University
	(1)	(2)	(3)
share D Reading (std)	−0.0022	−0.0011	0.0022
	(0.0029)	(0.0009)	(0.0016)
share D Math (std)	−0.0124***	−0.0030***	−0.0031***
	(0.0034)	(0.0010)	(0.0005)
share D Science (std)	−0.0089*	−0.0004	−0.0006
	(0.0044)	(0.0011)	(0.0012)
share D Social (std)	0.0089*	0.0006	0.0015
	(0.0042)	(0.0010)	(0.0010)
share D Raven's (std)	−0.0066	−0.0005	−0.0013
	(0.0061)	(0.0013)	(0.0022)
Campus FEs	✓	✓	✓
Year FEs	✓	✓	
N	95,873	95,873	50,608

Notes: Observations at the student level. Question order fixed across students, and the location of the correct answer randomized within the alternatives A, B, C and D in every question. Each applicant is randomly assigned to a booklet (14 types) that shares the structure of correct answer locations. By random chance, some booklets have a higher fraction of correct answers in location D. "Share D" refers to the standardized proportion of correct answers in option D relative to the total of questions in each exam subject. The dependent variable in Col (1) is the standardized measure of overall exam performance with mean 0 and SD 1. Col (2) is a binary indicator that equals one if the student was admitted to his/her first-choice major. Col (3) is a binary indicator that equals one if within two years of taking the CEE, the student attends any higher education intuition. All regressions include fixed effects for campus where applicants sought admission and year of testing, except for Col (3) in which we consider the 2020 cohort only. Standard errors are clustered by booklet. $*p < 0.10, **p < 0.05, ***p < 0.01$.