Cognitive Endurance, Talent Selection, and the Labor Market Returns to Human Capital

Germán Reyes*

March 2024

Abstract

Cognitive endurance—the capacity to sustain performance on a cognitively-demanding task over time—is thought to be a crucial productivity determinant. However, a lack of data on this variable has limited researchers' ability to understand its role for success in college and the labor market. This paper uses college-admission-exam records from 15 million Brazilian high-school students to measure cognitive endurance based on changes in performance during the exam. By exploiting exogenous variation in the order of exam questions, I first show that students are significantly more likely to correctly answer a given question when it appears at the beginning of the test versus the end. Motivated by this fact, I develop a method to decompose test scores into fatigue-adjusted ability and cognitive endurance. I then link these measures to college and employment records to quantify the association between endurance and long-run outcomes. I find that cognitive endurance has a significant wage return. Controlling for fatigue-adjusted ability and other student characteristics, an increase of one standard deviation in endurance predicts a 5.4% wage increase. This wage return is equivalent to a third of the wage return to fatigue-adjusted ability. I also document positive associations between endurance and college attendance, college graduation, firm quality, and other outcomes. Finally, I show that, due to systematic differences in endurance among students, the exam design can impact income-based test-score gaps and the informational content of the exam. I discuss the implications of these findings for designing more informative cognitive assessments to select talent and more effective interventions to build human capital.

*Middlebury College (e-mail: greyes@middlebury.edu). I thank especially my advisor Ted O'Donoghue for invaluable guidance. For helpful discussions and comments, I thank Ned Augenblick, Michèle Belot, Nicolas Bottan, Emily Breza, Aviv Caspi, Neel Datta, Stefano DellaVigna, Josh Dean, Christa Deneault, Rebecca Deranian, Gary Fields, Thomas Graeber, Ori Heffetz, David Huffman, Guy Ishai, Judd Kessler, Yizhou Kuang, Shengwu Li, Yucheng Liang, George Loewenstein, Michael Lovenheim, Suraj Malladi, Alejandro Martínez-Marquina, Francesca Molinari, Kevin Ng, Muriel Niederle, Ryan Oprea, Ricardo Perez-Truglia, Grace Phillips, Alex Rees-Jones, Evan Riehl, Jonah Rockoff, Seth Sanders, Paola Sapienza, Frank Schilbach, Heather Schofield, Peter Schwardmann, Dmitry Taubinsky, Emanuel Vespa, participants in the Cornell behavioral economics group, participants in the UC Berkeley psychology and economics group, and numerous seminar participants. I also thank Marco Pereira and other members of SEDAP for invaluable help using the secured data room. Special thanks to Elisa Taveras Pena and David Slichter for sharing the O*NET CBO crosswalks. Financial support from the National Science Foundation is gratefully acknowledged.

1 Introduction

The human capital framework posits that individuals' skills and knowledge act as a form of capital that improves productivity and, thus, labor earnings (Becker, 1962). The positive relationship between human capital and earnings is one of the most robust findings in the social sciences (Deming, 2022), and is supported by a large body of work (e.g., Mincer, 1958; Griliches, 1977; Card, 1999, 2001). While early studies focused on aggregate measures of human capital—like years of schooling—more recent work has focused on estimating the economic returns to specific skills, such as social skills (Deming, 2017) and cognitive skills (Hermo et al., 2022). Identifying skills that foster productivity is essential for the design of effective education and labor-market policies (Almlund et al., 2011; Kautz et al., 2014).

In this paper, I study one dimension of human capital that may be particularly important for knowledge workers: *cognitive endurance*, that is, the ability to sustain performance on a cognitively-demanding task for an extended duration. I first document that the performance of individuals on a college admission exam tends to decline, which allows me to measure cognitive endurance. I develop a method to decompose test scores into fatigue-adjusted ability and endurance. I use the decomposition to investigate the relationship between endurance and long-run outcomes. I show that endurance has a sizable wage return in the labor market, comparable to the wage return to ability. I also show that, due to systematic differences in endurance across students, seemingly neutral exam design choices, such as the exam length, can affect the test-score gap between different types of students and the predictive power of test scores.

Psychologists and self-help books have long hypothesized that cognitive endurance is an important productivity determinant. Research on the nature of expertise—popularized in influential books like *Focus* (Goleman, 2013) or *Deep Work* (Newport, 2016)—often identifies this skill as a key driver of performance on cognitive tasks. Relatedly, biographers of extraordinary achievers often ascribe their accomplishments to unusually-high endurance.[1] Consistent with this, researchers have documented the negative consequences of limited endurance for task performance in many settings.[2] The hypothesized link be-

[1] For example, in describing Newton's accomplishments, Keynes (1956) noted that his greatest skill was "the power of holding continuously in his mind a purely mental problem until he had seen straight through it." See Lykken (2005) for many other examples.

[2] Research shows that individual-level job performance tends to deteriorate over relatively short time spans. For example, over the course of a day: nurses are less likely to wash their hands (Dai et al., 2015; Steiny Wellsjo, 2022); doctors make more diagnostic mistakes (Chan et al., 2009; Linder et al., 2014; Kim et al., 2018); financial analysts make less accurate forecasts (Hirshleifer et al., 2019); and umpires make

tween endurance and productivity is also consistent with the large markets for endurance enhancers like coffee or nootropics (like Adderall or Ritalin).[3]

These observations suggest that cognitive endurance and task performance are intimately linked. Yet, despite this popular perception, empirical economists have had little to say about the role of endurance in the labor market, possibly because of a lack of data on this variable. I address this problem by using data from the national college admission exam in Brazil (called "ENEM") to create an individual-level measure of endurance that is based on performance declines throughout the exam (Borghans and Schils, 2018; Brown et al., 2022).

The ENEM is an ideal setting to study cognitive endurance for several reasons. First, the exam is administered under uniform conditions, and the scoring is standardized—two crucial properties for generating measures that are comparable across individuals (Almlund et al., 2011). Second, it is a high-stakes environment. Test scores largely determine the college options of the millions of high school students who take the ENEM every year. Since test-takers have incentives to exert maximal effort, limits to cognitive endurance are more likely to drive systematic declines in performance rather than low motivation (Duckworth et al., 2011; Gneezy et al., 2019). Third, the exam is grueling. The ENEM is ten hours long and is conducted over two consecutive days of testing. Thus, we might expect cognitive endurance to be an especially valuable skill in this setting and cross-person differences in endurance to be reflected in test performance.

My analysis takes advantage of three features of the ENEM. First, the dataset contains students' responses to each exam question, which enables me to measure student performance throughout the exam. Second, students are randomly assigned different test booklets. Each booklet has the same set of questions (or "items") but in a different order, which enables me to study how students perform on a given question when they are relatively "fresh" versus mentally fatigued. Third, the ENEM can be linked to other administrative datasets to measure students' long-run outcomes. I link the ENEM records to a census of all Brazilian college students and an employee-employer matched dataset that covers the universe of formal-sector workers in Brazil.

I measure cognitive endurance as the impact of a one-position increase in the order of a

more incorrect calls in baseball games (Archsmith et al., 2021).

[3]For example, in the US, 65% of adults drink coffee daily (Lampkin, 2012), and about 20% of college students report using nootropics without a prescription to enhance focus and cognition (Benson et al., 2015). Relatedly, over-the-counter focus-enhancing drugs have entire sections in chain drug stores, and there is a growing variety of products marketed as endurance training (e.g., brain-training games like "Lumosity" or interval-based training technologies like "Pomodoros").

given question on the likelihood of correctly answering the question. A potential-outcomes framework reveals that this measure captures the combined impact of two structural parameters: how cognitively fatigued an individual becomes throughout the exam and how an increase in cognitive fatigue affects test performance. These two parameters, and thus, my endurance measure, likely capture a variety of psychological mechanisms, including intrinsic motivation, grit, and attention capacity. One of the limitations of this paper is my inability to distinguish among various underpinning mechanisms that shape cognitive endurance.

Applying this framework, I first estimate *average* cognitive endurance using two empirical strategies. The first research design compares average student performance *on a given question* as a function of its position on each booklet, which I implement by regressing the fraction of students who correctly answer a question on its position on the exam, controlling for question fixed effects. This approach provides the more credible estimates of average cognitive endurance; however, since each student only receives one exam booklet, it cannot be used to estimate *individual-level* endurance. Thus, I also use a second research design that can be used to identify both average and individual-level endurance. The second approach consists of creating a position-adjusted measure of question difficulty, and then using this measure as a control variable instead of the question fixed effects. Both strategies deliver a similar-sized estimate of average cognitive endurance. A one-position increase in the order of a given question decreases the likelihood of correctly answering the question by 0.08 percentage points. Scaled by the number of questions per testing day, this estimate implies that daily performance decreases by 7.1 percentage points due to limited endurance (relative to a sample mean of 34.3%).

Next, I estimate the difficulty-adjusted regression separately for each individual. This allows me to decompose an individual's test score into a measure of cognitive endurance and a measure of fatigue-adjusted academic ability. My measure of cognitive endurance is the same as above but now estimated separately for each student. My measure of fatigue-adjusted ability is the residual of an individual's test score after subtracting from it the component explained by cognitive endurance. Using a sample of students who took the exam multiple times, I show that this measure of endurance has a test-retest reliability comparable to that of other commonly used constructs like risk aversion (Mata et al., 2018) or teacher value-added (Chetty et al., 2014a).

I first use the measures generated by the decomposition to investigate the importance of cognitive endurance for success in college and the labor market. I find that, holding

fixed fatigue-adjusted ability and other student characteristics, individuals with more cognitive endurance are more likely to attend college, enroll in higher-quality colleges, are more likely to graduate, earn higher wages, and work for higher-paying firms. The magnitudes of these associations are sizable. For example, ceteris paribus, a one standard-deviation (SD) increase in cognitive endurance predicts a 5.4% increase in early-career wages. The corresponding prediction for a one SD increase in fatigue-adjusted ability equals 15.4%. Hence, the wage return to endurance is about a third of the size of the return to fatigue-adjusted ability. Instrumental variable regressions show that the association between endurance and wages is larger after accounting for measurement error (on the order of 70% the size of the return to ability) and also reveal that the predicted effect is not driven by a mechanical relationship between endurance and test scores.

Then, I assess whether the value of endurance varies across jobs by estimating the wage return to ability and to endurance across college majors, occupations, and industries. On average, occupations and industries that pay higher wages also offer a higher wage return to ability and to endurance, suggesting a novel type of assortative matching between high-endurance workers and high-paying jobs. Furthermore, occupations and industries with a high wage return to endurance also tend to have a high wage return to ability, suggesting these two skills are complements in production. Some occupations with the highest wage return to endurance include those where lapses in sustained attention can have high production costs, like professionals in the aviation industry or facility operators in chemical plants. This suggests that the capacity to sustain focus on a task for a long time may be a psychological mechanism contributing to the reduced-form measure of cognitive endurance.

When two skills, ability and endurance, are combined into a single index (the test score) the market faces muddled information (Frankel and Kartik, 2019). Even though admission officers (or employers) would want to evaluate individuals mostly based on the skill that is predominantly needed to succeed, the information revealed about this skill is "contaminated" by less-relevant information about the other skill. Importantly, the extent to which the test score reveals information about ability vis-à-vis endurance depends on the exam design. Intuitively, not much endurance is required to perform well on a short exam.

In the final part of the paper, I focus on identifying the *distributional* and *informational* effects of an exam design that reveals more information about ability (and less about endurance). The distributional effect asks how the exam design impacts socioeconomic

status (SES) test-score gaps, an important determinant of inequity in college access. The informational effect asks how the exam design impacts the information content of the exam, an important determinant of the student-college match quality. I measure the exam information content by the correlation between test responses and long-run outcomes (or "predictive validity," for short). I quantify these effects by simulating the consequences of an exam reform that decreases the exam length by half, thereby reducing the importance of endurance for exam performance.

The exam reform would decrease test-score gaps by 1.3–4.8 percentage points (a 26%–29% reduction from pre-reform gaps, depending on the measure of SES) and increase the predictive validity of test responses for long-run outcomes by as much as 95%. Intuitively, the reform would reduce test-score gaps because, conditional on academic ability, low-SES students have lower endurance than high-SES students and, thus, perform disproportionally worse in questions toward the end of the exam. Similarly, the reform would increase the predictive validity of the exam partly because differences in performance at the beginning of the exam mainly reflect differences in ability (roughly, because most students are "fresh"), which are highly predictive of long-run outcomes. In contrast, performance differences towards the end of the exam also reflect the noise associated with mental fatigue, which reduces the information content of test responses.

My findings yield three broad lessons. First, cognitive endurance matters for success in college and the labor market. Thus, investing in the development of this skill, possibly during early ages, may have significant societal returns. Second, distinguishing between endurance and ability can improve how talent is selected and trained. Since the value of endurance varies among college majors, the student-major match may improve if majors where high endurance is required to succeed screen applicants partly based on this skill. Similarly, workers in endurance-intensive occupations may be more productive if the training necessary to enter into these occupations includes components aimed at building this skill. Third, seemingly neutral exam design decisions—the "choice architecture" of the exam—such as length or number of breaks, can have important consequences. By influencing the importance of endurance for test performance, the exam design can affect test-score gaps and predictive validity and, thus, the diversity of colleges' student bodies and the student-college match quality.

This paper relates to the literature that studies cognitive endurance and fatigue effects in field settings. Limited endurance and fatigue effects have been documented in a wide variety of environments (see footnote 2). Recent experimental evidence shows that cognitive

endurance can be trained in children, which leads to less pronounced performance declines (Brown et al., 2022). I contribute by linking individual-level endurance to long-run outcomes and establishing a novel set of associations. I do this in a high-stakes exam, which complements previous studies documenting performance declines in the low-stakes PISA test (e.g., Debeer et al., 2014; Borghans and Schils, 2018; Zamarro et al., 2018; Balart and Oosterveen, 2019). My findings provide a micro perspective to the results of Balart et al. (2018), who show that the average performance decline in the PISA test among a country's test-takers has a sizable predictive power in cross-country growth regressions.

This paper also contributes to a growing literature documenting the importance of different dimensions of human capital for long-run outcomes. A large body of work shows that cognitive skills are valuable in the labor market (e.g., Hanushek and Woessmann, 2008, 2012; Fe et al., 2022; Hermo et al., 2022). This work often uses test scores as a measure of cognitive skills. I show that, even in a high-stakes setting, test scores partly capture cognitive endurance and provide methods to decompose test scores into fatigue-adjusted ability and endurance. Relatedly, a growing body of work shows that skills other than intelligence and technical skills ("noncognitive skills") are also important predictors of long-run outcomes (Bowles et al., 2001; Heckman et al., 2006; Borghans et al., 2008; Almlund et al., 2011; Lindqvist and Vestman, 2011; Deming, 2017; Jackson, 2018; Buser et al., 2021; Edin et al., 2022). I document the strong predictive power of one noncognitive skill (endurance) for long-run outcomes and study how it relates to a measure of cognitive skills (fatigued-adjusted ability) in the labor market.

Finally, this paper contributes to the literature on the design of college admission exams (Rothstein, 2004; Ackerman and Kanfer, 2009; Bettinger et al., 2013; Hoxby et al., 2013; Bulman, 2015; Goodman, 2016; Goodman et al., 2020; Riehl, 2022). These exams are designed to rank a large number of applicants. This requires discerning small ability differences, and as a consequence, they tend to be long and arduous. I show that performance on college admission exams measures not only academic preparedness but also the capacity to endure mental fatigue. Hence, there is a limit to how much information an exam can extract about student academic achievement. A lengthier exam may not lead to more precise measures of ability but rather to a selection mechanism that puts more weight on endurance. This may be desirable for programs where endurance is crucial to succeed, but it may come at the cost of screening out high-ability low-endurance students.

The rest of the paper is structured as follows. Section 2 describes the context and data. Section 3 presents a statistical framework and describes my research design. Section

4 presents estimates of average cognitive endurance. Section 5 decomposes test scores into fatigue-adjusted ability and cognitive endurance. Section 6 examines the relationship between cognitive endurance and long-run outcomes. Section 7 studies the informational and distributional effects of a shorter exam. Section 8 concludes.

2 Institutional Context and Data

2.1 The ENEM exam

The High School Assessment Exam (*Exame Nacional do Ensino Médio*, or ENEM for short) is an achievement test created in 1998 by the Brazilian Ministry of Education to make high schools accountable for their students' progress. Some universities used the ENEM for college admissions; however, most institutions had university-specific admission exams. In 2009, the Ministry of Education expanded the ENEM to encourage universities to use it as their admission exam, and created a centralized admission system that uses ENEM scores to assign students to the highly-selective federal universities. Since then, many universities have started using the ENEM for admissions (Machado and Szerman, 2021; Otero et al., 2021).

The ENEM contains 180 multiple-choice questions equally divided across four subject tests (language arts, math, natural sciences, and social sciences) and an essay. The exam takes place over two consecutive days (two subjects per day, plus the essay on the second day). Test-takers have four and a half hours to complete the test on the first day and five and a half hours on the second day. There are no allocated breaks. To combat cheating, examinees randomly receive one of four different booklets each day. The order of the subjects and the set of questions is the same across booklets, but the order of the questions within a subject is randomized across booklets. A score for each subject is calculated based on item response theory (IRT).

The exam is simultaneously taken across the country once a year at the end of the year. It costs approximately $20 to take the exam, although this fee is waived for low-income applicants. Between 2009 and 2016, over 50 million individuals signed up to take the ENEM, making it the second-largest college admission exam globally. In Appendix C, I describe the main changes in the ENEM over time, explain how ENEM scores are used in the higher-education system other than for college admissions, and compare the ENEM to the US SAT and ACT exams.

2.2 Data

I combine three administrative databases from Brazil. The base dataset contains exam records from the ENEM from 2009–2016. This dataset contains both student-level and question-level information. The student-level data includes self-reported demographic and socioeconomic status (SES) measures, such as sex, race, high-school type (public/private), parental education, and family income. The question-level data includes each student's responses to each exam question, the position of the question, and skill tested.

To study individuals' trajectories through college and the labor market, I link the ENEM records to two other administrative datasets using individuals' national ID numbers (*Cadastro de Pessoas Físicas*). To measure college outcomes, I use Brazil's higher-education census from 2010–2019. This dataset includes information on all college enrollees' major, university, year of enrollment, number of credits, and year of graduation. To measure labor-market outcomes, I use an administrative employee-employer matched dataset called RAIS (*Relação Anual de Informações Sociais*) from 2016–2018. The RAIS covers the universe of formal-sector workers in Brazil, but it does not contain information on workers employed in the informal sector, self-employed individuals, or the unemployed. The RAIS contains both worker-level and firm-level information. Worker-level data includes educational attainment, occupation, and earnings. Firm-level data includes the number of employees, industry, and geographical location.

2.3 Samples and Summary Statistics

High-school-students sample. To construct this sample, I impose several sample restrictions. First, I only consider individuals who take the ENEM as high-school students. This restriction excludes individuals who take the exam after dropping out or graduating from high school. Second, I only include individuals with a non-zero non-missing score on each subject test. This restriction excludes, for example, students who missed one of the days of testing. I also exclude a small fraction of students with special accommodations, usually due to a disability. After these restrictions, the high-school-students sample contains information on approximately 15 million students who took the ENEM from 2009–2016. To examine students' long-run outcomes, I focus on 1.9 million high-school seniors in the first two cohorts in my data (2009–2010), for whom I observe college and labor-market outcomes 6–9 years after taking the exam.

Retakers sample. To assess the temporal stability of my measure of cognitive endurance,

I identify students who take the ENEM more than once, usually as high-school juniors to practice and again in their senior year to apply for college. Approximately 16% of test-takers in the high-school-students sample take the exam more than once.[4] I only include students with valid exam scores in all years. The retakers sample contains information on 1.5 million students or 3.1 million student-years.

Summary Statistics. Table 1 shows summary statistics on the samples. The average student in the high-school-students sample is 18.2 years old, 59.8% are female, 47.6% are white, and 22.2% went to a private high school (column 1). Over half of students (53.4%) have a high-school-educated mother, and 38.8% live in a household that earns an income above twice the minimum wage.[5] On average, students correctly respond to 34.3% of exam questions. High-school seniors from the 2009–2010 cohorts are slightly older, slightly more likely to be females, and white (column 2). Students in the retakers sample are slightly younger, their parents tend to have higher incomes, and they tend to perform better on the exam (column 3). Student characteristics are balanced across booklet colors (Appendix Table A1).

2.4 Definition of Main Outcomes

Test score. I define a student's exam score as the fraction of correct responses across all four academic subjects. The advantage of this measure is that it is intuitive and consistent with the existing literature (e.g., Zamarro et al., 2019). However, this measure differs from how the Brazilian testing agency calculates the ENEM score, which is based on IRT (see Appendix C.4). Reassuringly, the correlation between the fraction of correct responses and the IRT-based score is above 0.90 (Appendix Table C2).

College enrollment. I define college enrollment as an indicator for appearing in the higher-education census one year after taking the ENEM. The rest of the college outcomes are defined conditional on college enrollment.

College quality. I construct an earnings-based index of college quality. To do this, I group all college-educated workers in the RAIS (not just the workers in my sample) based

[4]Some high-school students take the ENEM more than two times in my sample, possibly because of grade repetition. I exclude a small fraction of students who take the ENEM more than three times.

[5]Students self-report their household income and other SES measures when they enroll to take the ENEM. Household income is elicited in ranges and expressed as a multiple of the minimum wage. For some analysis, I divide students into those whose household earns more than five minimum wages and those whose household earns less than twice minimum wage. Using the Brazilian National Household Survey, I find that the former households are in the top 30% of the national income distribution, while the latter households are in the bottom 30%.

on the university they attended and compute the average earnings of the graduates from each university.[6]

College degree quality. I create an index of college degree (or major) quality using the average earnings of the graduates of each college degree. To allow for international comparisons, I classify majors based on the International Standard Classification of Education (UNESCO, 2012).

Degree progress. I calculate the ratio between the number of credits completed at the end of each academic year and the total number of credits required to graduate. This variable is available starting in the 2015 higher-education census. Thus, I use student data from the cohort enrolled in 2015 to measure this outcome.

Likelihood of graduating. I define an indicator for graduating one to six years after enrolling in college. Most students who ever graduate do so within the first six years (Appendix Figure A1). As robustness, I define a measure of on-time graduation based on expected degree length. The higher-education census contains information on how long a student in good standing should take to graduate from each program. I use this information to define an indicator for graduating within the expected number of years.

Formal employment. I define formal employment as an indicator for appearing in the employee-employer matched dataset in any year in my sample. This variable is defined for all test-takers. The rest of the labor-market outcomes are defined conditional on formal employment. If an individual has multiple jobs, I use the data from the job with the highest number of hours. I use the job monthly earnings as a tiebreaker.

Monthly earnings. This variable represents the average salary of a worker across all months in a given year. To report this variable, firms have to calculate the worker's total earnings for the year and divide them by the number of months the firm employed the worker. If a worker appears in multiple years in the RAIS, I calculate the inflation-adjusted average monthly earnings across all years. I adjust earnings for inflation using the consumer price index.

Hourly wage. I calculate the hourly rate of each worker as the ratio between a worker's inflation-adjusted monthly earnings and the hours worked per month.[7] If a worker appears in multiple years in the RAIS, I calculate the average hourly wage across all years.

Firm, industry, and occupation mean wage. I calculate the average hourly wage at each

[6] This index is analogous to the college quality measure used by Chetty et al. (2011) and Chetty et al. (2014b) to study the long-term impacts of kindergarten quality and teachers, respectively.

[7] Firms do not record the number of hours individuals actually work each week. Instead, the data on hours indicates the number of hours per week that the worker is expected to work based on her contract.

firm, industry, and occupation. I use leave-one-out measures so that an individual's own employment outcomes do not affect the mean wage. I define firms using the 14-digit CNPJ,[8] industries using the Brazilian National Classification of Economic Activities (CNAE), and occupations using the Brazilian Occupational Code Classification (CBO). I calculate the wage indices separately for each year and use the average value across years.

I measure labor-market outcomes for the 2009–2010 cohort using employment data from 2016–2018. This means that, for the 2009 cohort, I measure outcomes 7–9 years after taking the ENEM, and for the 2010 cohort, 6–8 years after taking the ENEM. I account for this variation by controlling for an individual's potential years of experience throughout the analysis. I measure potential experience as the individual's age minus the years of schooling minus six.

3 Empirical Framework

This section lays out a simple potential-outcomes framework. I use the framework to formally define cognitive endurance in terms of empirical estimands and to clarify the identification assumptions.

3.1 Statistical Model

Let C_{ij} be the probability of individual i correctly answering question j. I model C_{ij} as a function of the student's level of cognitive fatigue, f_{ij}. Fatigue affects performance by impairing cognitive functions such as attention, memory, or reasoning (Ackerman, 2011). The effects can be manifested in many ways, including students forgetting a crucial formula, making a computation mistake, misinterpreting or ignoring an important aspect of a question, and filling in the wrong bubble in the multiple-choice sheet.

To build intuition, first consider an environment in which fatigue is binary: individuals can be either mentally "fresh" ($f_{ij} = 0$) or "fatigued" ($f_{ij} = 1$). Let $C_{ij}(0)$ be the likelihood of individual i correctly answering question j if she is fresh and $C_{ij}(1)$ the likelihood if she is fatigued. These two probabilities denote potential outcomes for different fatigue levels, but only one of the two outcomes is observed. The observed performance, $C_{ij}(f_{ij})$, can be

[8]The CNPJ is a tax identifier for legally incorporated identities. The first eight digits identify the company. The rest of the digits identify the branch or subsidiary of the company.

written in terms of these potential outcomes as

$$C_{ij}(f_{ij}) = C_{ij}(0) + \underbrace{\Big(C_{ij}(1) - C_{ij}(0)\Big)}_{\text{“Fatigue effect” } (\kappa_i)} f_{ij}, \tag{1}$$

where $C_{ij}(1) - C_{ij}(0) \equiv \kappa_i$ measures the effect of fatigue on performance, or “fatigue effect,” for short. I allow the fatigue effect to be heterogeneous across individuals, although for simplicity I assume that it is constant across types of questions.

Suppose for the moment that we observed whether the individual was fresh or fatigued when she answered each exam question. Then, one could compare i's average performance in questions she answered while fatigued ($\mathbb{E}[C_{ij}|f_{ij} = 1]$) to her average performance in questions she answered while rested ($\mathbb{E}[C_{ij}|f_{ij} = 0]$). This comparison can be written as

$$\begin{aligned}\mathbb{E}[C_{ij}|f_{ij} = 1] - \mathbb{E}[C_{ij}|f_{ij} = 0] = &\underbrace{(\mathbb{E}[C_{ij}(1)|f_{ij} = 1] - \mathbb{E}[C_{ij}(0)|f_{ij} = 1])}_{\text{Term 1: Fatigue effect}} \\ &+ \underbrace{(\mathbb{E}[C_{ij}(0)|f_{ij} = 1] - \mathbb{E}[C_{ij}(0)|f_{ij} = 0])}_{\text{Term 2: Selection bias}}.\end{aligned}$$

This expression shows that a comparison of average performance yields the sum of two terms. The first one is the fatigue effect for questions answered while fatigued. The second term is a selection bias that arises when comparing performance across different questions. For example, if individuals become fatigued over time, a selection bias might arise if questions become increasingly hard over the course of the exam. In this case, i's average performance would deteriorate even if she had not experience fatigued.

In practice, cognitive fatigue is not binary; rather, an individual can have different gradations of “tiredness.” In what follows, I assume f_{ij} is continuous and interpret κ_i as the impact of a unit change of cognitive fatigue on performance. Because cognitive fatigue cannot be directly observed, estimating κ_i is not feasible. In the empirical analysis, I use the position of question j on the version of the exam answered by i (Position_{ij}), under the reasoning that students become increasingly fatigued over the course of the exam. This notion is supported by research showing that time-on-task is a significant determinant of cognitive fatigue (e.g., Ackerman and Kanfer, 2009). To understand how cognitive fatigue relates to question position, consider a hypothetical linear projection of f_{ij} on Position_{ij}:

$$f_{ij} = \omega_i + \pi_i \text{Position}_{ij} + \eta_{ij}. \tag{2}$$

The intercept of the projection, ω_i, measures i's cognitive fatigue at the beginning of the test. The slope of the projection, π_i, measures the change in cognitive fatigue due to a one-position increase in the order of a given question. η_{ij} is a mean-zero projection error, uncorrelated with Position_{ij} by definition. If student i answers the exam in chronological order and finds the exam mentally taxing, we would expect $\pi_i > 0$. Using equation (2), it is possible to re-write equation (1) as a regression equation that can be estimated in observational data:

$$C_{ij} = \alpha_i + \beta_i \text{Position}_{ij} + \varepsilon_{ij}. \tag{3}$$

The intercept of the regression, $\alpha_i \equiv \mathbb{E}[C_{ij}(0)] + \kappa_i\omega_i$, measures i's expected performance on the test if she were fresh ($\mathbb{E}[C_{ij}(0)]$), plus the impact of her initial level of fatigue on performance ($\kappa_i\omega_i$). Henceforth, I interpret α_i as a measure of i's academic ability. The slope of the regression, $\beta_i \equiv \kappa_i\pi_i$, is the estimand of interest. I interpret β_i as i's cognitive endurance. This reduced-form measure is the product of two structural parameters, κ_i and π_i, that are likely determined by several psychological mechanisms. For example, the performance of some individuals may be less impaired by cognitive fatigue (captured by κ_i) due to, for example, high intrinsic motivation or grit. Similarly, students may not become cognitively fatigued throughout the exam (captured by π_i) due to, for example, high attention capacity or high conscientiousness (which may lead to more diligent test preparation and, thus, better test-taking strategies). A limitation of the analysis is my inability to distinguish between the different mechanisms underlying cognitive endurance.

The random part of performance, $\varepsilon_{ij} \equiv C_{ij}(0) - \mathbb{E}[C_{ij}(0)] + \eta_{ij}$, measures deviations of i's potential performance on question j from her average potential performance. Comparing i's performance across exam questions in different positions yields the sum of cognitive endurance plus a selection bias:

$$\mathbb{E}[C_{ij}|\text{Pos}_{ij} = p] - \mathbb{E}[C_{ij}|\text{Pos}_{ij} = p-1] = \beta_i + \underbrace{\mathbb{E}[C_{ij}(0)|\text{Pos}_{ij} = p] - \mathbb{E}[C_{ij}(0)|\text{Pos}_{ij} = p-1]}_{\text{Selection bias}}.$$

Next, I describe the two research designs that I use to deal with the selection bias.

3.2 Identifying Cognitive Endurance

In the empirical analysis, I first estimate the average cognitive endurance across all students, $\beta \equiv \mathbb{E}[\beta_i]$. This parameter represents the causal effect of increasing a question's

position on average student performance, $\bar{C}_j \equiv \mathbb{E}[C_{ij}]$. Rejecting the null hypothesis of $\beta = 0$ would demonstrate that average student performance partly depends on cognitive endurance (i.e., this would show that $\kappa_i \pi_i \neq 0$ for some students).

To identify β, I use two research designs. The first research design consists of assessing how average student performance *on a given question* varies as a function of the question's position. This approach is enabled by the fact that a given question is located in different positions across booklets. To illustrate this approach, Appendix Figure D1 displays student performance in a natural science question (Appendix Figure C2 shows the text of the question). This question appears as early as position 46 in the gray booklet and as late as position 87 in the blue booklet. Accordingly, the fraction of correct responses declines from 40.8% in the gray booklet to 29.9% in the blue booklet. Comparing student performance in these two booklets reveals that an increase of 41 positions reduces performance on this question by 10.9 percentage points. Analogous pairwise comparisons can be made for any two booklets.[9] I exploit this information using the following fixed effects specification:

$$\bar{C}_{jb} = \alpha_j + \beta \text{Position}_{jb} + \xi_{jb}, \tag{4}$$

where $\bar{C}_{jb}$ is the fraction of students who correctly answered question j in booklet b and α_j are question fixed effects. Appendix Figure A3 illustrates the mechanics of identification by plotting average student performance on selected questions as a function of their position on the four exam booklets and the corresponding best-fit lines. β is identified by first estimating the effect of question position on average student performance separately for each question and then aggregating these question-specific best-fit lines (like the ones plotted in the figure) using the OLS weights.

The advantage of this approach is that it relies on a weak identification assumption—the random allocation of booklets across students. However, since each student only receives one exam booklet, I cannot compare a student's performance across different booklets to identify β_i. Thus, I also use a second empirical strategy that can be used to identify both β and β_i.

The second empirical approach consists of controlling for question difficulty (Difficulty_j) in equation (4) instead of the question fixed effects. To estimate β, I assess how average

[9] Not all questions appear in a different position across all booklets. Appendix Figure A2 shows the variation in question position across all questions for every pairwise booklet combination.

student performance changes throughout the exam in regressions of the form:

$$\bar{C}_{jb} = \alpha + \beta \text{Position}_{jb} + \delta \text{Difficulty}_j + \mu_{jb}. \tag{5}$$

One challenge in implementing this approach is measuring question difficulty. An intuitive and often used measure of a question's difficulty is the fraction of students who correctly answered the question. However, a given question has a different fraction of correct responses depending on where it is located on the booklet. Thus, a question might appear to be more difficult simply because it is located later in the exam on average across booklets. To deal with this, I exploit the within-question position variation to construct a "position-adjusted" measure of a question's difficulty. This measure of question difficulty represents the fraction of correct responses we would expect to observe if question j appeared in the first position of the exam (see Appendix D for details). To avoid a spurious correlation, I calculate question difficulty using data from test-takers outside my sample.[10]

This strategy yields a consistent estimate of β under the assumption that unobserved determinants of student performance are conditionally independent of question position. Below, I provide evidence in support of this assumption. Importantly, as I describe in Section 5, this second empirical strategy can also be used to identify the cognitive endurance of each individual. In this case, there is one orthogonality assumption per student. The identification assumption requires any unobserved determinants of i's test performance to be uncorrelated with question position (conditional on question difficulty). I describe the consequences of violations of this assumption in Section 5.2. In the following two sections, I present estimates of average cognitive endurance (Section 4) and individual-level endurance (Section 5).

4 Cognitive Endurance and Test Performance

This section shows student performance trends over the course of the ENEM and presents estimates of average cognitive endurance using two research designs.

[10] These are mainly individuals who took the ENEM after graduating from high school. The results are very similar if I use my sample to generate the measures of question difficulty. The correlation between the measure of question difficulty estimated with test-takers in my sample and outside my sample is 0.98.

4.1 Student Performance over the Course of the ENEM

To motivate the analysis, I begin by studying student performance over the duration of the exam without controlling for question difficulty or any other performance determinant that may be changing throughout the exam. Figure 1 plots the fraction of students who correctly responded to each exam question (y-axis) against the position of the question in the test (x-axis). As a benchmark, the red dashed line shows the expected performance if students randomly guessed the answer to each question.

Figure 1 reveals a strong negative relationship between student performance and question position. Average performance decreases from about 45% at the beginning of the exam to about 24% at the end of the exam. A bivariate regression of the fraction of correct responses on question position indicates that average student performance declines by 21.4 percentage points over the course of each testing day ($p < 0.01$), as shown in Table 2, column 1. In addition, Figure 1 shows that average performance *increases* from about 30% at the end of the first day to about 45% at the beginning of the second day.[11]

Limited cognitive endurance can provide a parsimonious explanation of these patterns. As students advance through the exam, their mental resources may become increasingly taxed, and thus they become more prone to committing mistakes. Cognitive resources are replenished after taking a break (Sievertsen et al., 2016) and overnight via sleep (Baumeister, 2002; Lim and Dinges, 2008), which may explain why performance increases between the end of the first day and the beginning of the second day. Next, I implement the research designs described in Section 3.2 to identify average cognitive endurance.

4.2 Estimates of Average Cognitive Endurance

Table 2 presents the regression estimates from the two research designs. To facilitate the interpretation of the coefficients, I scale β so that it can be interpreted as the decrease in student performance over the course of each testing day.

Estimating the question-fixed-effects specification (equation 4) yields an average cognitive endurance $\beta = -0.072$ ($p < 0.01$), as shown in column 2. This estimate indicates that student daily performance decreases, on average, by 7.2 percentage points due to limited

[11] Interestingly, Figure 1 also shows that student performance seems to *increase* towards the end of each testing day. This pattern is not unique to the ENEM; a similar pattern has been found in the SAT (Mandinach et al., 2005) and the PISA test (Borghans and Schils, 2018). One possible explanation is what Mullainathan and Shafir (2013) refer to as the "the focus dividend," that is, the notion that when a resource is scarce (in this case, the time left to finish the exam), the mind becomes better at focusing and blocking distractions.

cognitive endurance. The difficulty-adjusted regression specification (equation 5) yields an estimate of average cognitive endurance $\beta = -0.058$ ($p < 0.01$), as shown in column 3. The similarity of this estimate relative to that obtained from the fixed effects specification suggests that controlling for question difficulty is adequate to account for differences in question characteristics. Moreover, the R-squared indicates that 97% of the variation in $\bar{C}_j$ is explained by a question's position and difficulty. This high R-squared suggests that there is little scope for unobservable variables to affect $\bar{C}_j$, further providing supporting evidence for the selection-on-observables assumption (Oster, 2019).

Figures 2 and 3 provide visual evidence on how limited endurance impacts student performance. Figure 2 plots average student performance over the course of the exam after removing the influence of question difficulty on performance. To construct this figure, I first regress $\bar{C}_{jb}$, the fraction of students who correctly answered question j in booklet b, on question difficulty, Difficulty_j, and estimate the residual from this regression, $\bar{C}^r_{jb} = \bar{C}_{jb} - \mathbb{E}[\bar{C}_{jb}|\text{Difficulty}_j]$. I add back the sample mean to $\bar{C}^r_{jb}$ to facilitate interpretation of units. Finally, I plot the mean value of $\bar{C}^r_{jb}$ across the exam. The figure shows that difficulty-adjusted performance tends to decline linearly throughout the exam. Daily performance decreases by about 5.2 percentage points each day, an effect consistent with the regression estimates.

Figure 3 plots the average percentage point change in the probability of correctly answering a question (y-axis) against the change in question position (x-axis) across all questions. The line is the predicted value from a linear regression estimated on the micro data. Its intercept is statistically equal to zero, indicating that a given question is, on average, equally likely to be answered if it appears in the same position in two different booklets. The slope indicates that, on average, a given question is 0.08 percentage points less likely to be correctly answered if it appears one position later in the test ($p < 0.01$). Thus, student performance decreases by about one percentage point roughly every 12 questions (or 36 minutes if students spend the exam time uniformly across questions). The implied daily change in performance due to limited endurance equals 7.2 percentage points ($p < 0.01$), an estimate quantitatively identical to the question-fixed-effects specification.

Taken together, the evidence indicates that average student performance decreases by about 5–7 percentage points per day due to limited cognitive endurance. This effect is sizable. The fixed-effects-specification estimate represents a 16% decrease of the estimated performance at the beginning of the exam (equal to 45%, Table 2, column 1) or about 60% of the standard deviation of overall test score (equal to 11.6 percentage points). The effect

is comparable to that of a decrease of half a standard deviation in teacher quality (Chetty et al., 2014a), an increase in the class size of about 16 pupils (Angrist and Lavy, 1999), or taking the exam under 66 degrees Fahrenheit hotter conditions (Park, 2022).

4.3 Limited Cognitive Endurance or Time Pressure?

Throughout this section, I have interpreted the causal effect of an increase in question position on performance as a manifestation of limited cognitive endurance. This interpretation is in line with the framework in Section 3. However, an estimate of $\beta < 0$ could also potentially be generated by students running out of time toward the end of the exam.

In Appendix B.1, I provide two pieces of evidence against this alternative interpretation. First, very few students leave any responses unanswered. Second, performance declines are present even when students respond to questions while they are likely not time-pressured (such as when responding to questions in the first half of each testing day). This evidence supports the interpretation of $\beta < 0$ as a consequence of limited cognitive endurance.

5 Decomposing Test Scores into Ability and Cognitive Endurance

The results in Section 4 demonstrate that test scores reflect not only students' academic preparedness ("ability") but also their capacity to endure mental fatigue ("cognitive endurance"). This section decomposes individuals' test scores into these two skills and examines the test-retest reliability of the generated measures.

5.1 Linear Decomposition

To quantify the relative influence of ability and endurance on a student's test score, I estimate the difficulty-adjusted regression specification separately for each student:

$$C_{ij} = \alpha_i + \beta_i \text{PosNorm}_{ij} + \delta_i \text{Difficulty}_j + \varepsilon_{ij} \quad \text{for } i = 1, ..., N, \tag{6}$$

where C_{ij} equals one if student i answered question j correctly, PosNorm_{ij} is question position normalized such that the first question of each day equals zero and the last question equals one, and Difficulty_j is the position-adjusted measure of question difficulty, normalized to have mean zero.[12] In the baseline specification, I estimate equation (6) pooling

[12]Because students may answer the exam in any order, the position of the questions on each booklet (the variable I observe) is a noisy measure of the order in which students answered the exam (which I do

student responses from both testing days and all academic subjects and show robustness to including day and subject fixed effects, to estimating the parameters separately by day and subject, and to estimating non-linear models.

Without further assumptions, $\hat{\alpha}_i$ and $\hat{\beta}_i$ simply describe how i's performance changes throughout the test. The intercept of each regression, $\hat{\alpha}_i$, measures the predicted performance of student i in the first exam question for a question of average difficulty. Thus, $\hat{\alpha}_i$ represents i's performance after accounting for the impact of a question's position and difficulty. The slope of each regression, $\hat{\beta}_i$, measures the predicted performance change between the first and last question of each testing day after accounting for question difficulty.[13] Importantly, equation (6) can be interpreted as an observational analog of the model (3). Under this model, $\hat{\alpha}_i$ measures i's academic ability and $\hat{\beta}_i$ measures i's cognitive endurance.

5.2 Limitations of Measuring Endurance using Standardized Tests

This approach to measuring cognitive endurance has advantages but also important limitations. The main advantage is that it is based on observed behavior ("revealed preference"). This deals with some of the well-known biases of measures based on self-reports ("stated preferences"). Examples include social-desirability bias (i.e., respondents want to look good in front of the interviewer), reference-group bias (i.e., respondents judge their behavior using different standards), and framing effects (i.e., slightly different ways of asking the same question can cause large changes in respondents' answers). The magnitude of some of these biases has been shown to be quantitatively significant for self-reported grit and self-control, two non-cognitive skills related to cognitive endurance (Lira et al., 2022).

However, there are at least three important concerns with the measure. First, estimating individual-level endurance requires one orthogonality condition per student. The identifying assumption is unlikely to hold exactly for *all* students. For example, some students may happen to be unprepared for the questions that appear at the end of the exam. Thus, their decline in performance would partly be driven by lack of preparation, leading to biased estimates of endurance. If the departures of the identification assumption are not systematic (e.g., some students are unprepared for questions at the end, but others are

not observe). Thus, my estimates should be interpreted as "intention-to-treat" estimates.

[13] In Appendix B.2, I derive the OLS estimate of β_i. The formula shows that $\hat{\beta}_i$ is calculated as a weighted average of deviations of i's performance on each exam question from i's average performance. Thus, $\hat{\beta}_i$ captures the intuition that a student who tends to do worse in the latter parts of the exam—relative to her average—has low endurance.

unprepared for questions at the beginning), then this issue is equivalent to measurement error, which would attenuate the effects documented below. Using the retakers sample, I provide evidence consistent with this interpretation. In addition, I show that the results are similar using several alternative measures of endurance (e.g., calculated separately for each academic subject and using the average).

Second, my endurance measure is predicated on the assumption that students answer the exam in chronological order. It is worth highlighting that the order in which students answer the exam is endogenous. Some students may leave the questions they find harder until the end. Thus, regressing test performance on the order in which students answered the exam would show that performance tends to decline over time simply because these students are strategically skipping the questions they find challenging. Using the order in which questions are positioned on the exam as a regressor deals with this problem but may lead to attenuated endurance estimates.

Finally, floor or ceiling effects can bias my estimates of endurance. For example, individuals with extremely low ability or endurance may randomize their responses throughout the entire exam and show up in the data as having high measured endurance due to their stable performance. While this issue is not specific to my measure of endurance, it may be a concern for the empirical analysis. Below, I show that the results are robust to excluding students in the tails of the ability and the endurance distributions (i.e., students for whom floor and ceiling effects are more likely to be binding).

5.3 Assessing the Reliability of the Cognitive Endurance Measure

Are the measures of fatigued-adjusted ability and cognitive endurance generated by the decomposition reliable? To assess the reliability of a construct, researchers typically measure the construct multiple times and calculate the "temporal stability" or correlation between these measures (Miller et al., 2009). The size of the correlation is a measure of construct reliability; the higher the correlation, the more reliable the construct.

I compute two measures of test-retest reliability. First, I estimate ability and endurance separately for each testing day and calculate the correlation between consecutive days. The advantage of this approach is that it can be implemented in my main sample. The drawback is that the academic subjects tested vary each day, which could affect the reliability estimates.[14] Second, I estimate the temporal stability of ability and endurance between

[14]For example, students who are good at natural science (a subject test on the first day) might not be as good at language arts (a subject test on the second day). This would lead to an imperfect between-day

consecutive years. This analysis produces more comparable estimates, but it can only be done using the smaller sample of retakers.

The test-retest reliability of academic ability and cognitive endurance is comparable to that of other well-known constructs. Figure 4 show a series of binned scatterplots plotting the average $t+1$ estimate of ability/endurance as a function of the time t estimate. The temporal stability of ability ranges from 0.61 (between consecutive days) to 0.77 (between consecutive years). The temporal stability of cognitive endurance ranges from 0.14 (between consecutive days) to 0.30 (between consecutive years). The test-retest reliability of fatigue-adjusted ability and cognitive endurance is comparable to the reliability of other well-known psychological and economic constructs.[15]

5.4 Summary Statistics on Ability and Cognitive Endurance

Average cognitive endurance is $\hat{\beta} = -0.058$, meaning that, due to limited endurance, the performance of the average student decreases by 5.8 percentage points over the course of the exam. This estimate is consistent with the quasi-experimental results shown in Section 4. The standard deviation of $\hat{\beta}_i$ is $\sigma_{\hat{\beta}} = 14.4$ percentage points.[16] Because of sampling error in $\hat{\beta}_i$, this raw standard deviation overstates the variability of true latent β_i, σ_β. Following Angrist et al. (2017), I estimate σ^2_β as

$$\hat{\sigma}^2_\beta = \sigma^2_{\hat{\beta}} - \mathbb{E}[\text{SE}^2_{\hat{\beta}}],$$

where $\mathbb{E}[\text{SE}^2_{\hat{\beta}}]$ is the average squared standard error of $\hat{\beta}_i$. I construct an analogous estimate for the standard deviation of latent ability, $\hat{\sigma}_\alpha$ (see Appendix B.3 for details).

The standard deviation (SD) of β_i is $\hat{\sigma}_\beta = 0.088$. This means that an increase of one SD in cognitive endurance predicts an 8.8 percentage point increase in the test score. The corresponding estimate for fatigue-adjusted ability is $\hat{\sigma}_\alpha = 0.132$. Hence, $\hat{\sigma}_\beta$ is about two-thirds the magnitude of $\hat{\sigma}_\alpha$, meaning that ability is more dispersed than endurance across students. These estimates can be translated into percentage effects by dividing by

correlation.

[15] Appendix Table A2 includes examples of reliability estimates for some well-known economic and psychological constructs. IQ is the construct with the highest known reliability, with correlations on the order of 0.80 (Hopkins and Bracht, 1975; Schuerger and Witt, 1989). Other commonly used constructs have lower temporal stability. For example, reliability estimates of risk aversion range 0.20–0.40 (Mata et al., 2018); big five personality range 0.49–0.70 (Wooden, 2012); and teacher value-added range 0.23–0.47 (Chetty et al., 2014a).

[16] Appendix Figure A5 shows the distribution of estimated ability (Panel A) and endurance (Panel B).

the average test score of 0.344 (Table 1, Panel D). Under this rescaling, the estimates imply that a one SD increase in endurance leads to a 25.6% increase in test score. The corresponding impact of ability equals 38.3%.

Figure 5 shows the joint distribution of estimated ability and endurance. The red diamonds show a binned scatterplot of mean endurance as a function of ability, calculated by dividing students into 100 equally-sized ability bins. The gray circles display a scatterplot of $\hat{\beta}_i$ against $\hat{\alpha}_i$ for a randomly-selected one percent of my sample.

Figure 5 reveals two important patterns. First, there is substantial variation in individuals' ability-endurance combination.[17] Second, there is a negative relationship between $\hat{\alpha}$ and $\hat{\beta}$. On average, individuals with low values of $\hat{\alpha}_i$ tend to have higher values of $\hat{\beta}_i$. This relationship is largely mechanical and it is driven by floor and ceiling effects. Low-ability individuals have a limited margin to decrease their performance throughout the exam because test scores are bounded. An analogous argument holds for high-ability individuals. This generates a "missing mass" of individuals with low-ability low-endurance and high-ability high-endurance, inducing a negative correlation between the two variables.[18]

In the analysis below, I always control for both variables to account for their mechanical relationship and show robustness to excluding individuals in the tails of the ability/endurance distribution.

In the following two sections, I use the estimates of fatigued-adjusted ability and endurance to (i) revisit the association between test scores and long-run outcomes through the lens of the ability-endurance decomposition and (ii) characterize how systematic differences in endurance across students affect test-score gaps and the information content of test scores.

6 Cognitive Endurance and Student Outcomes in Adulthood

In this section, I use the decomposition to separately quantify the contribution of ability and endurance to the well-known association between test scores and long-run outcomes (e.g., Bishop, 1989; Hanushek and Woessmann, 2008, 2012).

[17] For example, for individuals with $\hat{\alpha}_i \simeq 0.50$, their estimates of endurance ranges from $\hat{\beta}_i = -0.50$ (a value roughly in the bottom one percent of the endurance distribution) to $\hat{\beta}_i = 0.50$ (a value in the top one percent).

[18] For individuals with intermediate values of ability (for whom ceiling and floor effects are less likely to be binding), the correlation is negligible. For example, the correlation between the two measures is -0.08 for individuals with estimated ability between 0.50 and 0.60.

6.1 Estimating the Return to Academic Ability and Cognitive Endurance

To assess how test scores and their component skills (ability and endurance) relate to college and labor-market outcomes, I estimate regressions of the form:

$$Y_i = \phi + \lambda X_i + \psi_T \text{TestScore}_i + \nu_i \tag{7}$$

$$Y_i = \tilde{\phi} + \tilde{\lambda} X_i + \psi_A \text{Ability}_i + \psi_E \text{Endurance}_i + \tilde{\nu}_i, \tag{8}$$

where Y_i is an outcome of student i; Ability_i and Endurance_i are the measures of academic ability and cognitive endurance estimated in Section 5; and X_i is a vector that contains demographic variables and socioeconomic status.[19] For labor-market outcomes, I additionally control for educational attainment and potential years of experience. Because students can enroll in multiple college degrees, each observation denotes a student–degree combination. I account for the fact that an individual can appear multiple times in the dataset by clustering the standard errors at the individual level.

To compare the magnitude of the predicted effect of endurance on a given outcome with the corresponding effect of academic ability, I normalize both variables such that their coefficients represent the effect of a one SD increase on a given outcome.

6.2 Baseline Estimates

Table 3 presents estimates of equations (7) and (8) using as dependent variables college outcomes (Panel A) and labor-market outcomes (Panel B). I first discuss college outcomes and then turn to labor-market outcomes.

6.2.1 College outcomes. Consistent with a sizable literature on the strong predictive power of test scores, I find that students with higher test scores tend to have better college outcomes. Students with a one SD higher test score are 8.8 percentage points more likely to enroll in college (relative to a mean of 24.4%, column 1). Conditional on enrolling in college, the quality of their institution and college major—as measured by the average earnings of previous graduates—is 8.2%–11.7% higher (columns 2–3), the share of total credits they complete by the end of their first year is 1.4 percentage points higher (an 8.8% increase relative to the mean of 15.8%, column 4), and they are 6.0 percentage points more

[19] For students with a missing value for a control variable, I define the missing value as equal to the sample mean value and include a dummy for missing student characteristics in the regressions.

likely to graduate (column 5). Conditional on graduating, they take 0.12 fewer years to graduate (a 3.1% decrease relative to the mean of 3.4 years, column 6).[20]

Both ability and endurance have a sizable predicted impact on college outcomes. The predicted effect of fatigued-adjusted ability on college outcomes is stronger than the corresponding effect of test scores. More interestingly, cognitive endurance has an economically and statistically significant effect on college outcomes. A one SD increase in endurance predicts a 2.9 percentage points increase in the likelihood of enrolling in college ($p < 0.01$); a 8.2% increase in the college quality ($p < 0.01$), and a 6.0 percentage point increase in the six-year graduation rate ($p < 0.01$). To benchmark the size of these associations, I compute the ratio between the predicted effect of endurance on an outcome and the predicted effect ability ($\hat{\psi}_E/\hat{\psi}_A$). This ratio is shown in the third-to-last row in Panel A. The effect of endurance as a percent of the effect of ability ranges from 31.6%–36.2%, depending on the outcome.

Figure 6, Panels A–C present binned scatterplots of selected college outcomes against endurance. To construct each panel, I first regress Y_i and Endurance_i on student-level characteristics and ability, and estimate the residuals from these regressions, Y_i^r and Endurance_i^r (adding back the unconditional sample mean to facilitate the interpretation of units). Then, I group individuals into 10 equally-sized bins (deciles) based on Endurance_i^r and plot the mean value of Y_i^r for each bin. Consistent with the regression results, there is a strong relationship between endurance and college enrollment (Panel A), college quality (Panel B), and the six-year graduation rate (Panel C).

These results indicate that both academic ability and endurance are key predictors of college success. While the importance of academic ability has been widely documented, the results suggest that endurance plays a commensurate role in college success. Moreover, the results show that traditional estimates of the impact of test scores (often used as a proxy for cognitive skills) on long-run outcomes partly measure the effect of endurance on those outcomes.

6.2.2 Labor-market outcomes. Students with higher test scores tend to have better labor-market outcomes. On average, students with a one SD higher test scores are 0.1 percentage points more likely to have a formal-sector job (column 1), have a 12.7% higher

[20] These estimates are comparable to those in the literature. For example, Chetty et al. (2014b) estimates that a one SD increase in test scores is associated with a 5.5 percentage point increase in college enrollment at age 20, a 7.8% increase in college quality as measured by the earnings of previous graduates, and an 11.9% increase in earnings at age 28 (see their Appendix Table 3, row 2).

hourly wage (column 2), earn a 10.9% higher monthly salary (column 3), work in firms that pay 9.1% higher wages (column 4), choose occupations that pay 4.1% higher wages (column 5), and work in industries that pay 1.3% higher wages (column 6).

These associations reflect both the predicted impact of academic ability and cognitive endurance, both of which have statistically and economically significant effects on labor-market outcomes. For example, a one SD increase in endurance predicts a 5.4% increase in hourly wages ($p < 0.01$), a 5.2% increase in monthly earnings ($p < 0.01$), and a 3.6% increase in average firm wage ($p < 0.01$). The strong relationship between cognitive endurance and these three outcomes is illustrated in binned scatterplots in Figure 6, Panels D–F. These figures show that mean wages and earnings increase roughly linearly with endurance. Depending on the outcomes, the predicted effect of cognitive endurance as a percent of the predicted effect of ability ranges from 25.5%–38.7%.

These results indicate that endurance has a sizable wage return in the labor market. Under complete information and frictionless markets, the price of a skill equals the present value of the future returns generated by the skill (Abraham and Mallatt, 2022). Thus, the sizable wage return to endurance suggests that this skill is a key productivity determinant. The positive wage return to ability and to endurance are consistent with models in which firms pay workers according to their productivity, and output is generated by combining ability with cognitive effort. Cognitive endurance enables workers to sustain effort for a longer time, allowing them to produce a higher total output. The results also reveal a novel type of assortative matching in the labor market: workers with high cognitive endurance are more likely to work for high-paying firms. This is relevant given that the sorting between workers and firms is an important driver of labor-market outcomes (Card et al., 2018).

6.2.3 Robustness. Appendix B.4 presents a series of robustness and specification checks. The baseline results are robust to computing the effects nonparametrically, estimating ability and endurance with alternative specifications (e.g., with day or subject fixed effects), and imposing several sample restrictions (e.g., excluding the tails of the ability or endurance distribution).

6.3 Why is Cognitive Endurance related to Earnings?

An important question is whether the predicted impact of endurance on long-run outcomes is due to the mechanical relationship between this variable and the test score. When

holding ability constant, an increase in endurance leads to a higher test score. The effects documented above could result from the opportunities created by having a better test score, rather than endurance being a valuable skill itself.

To shed light on this, I use a measure of cognitive endurance that is not mechanically related to the test scores students use to apply for college. ENEM scores are only valid for one year, which means students cannot use their scores from previous years to apply for college. Thus, I instrument the year t measure of endurance (and ability) with the year $t-1$ measures. Using repeated measures of a skill as an instrument additionally helps to deal with measurement error (e.g., Gronqvist et al., 2017; Edin et al., 2022).

Appendix Tables A3 and A4 present the results. Panel A reports OLS estimates on the retakers sample and Panel B the instrumental variables (IV) estimates. The OLS coefficients estimated on the retakers sample are comparable to those estimated on the main sample. The IV estimates tend to be larger than the OLS estimates. For example, the OLS estimate of the effect of a one SD increase in endurance [ability] on wages is 12.1% [23.1%], while the IV estimate is 18.8% [25.0%]. Hence, the IV estimates suggest that the wage return to endurance—as a percent of the wage return to ability—is significantly higher, on the order of 75%. The difference between the IV and OLS estimates tends to be larger for the endurance effects than for the ability effects, consistent with the endurance measure containing more measurement error than ability. In sum, these results indicate that the predicted impact of endurance on long-run outcomes is not due to a mechanical relationship between endurance and test scores.

6.4 The Value of Endurance across Degrees, Occupations, and Industries

The task-based approach to labor markets highlights that workers produce output by performing job tasks, and tasks differ in their skill requirements (Acemoglu and Autor, 2011). Consequently, the value of endurance should vary according to the tasks individuals have to accomplish in a given job and the importance of endurance in the production function of those tasks. For example, endurance may be particularly important for some jobs because mistakes due to attentional lapses can dramatically reduce the output value, as in "O-ring" production functions (Kremer, 1993).

To assess this, I estimate the wage return to endurance separately for each college degree, occupation, and industry. If workers are paid according to their productivity, the wage return to endurance should reflect the increase in productivity due to an increase in this skill. Thus, a high wage return to endurance in a given occupation would indicate

that this skill is particularly valuable in the production function of the tasks required by such an occupation.[21]

Figure 7 plots the distribution of wage returns across college degrees (Panel A), occupations (Panel C), and industries (Panel E). There is substantial heterogeneity in the wage return to ability and to endurance. For example, while the average return to endurance across degrees is 4.9%, the return across degrees in the bottom decile of the return distribution is 0.1% and in the top decile is 9.8%. This suggests that cognitive endurance is more valuable for success in some college degrees. Occupations and industries also exhibit substantial heterogeneity in wage returns.

Figure 7 also show that degrees, occupations, and industries that tend to pay higher average wages tend to offer higher returns to ability and to endurance (Panels B, D, and F). For example, the return to endurance among the top-ten-percent-paying occupations is about three times higher than the return to endurance among the bottom-ten-percent-paying occupations (4.9% vs. 1.6%, respectively). This finding is consistent with high-paying jobs requiring high-endurance workers, suggesting that the value of this skill is higher in high-paying jobs.

Figure 8 shows the joint distribution of the wage return to ability and the wage return to endurance across college degrees (Panel A), occupations (Panel B), and industries (Panel C). The figure reveals a strong association between the wage return to ability and the wage return to endurance. For example, on average, a 10%-increase in the wage return to endurance across occupations predicts a 22.1% increase in the wage return to ability ($p < 0.01$). This finding suggests that ability and endurance are complementary skills in production. The more endowed workers are with one skill, the greater the value derived from the other skill.

To make tangible some of the real-world tasks for which endurance may be particularly valuable, Table 4 list the top-five degrees, occupations, and industries with the highest wage return to endurance. The list includes occupations where attentional lapses may be extremely costly, such as facility operators in petrochemical plants or air navigation professionals (Panel B). The list also includes degrees conducive to these occupations (e.g., aeronautics, Panel A) and related industries (e.g., oil extraction, Panel C). This list provides suggestive evidence that one of the psychological mechanisms behind the reduced-

[21] There are two important caveats with this approach to measuring the value of endurance. The first one is that individuals may select into degrees, occupations, and industries partly based on their endurance. The second one is that an increase in productivity may not lead to a corresponding increase in wages in some occupations or industries due to institutional factors (e.g., collective bargaining).

form measure of cognitive endurance is the capacity to sustain attention on a task for a long time.

7 Endurance, Test-score Gaps, and Exam Informativeness

Test scores muddle information about an applicant's ability and endurance. While a top score reveals that a student has both high ability and high endurance, an average test score might come from a student with high ability and low cognitive endurance, a student with low ability and high endurance, or a student who is average on both dimensions. Even though admission officers or employers would want to assess candidates primarily on the skill most critical to success, the information revealed about this skill is obfuscated by the less-relevant information about the other skill.

This informational problem could be overcome by reporting separate sub-scores for fatigue-adjusted ability and endurance. However, this may not be feasible due to institutional constraints. With a one-dimensional test score, the exam design will influence the extent to which exam performance reveals information about ability relative to endurance.

In this section, I focus on identifying the *distributional* and *informational* effects of an exam design that reveals more information about ability (and less about endurance). The distributional effect asks how the exam design impacts socioeconomic status (SES) test-score gaps, an important determinant of inequity in college access. The informational effect asks how the exam design impacts the information content of the exam, an important determinant of the student-college match quality.

7.1 Cognitive Endurance and Test-Score Gaps

Standardized tests often exhibit large racial and income test-score gaps (e.g., Fryer Jr and Levitt, 2006; Card and Rothstein, 2007; Riehl, 2022). In the context of college admission exams, these gaps lead to inequitable college access and amplify earnings disparities (Chetty et al., 2020). Understanding the sources of these gaps is an active area of research. Next, I examine the contribution of differences in cognitive endurance to these gaps.

7.1.1 Decomposing Test-Score Gaps. To begin with, notice that the linear decomposition (6) can be used to parsimoniously summarize an individual's test score, $\text{TestScore}_i \equiv \mathbb{E}[C_{ij}]$, as a linear combination of fatigue-adjusted ability and endurance:

$$\text{TestScore}_i = \hat{\alpha}_i + \hat{\beta}_i \overline{\text{Position}}.$$

Let $X \in \{0, 1\}$ be a student observable characteristic. For example, $X = 1$ may denote high-income students and $X = 0$ low-income students. The average test score of students with characteristic x can be written as

$$\mathbb{E}[\text{TestScore}_i | X_i = x] = \mathbb{E}[\hat{\alpha}_i | X_i = x] + \mathbb{E}[\hat{\beta}_i | X_i = x] \overline{\text{Position}}.$$

Using this expression, the test-score gap, ScoreGap, can be decomposed into differences in average academic ability and differences in average cognitive endurance:

$$\text{ScoreGap} = \underbrace{\alpha_1 - \alpha_0}_{\substack{\text{Difference in average} \\ \text{academic ability} \\ \text{between groups}}} + \underbrace{(\beta_1 - \beta_0)\overline{\text{Position}}}_{\substack{\text{Difference in average} \\ \text{cognitive endurance} \\ \text{between groups}}}, \tag{9}$$

where $\alpha_x \equiv \mathbb{E}[\hat{\alpha}_i | X_i = x]$ and $\beta_x \equiv \mathbb{E}[\hat{\beta}_i | X_i = x]$.

Equation (9) shows that, in the absence of systematic differences in limited endurance ($\beta_1 = \beta_0$), test scores gaps would be purely a reflection of gaps in academic ability. Thus, exam design features that put a higher or lower weight on endurance, such as the length of the exam or the number of breaks, should not affect test-score gaps. This is no longer true in the presence of systematic differences in endurance. If student-level characteristics are associated with endurance, then an exam design that puts more weight on endurance will affect test-score gaps.

I focus on estimating how an exam reform that decreases by half the length of the test impacts test-score gaps. This reform would decrease the average question position (from $\overline{\text{Position}}$ to $\overline{\text{Position}}/2$), thereby decreasing the influence of endurance gaps on test-score gaps.[22] This reform would be equivalent to changing the ENEM from its current length to roughly the length of the ACT exam.[23]

[22] An important concern is that, by reducing the number of questions, the exam would determine the place in the score distribution of any one student with less precision. However, the reform could be achieved without sacrificing much precision by using an adaptive exam that selects questions based on the student's ability level.

[23] While I focus on test length, other exam features can also affect the influence of endurance for test-score gaps. For example, Figures 1 and 2 show that student performance starkly increases between the end of the first day and the beginning of the second day, suggesting that giving students more breaks would decrease the importance of endurance for test-scores gaps. Thus, the reform can also be interpreted as,

Using equation (9), I estimate the effects of the reform on test-score gaps between:

1. Male and female students,
2. White and non-white (Black, Brown, and Indigenous) students,
3. Students in households in the top 30% and bottom 30% of the income distribution,
4. Students with a college-educated mother and non-college-educated mother,
5. Students enrolled in a private high school and public high school.

7.1.2 The Impact of an Exam Reform on Test-Score Gaps. Table 5 shows estimates of the contribution of gaps in ability and endurance to test-score gaps. Column 1 shows the difference in average test scores between the groups of students listed in the row header, column 2 shows the difference in average academic ability (in a regression that controls for endurance), and column 3 shows the difference in average cognitive endurance (controlling for ability).

By reducing the contribution of endurance gaps to test-score gaps by half, the reform would: (i) Reduce the gender test-score gap by 0.85 percentage points (a 32% decrease from the pre-reform gap of 2.6 percentage points); (ii) Reduce the racial test-score gap by 0.08 percentage points (a 14% decrease from the pre-reform gap of 5.7 percentage points); and (iii) Reduce the SES test-score gap by 1.3–3.1 percentage points (a 13%–16% decrease from pre-reform gaps), depending on the SES measure.

The predicted impact of the exam reform is robust to (i) measuring the gaps in percentiles (Appendix Table A5); (ii) Estimating ability and endurance with alternative specifications (Appendix Table A6); (iii) Using different measures of position-adjusted question difficulty when estimating ability and endurance (Appendix Table A7); (iv) Excluding individuals in the tails of the ability or endurance distributions (Appendix Table A8); and (v) Using precision-weighted estimates (Appendix Table A9).

7.2 Cognitive Endurance and Exam Informativeness

Admission officers use test scores to screen applicants partly because they are informative about which applicants will succeed in college. The standard approach to assess the informative content of an exam is to calculate the cross-individual correlation between test

for example, introducing a long break in the middle of each testing day.

scores and a long-run outcome that colleges want to screen their applicants based on (such as first-year college GPA or on-time graduation). This correlation is known as the *predictive validity* of an exam (Rothstein, 2004). Next, I study how an exam's predictive validity depends on cognitive endurance.

7.2.1 Decomposing Predictive Validity. The predictive validity of test scores for outcome Y, ρ^Y, can be written as a weighted average of the predictive validity of each exam question $j \in \{1, ..., J\}$:

$$\begin{aligned}\rho^Y &\equiv \text{Corr}(Y_i, \text{TestScore}_i) \\ &= \frac{1}{J}\sum_{j=1}^{J} \frac{\sigma_{C_j}}{\sigma_T}\rho_j^Y,\end{aligned}$$

where σ_T and σ_{C_j} are the standard deviations of test scores and question j responses, and $\rho_j^Y \equiv \text{Corr}(Y_i, C_{ij})$ is the predictive validity of question j.

Limited endurance affects a question's informativeness by changing the skill composition of students who correctly answer the question. To see this, notice that ρ_j^Y can be written as a function of the gap in average outcomes between students who correctly and incorrectly responded to question j:

$$\rho_j^Y = \Big(\mathbb{E}[Y_i|C_{ij} = 1] - \mathbb{E}[Y_i|C_{ij} = 0]\Big)\frac{\sigma_{C_j}}{\sigma_Y}.$$

As the empircal framework in Section 3 highlights, both ability and endurance are required to correctly answer questions, but the importance of these two skills varies throughout the exam. Loosely speaking, differences in performance at the beginning of the exam are mainly driven by differences in ability since all students are "fresh.". Toward the end of the exam, differences in performance depends on both differences in ability and differences in endurance.

I estimate how the length of an exam affects its informativeness by exploiting random variation in whether a given question is presented when students are relatively fresh or cognitively fatigued. Intuitively, if a given question if more predictive the later it appears on the exam, then we may expect a longer exam to be more informative since the additional questions contained by the exam would be especially informative. The opposite is true if a given question is less predictive the later it appears.

I empirically assess this by estimating regressions of the form:

$$\rho_{jb}^{Y} = \alpha_j + \gamma^{Y} \text{Position}_{jb} + \eta_{jb}. \tag{10}$$

where ρ_{jb}^{Y} is the predictive validity of question j in booklet b, α_j are question fixed effects, and Position_{jb} is the position of question j in booklet b. The coefficient of interest is γ^{Y}, which measures the impact of a one-position increase in the order of a given question on the question's predictive validity for outcome Y. I scale γ^{Y} so that it represents the change in predictive validity due to a reform that decreases the average question position by half.

I estimate the effect of the reform for eight main outcomes: test score (calculated without the contribution of question j to avoid mechanical effects), college enrollment, college quality, degree progress, six-year graduation rate, hourly wage, monthly earnings, and firm leave-individual-out mean earnings. Since the dependent variable is an estimate, I weight each observation using the inverse square of its standard error. I cluster standard errors at the question position level.

7.2.2 The Impact of an Exam Reform on the Test's Predictive Validity. Table 6 presents the results. Panel A shows the average predictive validity across all test questions. The average test question is predictive of long-run outcomes, although the size of the correlations tends to be small. For example, on average across all questions, correctly responding to a question has a 0.05 positive correlation with enrolling in college (column 2, $p < 0.01$), 0.10 correlation with college quality (column 3, $p < 0.01$), and 0.10 correlation with wages and earnings (columns 6–7, $p < 0.01$).

Panel B reports the estimates of equation (10). The exam reform would generate modest increases in the predictive validity of the exam for the majority of outcomes. For example, the exam reform would increase the average predictive validity of test responses for college enrollment by 0.05 points (a 95.2% increase relative to the pre-reform mean, $p < 0.01$), for college quality by 0.09 points (a 91.1% increase relative to the pre-reform mean), and for earnings by 0.07–0.08 points (a 75.7%–79.6% increase relative to the pre-reform mean, $p < 0.01$). The predicted effect for the six-year graduation rate is also positive but not statistically different from zero. The reform would decrease the predictive validity for degree progress.

These findings can be seen visually in Appendix Figure 9, which shows binned scatterplots plotting the change in the predictive ability of a question (y-axis) against the change in the question position (x-axis) for selected outcomes. In all cases, the average predictive

validity of test questions tends to decrease if the question appears later in the test.

The decreasing informativeness of questions during the exam can help explain puzzling empirical findings in the literature. Kobrin et al. (2008) study how the predictive validity of the SAT changed after the exam increased the number of questions in 2005. Intuitively, more test questions should lead to more precise student ability estimates, and thus more predictive test scores. Yet, the predictive validity of the exam remained unchanged. This finding can be explained by cognitive fatigue eroding the predictive power of test responses towards the end of the exam. Bettinger et al. (2013) show that performance on the English and Math sections of the ACT predict college outcomes, while performance on the Science and Reading sections do not. Notably, the Science and Reading questions are the last to appear in the ACT. Hence, the lack of predictive power of these two subjects may be driven by students being too fatigued by the time they reach those sections for their responses to be informative of their abilities, rather than by the skills assessed by Science and Reading questions being irrelevant for long-term outcomes.

7.3 Discussion

In summary, this section shows that, due to systematic differences in cognitive endurance among students, the design of college admission exams can have equity and efficiency consequences. I estimate that a reform that halves the exam length would reduce SES gaps by 26%–29%, possibly leading to a more diverse college student body. In addition, the shorter exam would be more informative about the quality of each applicant (as measured by its predictive validity), possibly leading to a better allocation of students to colleges.

The first result is driven by the fact that, conditional on academic ability, low-SES students have lower endurance than high-SES students; thus, their performance declines at a steeper rate throughout the exam. The second result is driven by the fact that differences in student performance at the beginning of the exam mainly reflect differences in ability (roughly, because most students are "fresh"), whereas performance differences towards the end of the exam increasingly reflect differences in endurance. Since ability is a stronger predictor of long-run outcomes than endurance, the predictive validity of a given question decreases if it appears later on the exam.

8 Conclusion

Just like individuals differ in preferences and personality traits, they also differ in their capacity to endure mental fatigue. This paper shows that cognitive endurance affects student performance in college admission exams and has a substantial earnings return in the labor market.

My findings have implications for investments in different types of human capital. I find that endurance is highly valued in the labor market. Yet, a typical school curriculum does not include any material directly aimed at building this skill. Policymakers should consider investing in the development of cognitive endurance, possibly during early ages when neuroplasticity is higher.[24] An important caveat in my analysis is the lack of exogenous variation in cognitive endurance. While my findings provide evidence of a positive link between endurance and earnings, these estimates may be misleading if the available control variables are inadequate to provide meaningful estimates of the causal effect of this skill on long-run outcomes.

The findings also have implications for designing more informative standardized tests. In a typical test, all questions contribute equally to an individual's score. However, questions that appear early in the exam are more predictive of long-run outcomes. A mechanism for aggregating individuals' test responses that takes into account students' varying fatigue levels throughout the exam may lead to more informative test scores. For example, testing agencies can weight each question based on the position in which it was answered, assigning more weight to questions students responded to earlier. Alternatively, testing agencies could report separate sub-scores of endurance and ability for each student. Admission officers could then decide how much weight to put on each of these two sub-scores, depending on the relative importance of these two skills for success in a given program.

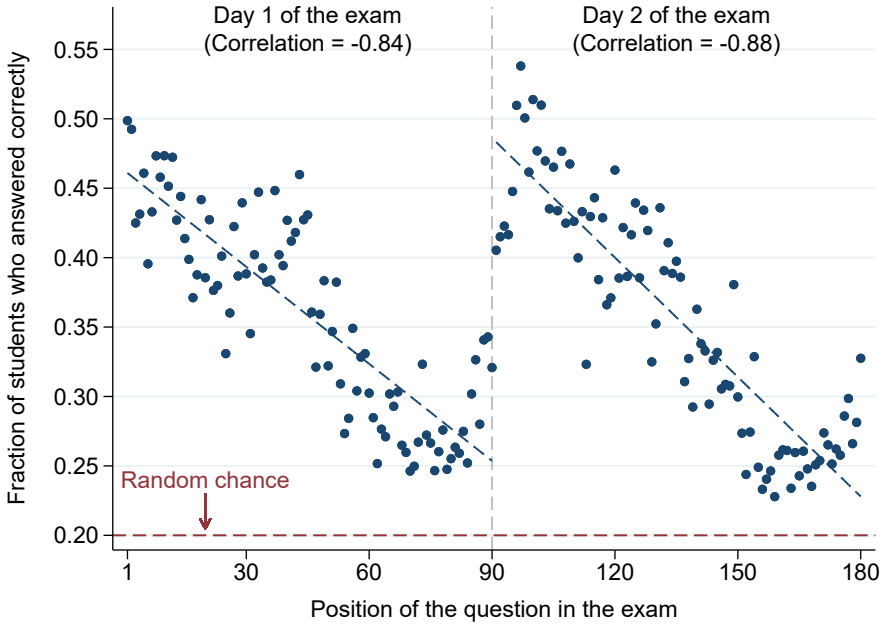
The ability-endurance score decomposition developed in this paper generates directions for future research. Test scores are commonly used in economics research, for example, as measures of cognitive skills (Hanushek and Woessmann, 2008, 2012); as a "surrogate" variable to measure the impact of an intervention on long-run outcomes (Athey et al., 2019); and to measure the effectiveness of educational inputs (Chetty et al., 2011; Dobbie

[24] While research in this area is in its infancy, some examples of protocols that build cognitive endurance include mindfulness meditation (Levy et al., 2012; Goleman and Davidson, 2017), spending time engaging in cognitively-effortful activities (Brown et al., 2022), and the restriction of smartphones in learning environments (Thornton et al., 2014). Some of these protocols are already being implemented in the private sector. For example, meditation practices are commonly used among tech companies in Silicon Valley to enhance worker productivity (Shachtman, 2013).

and Fryer Jr, 2015; Angrist et al., 2016). The decomposition allows researchers to explore the role of cognitive endurance in these and other applications. For instance, researchers can use conventional value-added methods to identify teachers who might be particularly effective at building cognitive endurance.

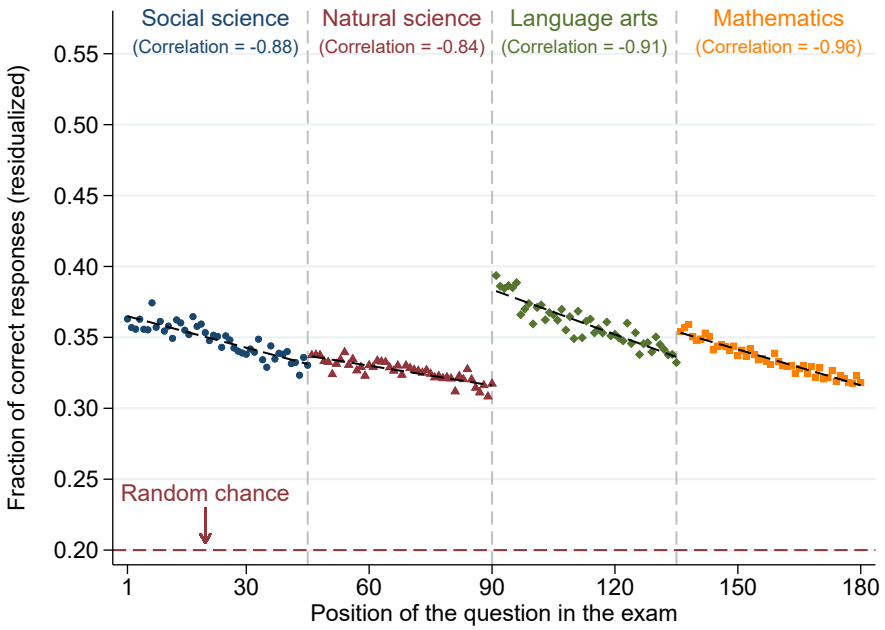
Figures and Tables

Figure 1: Average student performance over the course of the ENEM



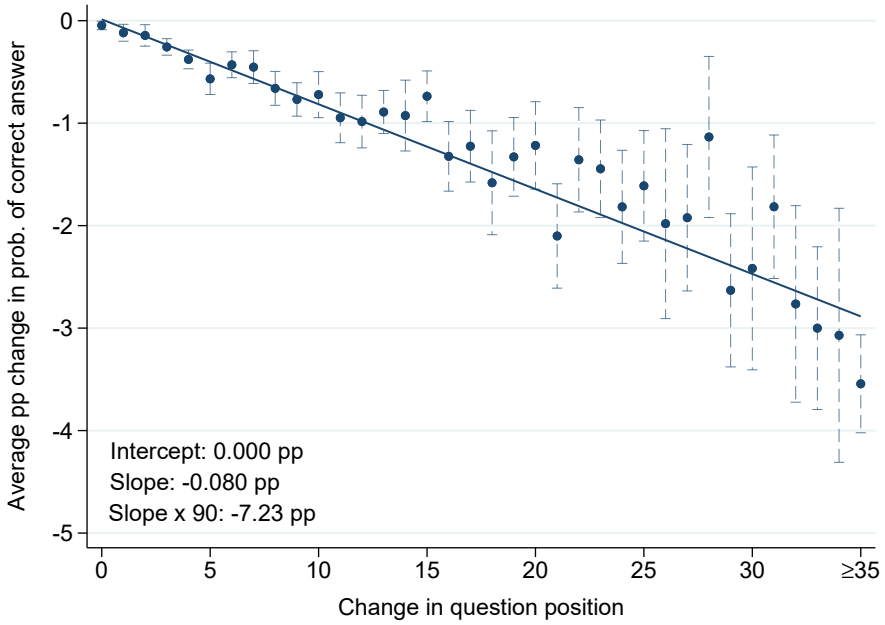
Notes: This figure shows student performance over the course of each testing day in the ENEM. The y-axis displays the fraction of students who correctly responded to each question, averaged across all years in my sample. The x-axis displays the position of each question in the exam. The dashed lines are predicted values from a linear regression estimated separately for each testing day. The horizontal red dashed line shows the expected performance if students randomly guessed the answer to each question.

Figure 2: Performance residuals after controlling for question difficulty



Notes: This figure shows student performance over the course of each testing day after removing the influence of question difficulty on performance. The y-axis displays the residuals of a regression of (i) $\bar{C}_{jb}$, the fraction of students who correctly answered question j in booklet b on (ii) Difficulty$_j$, a position-adjusted measure of question difficulty (adding back the sample mean to facilitate interpretation of units). The x-axis displays the position of each question in the exam. Marker colors denote each academic subject tested. Appendix D describes how I construct the measure of question difficulty. The dashed lines are predicted values from a linear regression estimated separately for each academic subject. The horizontal red dashed line shows the expected performance if students randomly guessed the answer to each question.

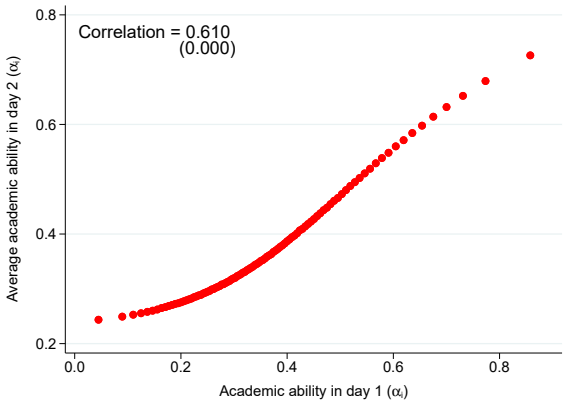
Figure 3: The effect of an increase in the order of a given question on student performance



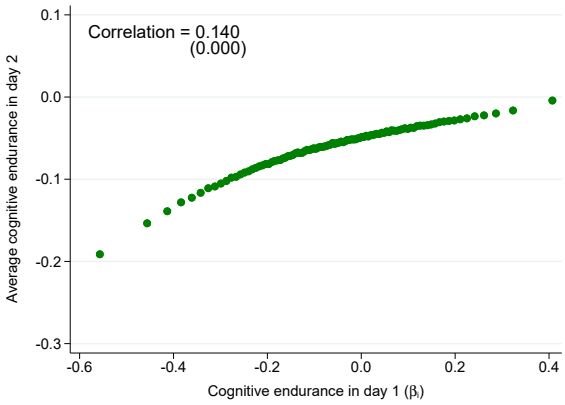
Notes: This figure shows estimates of the impact of an increase in the order of a given question on the fraction of students who correcly answer the question. The y-axis plots the average change (in percentage points) in the fraction of students who correctly respond to a question. The x-axis displays changes in a question position between each possible booklet pair. See Appendix Figure A2, Panel A for a histogram of the values in the x-axis. To construct this figure, I first compute the change in student performance and the distance in a question's position between each possible booklet pair. Then, I calculate the average change in performance for each observed distance. The solid line denotes predicted values from a linear regression estimated on the plotted points, using as weights the number of questions used to estimate each point. The vertical dashed lines denote 95% confidence intervals, estimated with heteroskedasticity-robust standard errors.

Figure 4: The temporal stability of ability and endurance estimates

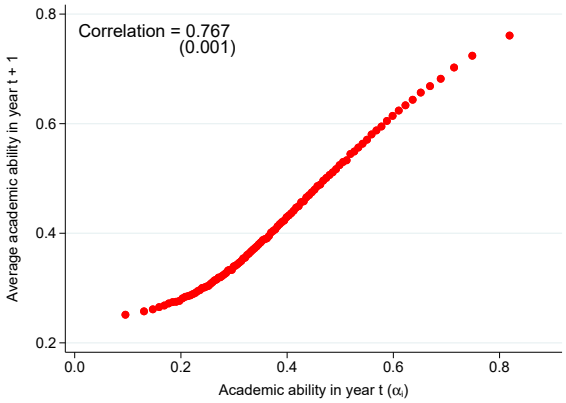
Panel A. Ability in day d and day $d+1$



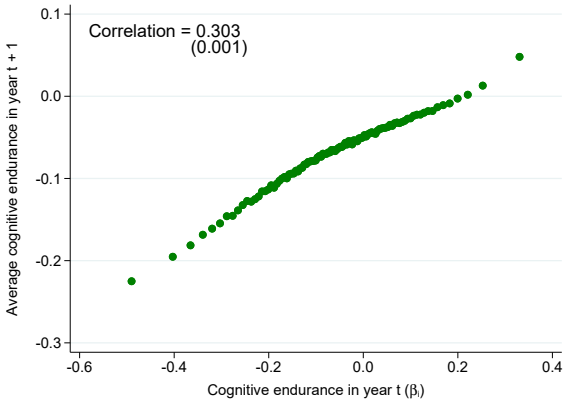
Panel B. Endurance in day d and day $d+1$



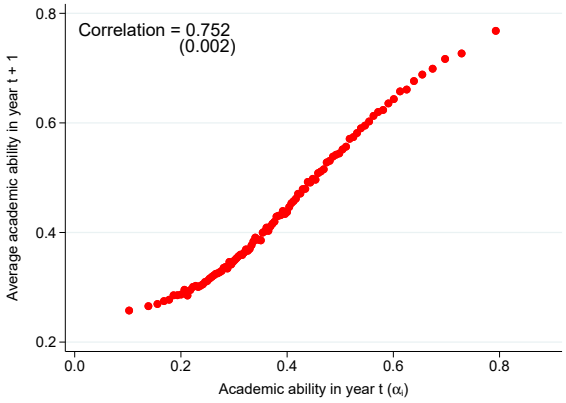
Panel C. Ability in year t and year $t+1$



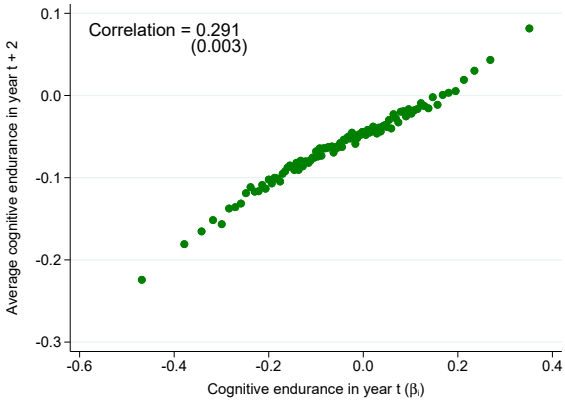
Panel D. Endurance in year t and year $t+1$



Panel E. Ability in year t and year $t+2$

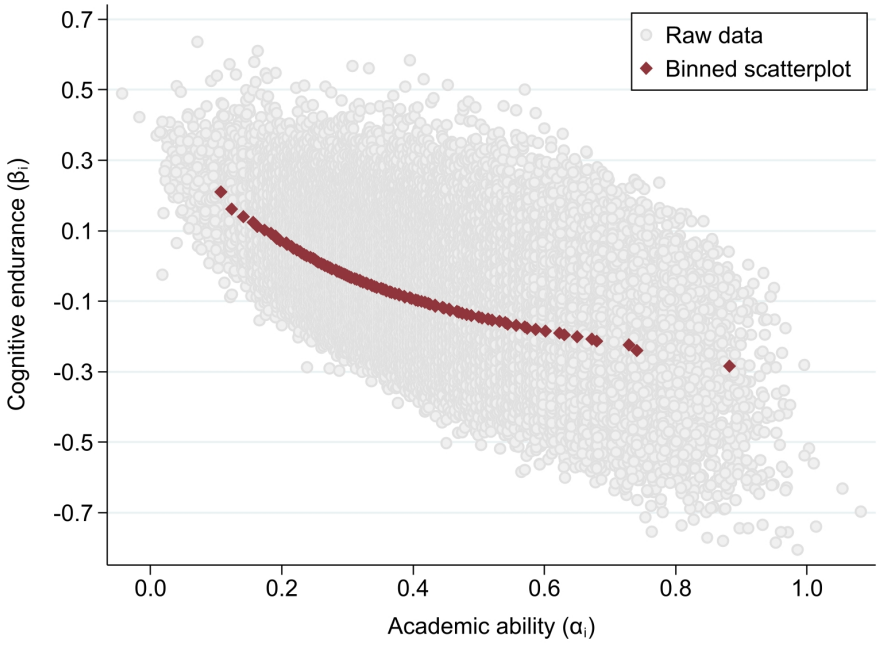


Panel F. Endurance in year t and year $t+2$



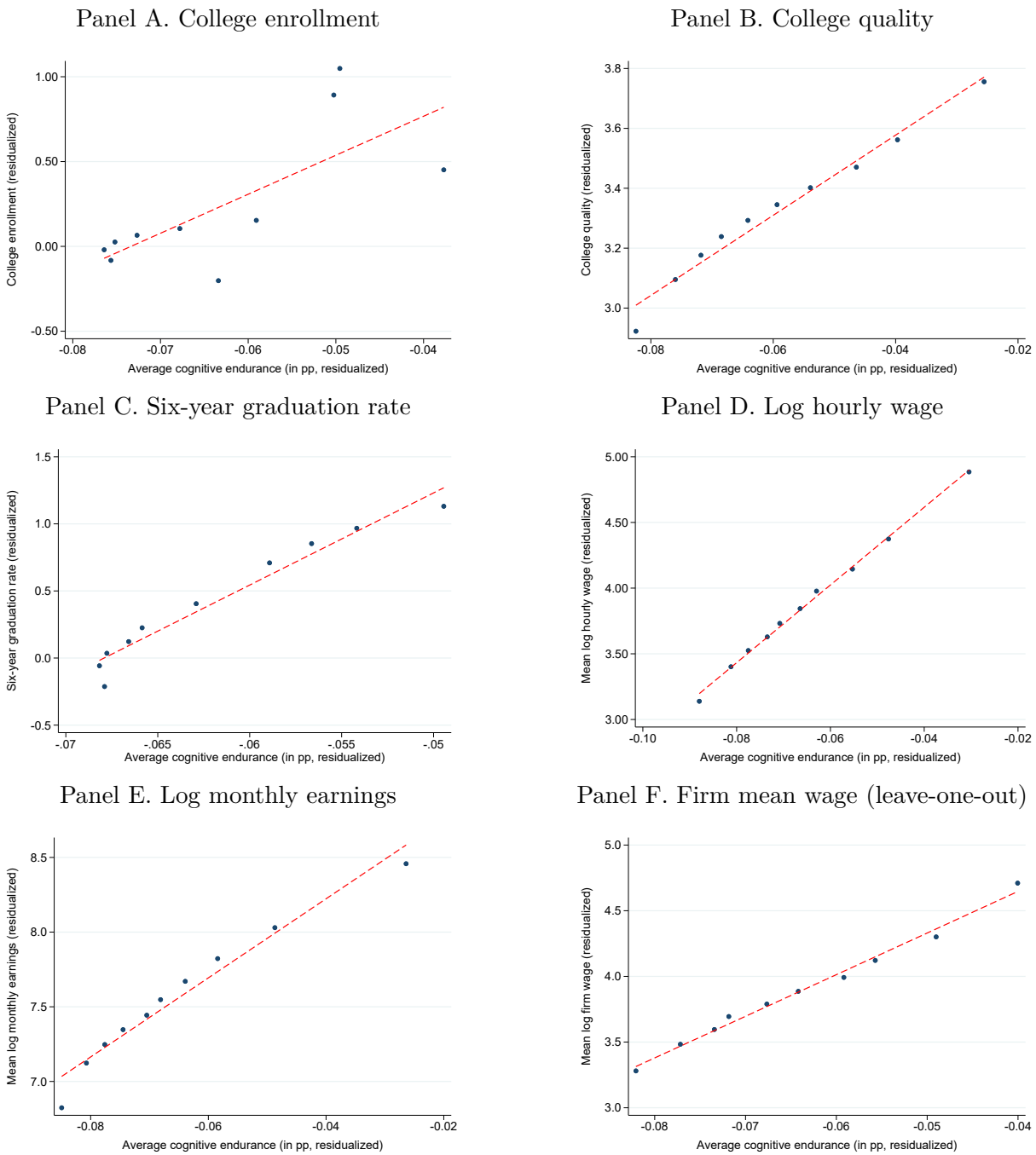
Notes: This figure shows the correlation between the measures of academic ability and cognitive endurance measured at two different points in time. Each panel shows a binned scatterplot plotting the estimates of ability/endurance at two different times. To construct this figure, I first divide students into 100 equally-sized bins based on their ability/endurance at time t. Then, I calculate the average ability/endurance at time $t' > t$ for students in each bin. The panel title indicates the two time periods in which I measure ability and endurance.

Figure 5: Joint distribution of ability and endurance estimates



Notes: This figure shows estimates of the relationship between academic ability and cognitive endurance. Gray circles display a scatterplot of $\hat{\beta}_i$ against $\hat{\alpha}_i$ for a randomly-selected one percent of my sample. The red diamonds show a binned scatterplot of average endurance as a function of ability. To construct the binned scatterplot, I first divide students into 100 equally-sized bins based on their ability. Then, I calculate the average endurance for students in each bin. Finally, I plot average endurance against ability in each bin.

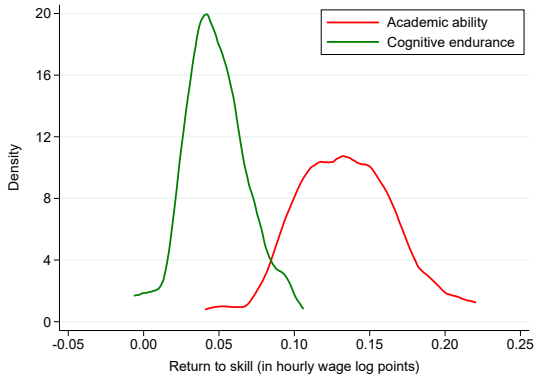
Figure 6: The relationship between cognitive endurance and long-run outcomes



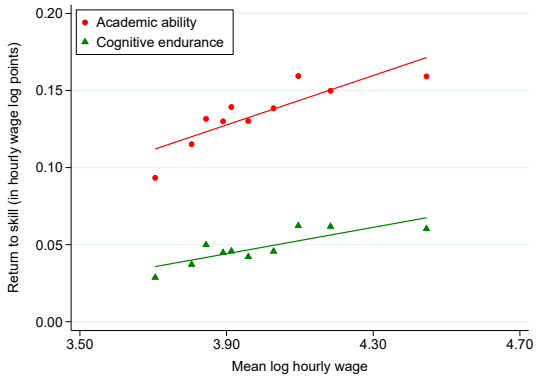
Notes: This figure shows the relationship between cognitive endurance and selected college and labor-market outcomes. Each panel shows a binned scatterplot plotting the average value of the outcome (y-axis) against cognitive endurance (x-axis). To construct this figure, I first residualize cognitive endurance and each outcome on student-level characteristics and academic ability. I add back the unconditional sample mean to facilitate interpretation. Then, I divide students into 10 equally-sized bins (deciles) based on their residualized endurance and plot the average outcome for students of each bin. The red dashed lines are predicted values from a linear regression on the plotted points. Each panel shows the results for the outcome listed in the panel title. See Section 2.4 for variable definitions.

Figure 7: Heterogeneity in the wage return to ability and cognitive endurance

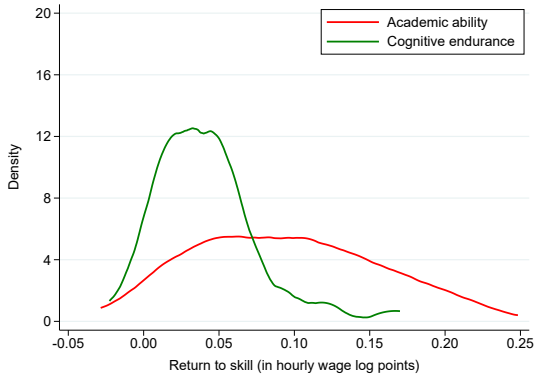
Panel A. Distribution of wage returns across college degrees



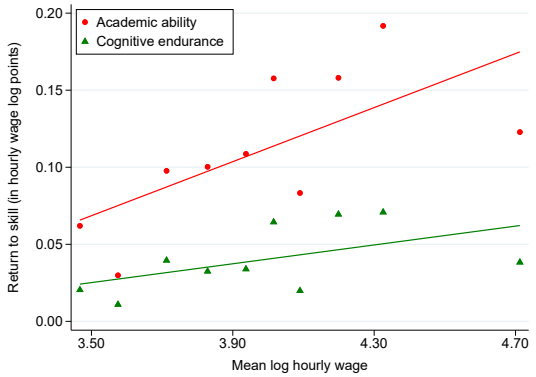
Panel B. Return to ability/endurance vs. average wage across college degrees



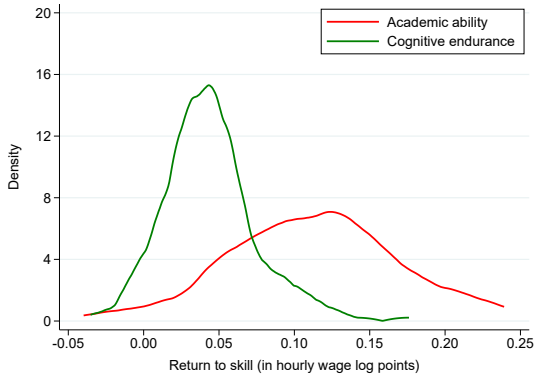
Panel C. Distribution of wage returns across occupations



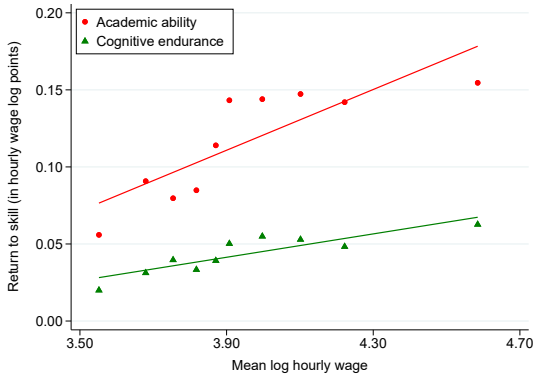
Panel D. Return to ability/endurance vs. average wage across occupations



Panel E. Distribution of wage returns across industries



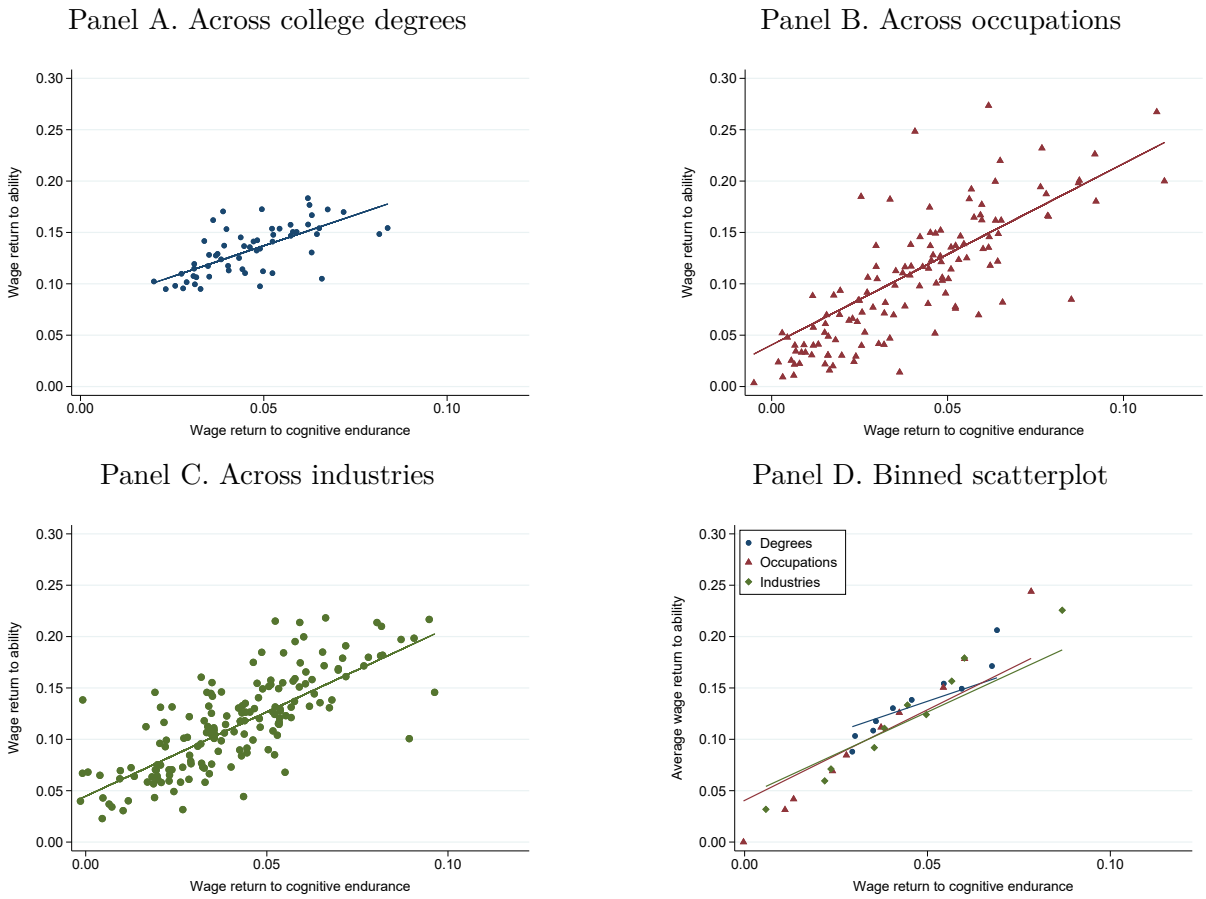
Panel F. Return to ability/endurance vs. average wage across across industries



Notes: Panels A, C, and E show nonparametric estimates of the distribution of the wage return to ability (red line) and the wage return to endurance (green line) across degrees, occupations, and industries. The wage return to ability and endurance are the coefficients ψ_A and ψ_E in equation (8) using log hourly wage as outcome, estimated separately for each degree, occupation, and industry. The figure excludes outliers (i.e., estimates of the returns below -0.05 or above 0.35).

Panels B, D, and F display a series of binned scatterplots plotting the wage return to ability/endurance (y-axis) against the mean hourly wage in bins (x-axis). To construct this figure, I first divide degrees, occupations, and industries into 10 equally-sized bins based on their mean wage. Then, I estimate the average return to ability/endurance in each bin. Finally, I plot the average return to ability/endurance against the mean wage in each bin.

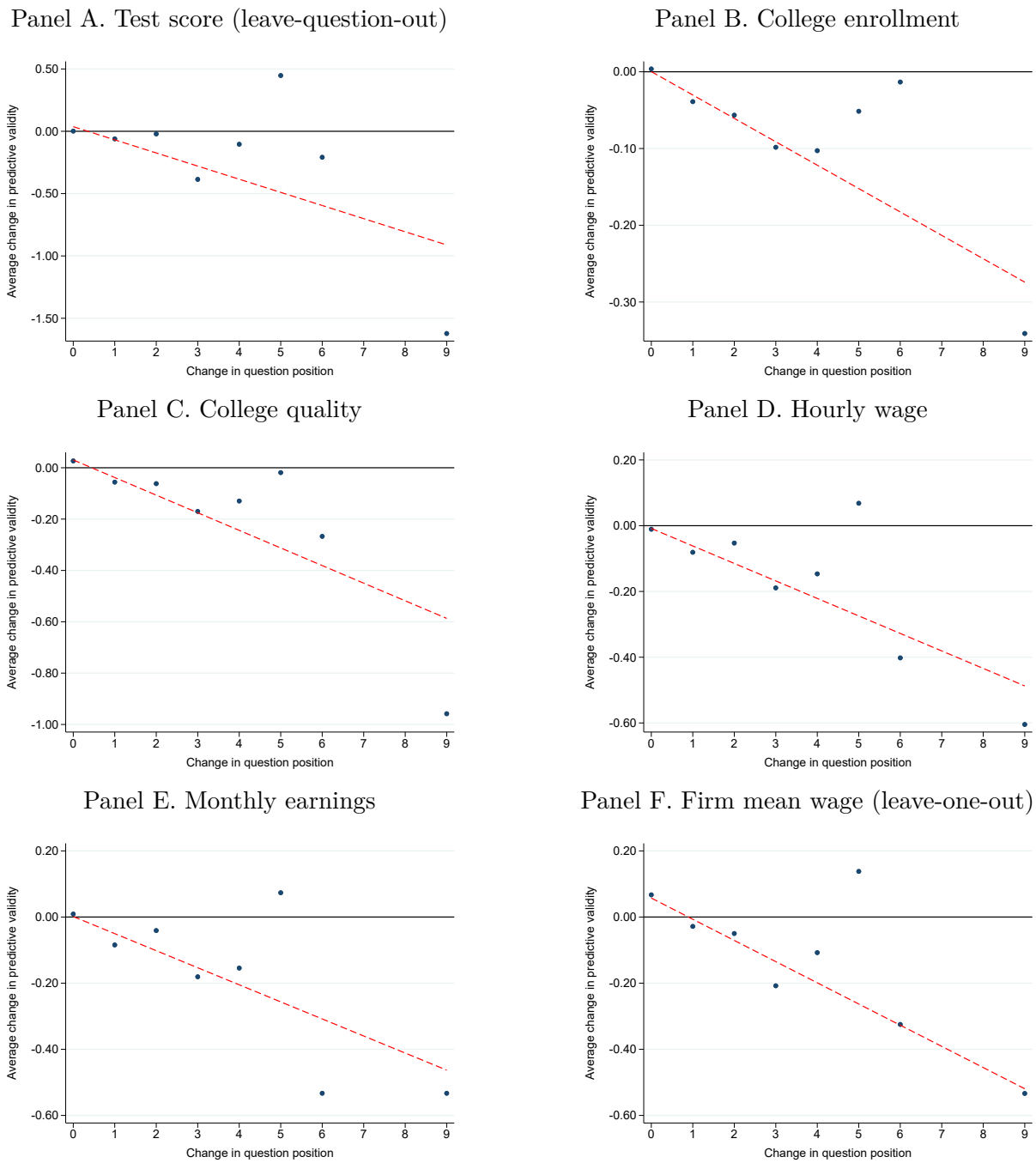
Figure 8: The relationship between the wage return to ability and endurance



Notes: This figure shows the relationship between the wage return to ability (y-axis) against the wage return to endurance (x-axis). Panels A–C show scatterplots of the wage return to ability in a given college degree (Panel A), occupation (Panel B), and industry (Panel C), against the wage return to endurance. The scatterplots exclude outliers (wage returns in the bottom 5% or top 5% of the distribution). The solid lines denote predicted values from linear regressions estimated on the microdata (including all observations).

Panel D shows a binned scatterplot plotting the mean wage return to ability against the wage return to endurance. To construct this figure, I first divide degrees (blue circles), occupations (red triangles), and industries (green diamonds) into 10 equally-sized bins based on their wage return to endurance. Then, I calculate the average wage return to ability in each bin, using the number of individuals in each bin as weights. The solid lines denote predicted values from linear regressions estimated on the plotted points.

Figure 9: Change in a question's position and change in predictive validity



Notes: This figure displays estimates of the effect of an increase in the order of a given question on the question's predictive validity. Each panel shows a binned scatterplot plotting the average change in the predictive validity of a test question on a given outcome (y-axis) against the change in the position of the question on the exam (x-axis). Each panel shows the results for the outcome listed in the panel title. See Section 2.4 for variable definitions. The red dashed lines are predicted values from a linear regression on the microdata. See Appendix Figure A2, Panel B for a histogram of the values in the x-axis.

Table 1: Summary statistics of the samples

	High-school-students sample		Retakers sample
	All	2009-2010	All
	(1)	(2)	(3)
Panel A. Demographic characteristics and race			
Age	18.204	19.151	18.062
Female	0.598	0.611	0.618
White	0.476	0.510	0.504
Black/Brown	0.505	0.450	0.483
Panel B. SES and household characteristics			
Attends a private HS	0.222	0.222	0.342
Mom completed high school	0.534	0.506	0.606
Mom completed college	0.205	0.186	0.270
Family earns above 2x M.W.	0.388	0.379	0.432
Family earns above 5x M.W.	0.062	0.071	0.087
Panel C. Exam preparation			
Took a foreign lang. course	0.241	0.269	0.263
Took a test prep course	0.119	0.167	0.160
Panel D. Fraction of correct responses			
Natural Science	0.283	0.333	0.317
Social Science	0.398	0.388	0.446
Language	0.408	0.449	0.468
Math	0.283	0.287	0.320
Average	0.343	0.364	0.388
Panel E. Geographical location			
Lives in the North	0.089	0.081	0.082
Lives in the Northeast	0.305	0.261	0.354
Lives in the Southeast	0.389	0.426	0.365
Lives in the South	0.131	0.150	0.113
Lives in the Midwest	0.086	0.081	0.085
Number of test-takers	14,941,156	1,910,502	1,519,842

Notes: This table shows summary statistics on all test-takers in the high-school-students sample (column 1), those who took the exam in 2009–2010 as high-school seniors (column 2), and students in the retakers sample (column 3). For students who took the exam multiple times, I compute the summary statistics using data from the last year in which I observe them in my sample. See Section 2.3 for sample definitions.

Table 2: The effect of question position on test performance

	Outcome: Correctly responded the question		
	(1)	(2)	(3)
Question position (normalized)	-0.214^{***}	-0.071^{***}	-0.058^{***}
	(0.013)	(0.004)	(0.002)
Constant	0.450^{***}		
	(0.008)		
N (Item−Booklets)	5,896	5,896	5,896
N (Students)	14,940,464	14,940,464	14,940,464
N (Question responses)	2,689,345,707	2,689,345,707	2,689,345,707
R−squared	0.85	0.99	0.97
Question fixed effects	No	Yes	No
Controls for question difficulty	No	No	Yes

Notes: This table displays estimates of the effect of a question position on the likelihood of correctly answering the question.

Each column displays an estimate from a different specification. Column 1 presents estimates from a bivariate regression of average student performance on question position. Column 2 presents estimates from equation (4), which includes question fixed effects. Column 3 presents estimates from equation (5), which controls for question difficulty. I normalize question position such that the first question in each testing day is equal to zero and the last question is equal to one.

Heteroskedasticity-robust standard errors clustered at the question level in parentheses. ***, ** and * denote significance at 10%, 5% and 1% levels, respectively.

Table 3: The effect of academic ability and cognitive endurance on long-run outcomes

Panel A. College outcomes

	Dependent variable					
	Enrolled college (1)	College quality (2)	Degree quality (3)	1st-year credits (4)	Grad. rate (5)	Time to grad. (6)
Test score	0.088***	0.082***	0.117***	0.014***	0.060***	−0.119***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.002)
Endurance	0.032***	0.030***	0.051***	0.006***	0.026***	−0.048***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Ability	0.102***	0.095***	0.140***	0.016***	0.072***	−0.140***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.002)
Ratio coef.	0.310***	0.319***	0.365***	0.358***	0.361***	0.342***
	(0.002)	(0.001)	(0.001)	(0.004)	(0.003)	(0.005)
Mean DV	0.244	3.326	3.244	0.158	0.418	3.817
N	2,501,519	1,800,546	1,768,707	1,124,972	1,472,916	793,822

Panel B. Labor-market outcomes

	Dependent variable					
	Formal sector (1)	Hourly wage (2)	Monthly earnings (3)	Firm wage (4)	Occup. wage (5)	Industry wage (6)
Test score	0.001***	0.129***	0.111***	0.092***	0.041***	0.013***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
Endurance	0.000***	0.054***	0.052***	0.036***	0.017***	0.004***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)
Ability	0.002***	0.154***	0.135***	0.108***	0.049***	0.014***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
Ratio coef.	0.276***	0.351***	0.387***	0.330***	0.346***	0.255***
	(0.015)	(0.003)	(0.003)	(0.004)	(0.006)	(0.010)
Mean DV	0.326	3.865	7.551	3.885	3.886	3.858
N	2,523,029	818,590	818,590	692,880	818,374	818,590

Notes: This table displays estimates of the relationship between ability/endurance and college outcomes (Panel A) and labor market-outcomes (Panel B).

The first row of each panel shows estimates of the association between test scores and the outcome listed in the column header (coefficient ψ_T in equation (7)). The following rows show estimates of the association between ability and cognitive endurance and a given outcome (coefficients ψ_A and ψ_E in equation (8)). All regressions control for age, gender, race, high school type, parental income, cohort fixed effects, and municipality fixed effects. In addition to the baseline controls, the regressions in Panel A, columns 4–6, include college-degree fixed effects to remove the influence of a student's program choice, while the regressions in Panel B control for potential years of experience and years of education. Heteroskedasticity-robust standard errors clustered at the individual level in parentheses. See Section 2.4 for outcome definitions.

The third-to-last row in each panel shows the ratio between the predicted effect of academic ability and the effect of cognitive endurance on a given outcome. Standard errors estimated through the delta method in parentheses. ***, ** and * denote significance at 10%, 5% and 1% levels.

Table 4: Degrees, occupations, and industries with the largest return to endurance

	Return ability (1)	Return endur. (2)	Ratio returns (3)	Wage pctil. (4)	Sample size (5)
Panel A. Top five degrees					
1. Aeronautics and related degrees	0.173	0.106	0.613	69.7	670
	(0.026)	(0.015)	(0.077)		
2. Music and performing arts	0.221	0.093	0.420	59.3	502
	(0.060)	(0.035)	(0.110)		
3. Religious studies	0.186	0.088	0.471	49.0	356
	(0.047)	(0.031)	(0.097)		
4. History and archeology	0.211	0.087	0.414	60.0	988
	(0.043)	(0.022)	(0.064)		
5. Forestry engineering	0.154	0.084	0.543	52.3	397
	(0.045)	(0.024)	(0.115)		
Panel B. Top five occupations					
1. Public tax auditors	0.454	0.168	0.369	49.2	255
	(0.106)	(0.057)	(0.084)		
2. Professionals in air	0.278	0.114	0.412	88.3	347
navigation, sea and fluvial	(0.049)	(0.028)	(0.072)		
3. Technicians in operation of radio,	0.200	0.112	0.558	66.3	658
TV systems and video producers	(0.047)	(0.027)	(0.091)		
4. Plant operators in chemical,	0.267	0.109	0.409	62.8	1,221
petrochemical and related occup.	(0.031)	(0.020)	(0.057)		
5. Instrument and precision	0.180	0.092	0.511	70.8	203
equipment repairers	(0.068)	(0.040)	(0.194)		
Panel C. Top five industries					
1. Oil extraction and	0.276	0.135	0.488	91.5	948
related services	(0.034)	(0.021)	(0.052)		
2. Financial intermediation and	0.215	0.087	0.402	54.7	6,177
and insurance (aux. services)	(0.013)	(0.007)	(0.024)		
3. Research and development	0.198	0.084	0.426	73.7	1,323
	(0.024)	(0.015)	(0.052)		
4. Electricity, gas and hot water	0.223	0.083	0.373	78.1	1,966
	(0.020)	(0.011)	(0.036)		
5. Manufacture of office machinery	0.132	0.073	0.553	54.7	920
	(0.033)	(0.019)	(0.095)		

Notes: This table lists the top five 3-digit academic degrees (Panel A), 3-digit occupations (Panel B), and 2-digit industries (Panel C) with the highest wage return to cognitive endurance (column 2).

Column 1 shows the wage return to ability. Column 3 shows the ratio between the wage return to endurance and the wage return to ability. Column 4 shows the average wage percentile of workers in each degree, occupation, or industry. Column 5 shows the sample size used to estimate each wage return.

The wage return to ability and endurance are the coefficients ψ_A and ψ_E in equation (8) using as outcome log hourly wage, estimated separately for each degree, occupation, and industry.

Heteroskedasticity-robust standard errors clustered at the individual level in parentheses.

Table 5: The contribution of gaps in ability and endurance to test-score gaps

	Gap between				
	Male / Female	White / Non-white	Priv HS / Public HS	Mom coll / No coll	High-inc / Low-inc
	(1)	(2)	(3)	(4)	(5)
Panel A. Difference in average test score					
Test-score gap	0.026***	0.057***	0.130***	0.098***	0.192***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Panel B. Contribution of gaps in ability and endurance to test-score gaps					
Ability gap	0.030***	0.056***	0.127***	0.095***	0.188***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Endurance gap	0.017***	0.016***	0.038***	0.026***	0.063***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Panel C. Impact of a reform that halves the exam length on test-score gaps					
P.p. change gap	-0.008***	-0.008***	-0.019***	-0.013***	-0.031***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Pct. change gap	-0.322***	-0.137***	-0.144***	-0.130***	-0.163***
	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
N (Students)	14,941,097	14,565,550	9,924,652	14,290,759	9,996,959

Notes: This table shows test-score gaps in the ENEM and the contribution of differences in ability and endurance to those gaps.

Each column shows the result for a different test-score gap. Column 1 shows gaps between male and female students. Column 2 shows gaps between white and non-white (Black, Brown, and Indigenous) students. Column 3 shows gaps between students enrolled in a private high school and public high school. Column 4 shows gaps between students with a college-educated mother and non-college-educated mother. Column 5 shows gaps between students in households in the top 30% and bottom 30% of the income distribution.

Panel A shows the average test score difference between the two groups displayed in the column header, $\mathbb{E}[\text{TestScore}_i|X_i = 1] - \mathbb{E}[\text{TestScore}_i|X_i = 0]$.

Panel B shows the contribution of differences in ability and differences in endurance to the test-score gap. The ability gap is the average difference in ability, controlling for endurance, $\mathbb{E}[\hat{\alpha}_i|X_i = 1, \hat{\beta}_i] - \mathbb{E}[\hat{\alpha}_i|X_i = 0, \hat{\beta}_i]$. The endurance gap is the average difference in endurance, controlling for ability and scaled by the average question position, $\Big(\mathbb{E}[\hat{\beta}_i|X_i = 1, \hat{\alpha}_i] - \mathbb{E}[\hat{\beta}_i|X_i = 0, \hat{\alpha}_i]\Big) \times \overline{\text{Position}}$.

Panel C shows estimates of the impact of a reform that changes the length of the exam from $\overline{\text{Position}}$ to $\overline{\text{Position}}/2$. The first row shows the percentage point change in the test-score gap due to the reform, which is equal to $-\Big(\mathbb{E}[\hat{\beta}_i|X_i = 1, \hat{\alpha}_i] - \mathbb{E}[\hat{\beta}_i|X_i = 0, \hat{\alpha}_i]\Big) \times \overline{\text{Position}}/2$. The second row shows the percentage change in the test-score gap, which equals the percentage point change in the gap divided by the pre-reform test-score gap (shown in Panel A). Standard errors estimated through the delta method in parentheses. ***, ** and * denote significance at 10%, 5% and 1% levels, respectively.

Table 6: The effect of an exam reform that halves the exam length on its predictive validity

	Outcome: Predictive validity of question j for							
	Test score (1)	College enrol. (2)	College quality (3)	Degree progress (4)	Grad. rate (5)	Hourly wage (6)	Monthly earnings (7)	Firm wage (8)
Panel A. Average predictive validity								
Constant	0.285***	0.057***	0.109***	0.003***	0.016***	0.106***	0.101***	0.084***
	(0.007)	(0.002)	(0.003)	(0.001)	(0.001)	(0.003)	(0.003)	(0.002)
Panel B. Effect of the exam reform								
Change in Pred. Val.	0.115*	0.055***	0.099***	−0.005**	0.003	0.084**	0.077**	0.072**
	(0.069)	(0.018)	(0.032)	(0.002)	(0.013)	(0.034)	(0.032)	(0.030)
Chg. Val./Mean	0.404***	0.952***	0.911***	−1.500***	0.201	0.796***	0.757***	0.855***
	(0.110)	(0.191)	(0.144)	(0.494)	(0.662)	(0.193)	(0.194)	(0.239)
N (Item−Booklets)	1,416	1,416	1,416	700	1,416	1,416	1,416	1,416

Notes: This table displays the estimated effect of an exam reform that changes the exam length from $\overline{\text{Position}}$ to $\overline{\text{Position}}/2$ on the predictive validity of the exam questions for long-run outcomes.

Each column displays the estimates of equation (10) for a different outcome. In Panel A, the regression only includes a constant. In Panel B, the regression includes question fixed effects. I the coefficients so that they can be interpreted as the effect of decreasing the exam length by half. See Section 2.4 for outcome definitions.

Heteroskedasticity-robust standard errors clustered at the question level in parentheses. ***, ** and * denote significance at 10%, 5% and 1% levels, respectively.

References

Abraham, K. G. and Mallatt, J. (2022). Measuring human capital. *Journal of Economic Perspectives*, 36(3):103–30.

Acemoglu, D. and Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. In *Handbook of labor economics*, volume 4, pages 1043–1171. Elsevier.

Ackerman, P. L. (2011). *Cognitive fatigue: Multidisciplinary perspectives on current research and future applications.* American Psychological Association.

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

Almlund, M., Duckworth, A. L., Heckman, J., and Kautz, T. (2011). Personality psychology and economics. In Hanushek, E. A., Machin, S., and Woessmann, L., editors, *Handbook of the Economics of Education*, volume 4 of *Handbook of The Economics of Education*, pages 1–181. DOI: 10.1016/B978-0-444-53444-6.00001-8.

Angrist, J. D., Cohodes, S. R., Dynarski, S. M., Pathak, P. A., and Walters, C. R. (2016). Stand and deliver: Effects of boston's charter high schools on college preparation, entry, and choice. *Journal of Labor Economics*, 34(2):275–318.

Angrist, J. D., Hull, P. D., Pathak, P. A., and Walters, C. R. (2017). Leveraging lotteries for school value-added: Testing and estimation. *The Quarterly Journal of Economics*, 132(2):871–919.

Angrist, J. D. and 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.

Anusic, I. and Schimmack, U. (2016). Stability and change of personality traits, self-esteem, and well-being: Introducing the meta-analytic stability and change model of retest correlations. *Journal of Personality and Social Psychology*, 110(5):766.

Archsmith, J., Heyes, A., Neidell, M., and Sampat, B. (2021). The Dynamics of Inattention in the (Baseball) Field. National Bureau of Economic Research.

Athey, S., Chetty, R., Imbens, G. W., and Kang, H. (2019). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely.

Balart, P. and Oosterveen, M. (2019). Females show more sustained performance during test-taking than males. *Nature Communications*, 10(3798). DOI: 10.1038/s41467-019-11691-y.

Balart, P., Oosterveen, M., and Webbink, D. (2018). Test scores, noncognitive skills and economic growth. *Economics of Education Review*, 63:134–153. DOI: 10.1016/j.econedurev.2017.12.004.

Baumeister, R. F. (2002). Yielding to Temptation: SelfControl Failure, Impulsive Purchasing, and Consumer Behavior. *Journal of Consumer Research*, 28(4):670–676. DOI: 10.1086/338209.

Becker, G. S. (1962). Investment in human capital: A theoretical analysis. *Journal of political economy*, 70(5, Part 2):9–49.

Benson, K., Flory, K., Humphreys, K. L., and Lee, S. S. (2015). Misuse of stimulant medication among college students: a comprehensive review and meta-analysis. *Clinical child and family psychology review*, 18(1):50–76.

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

Bishop, J. H. (1989). Is the test score decline responsible for the productivity growth decline? *The American Economic Review*, pages 178–197.

Borghans, L., Duckworth, A. L., Heckman, J. J., and Weel, B. t. (2008). The Economics and Psychology of Personality Traits. *Journal of Human Resources*, 43(4):972–1059. DOI: 10.3368/jhr.43.4.972.

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

Bowles, S., Gintis, H., and Osborne, M. (2001). The Determinants of Earnings: A Behavioral Approach. *Journal of Economic Literature*, 39(4):1137–1176. DOI: 10.1257/jel.39.4.1137.

Brown, C., Kaur, S., Kingdon, G., and Schofield, H. (2022). Cognitive endurance as human capital. Working Paper.

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

Buser, T., Niederle, M., and Oosterbeek, H. (2021). Can competitiveness predict education and labor market outcomes? Evidence from incentivized choice and survey measures. *NBER Working Paper 28916*. DOI: 10.3386/w28916.

Card, D. (1999). The causal effect of education on earnings. *Handbook of labor economics*, 3:1801–1863.

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica*, 69(5):1127–1160.

Card, D., Cardoso, A. R., Heining, J., and Kline, P. (2018). Firms and labor market inequality: Evidence and some theory. *Journal of Labor Economics*, 36(S1):S13–S70.

Card, D. and Rothstein, J. (2007). Racial segregation and the black–white test score gap. *Journal of Public Economics*, 91(11-12):2158–2184.

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

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., and Yagan, D. (2011). How does your kindergarten classroom affect your earnings? evidence from project star. *The Quarterly journal of economics*, 126(4):1593–1660.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014a). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American economic review*, 104(9):2593–2632.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014b). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American economic review*, 104(9):2633–79.

Chetty, R., Friedman, J. N., Saez, E., Turner, N., and Yagan, D. (2020). Income segregation and intergenerational mobility across colleges in the United States. *The Quarterly Journal of Economics*, 135(3):1567–1633.

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

Debeer, D., Buchholz, J., Hartig, J., and Janssen, R. (2014). Student, School, and Country Differences in Sustained Test-Taking Effort in the 2009 PISA Reading Assessment. *Journal of Educational and Behavioral Statistics*, 39(6):502–523.

Debeer, D. and Janssen, R. (2013). Modeling item-position effects within an irt framework. *Journal of Educational Measurement*, 50(2):164–185.

Deming, D. J. (2017). The Growing Importance of Social Skills in the Labor Market. *The Quarterly Journal of Economics*, 132(4):1593–1640. DOI: 10.1093/qje/qjx022.

Deming, D. J. (2022). Four facts about human capital. *Journal of Economic Perspectives*, 36(3):75–102.

Dobbie, W. and Fryer Jr, R. G. (2015). The medium-term impacts of high-achieving charter schools. *Journal of Political Economy*, 123(5):985–1037.

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., and Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences*, 108(19):7716–7720.

Edin, P.-A., Fredriksson, P., Nybom, M., and Ockert, B. (2022). The rising return to noncognitive skill. *American Economic Journal: Applied Economics*, 14(2):78–100.

Fe, E., Gill, D., and Prowse, V. (2022). Cognitive skills, strategic sophistication, and life outcomes. *Journal of Political Economy (forthcoming)*.

Frankel, A. and Kartik, N. (2019). Muddled information. *Journal of Political Economy*, 127(4):1739–1776.

Fryer Jr, R. G. and Levitt, S. D. (2006). The black-white test score gap through third grade. *American law and economics review*, 8(2):249–281.

Gneezy, U., List, J. A., Livingston, J. A., Qin, X., Sadoff, S., and Xu, Y. (2019). Measuring Success in Education: The Role of Effort on the Test Itself. *American Economic Review: Insights*, 1(3):291–308.

Goleman, D. (2013). *Focus: The Hidden Driver of Excellence*. A&C Black.

Goleman, D. and Davidson, R. J. (2017). *Altered traits: Science reveals how meditation changes your mind, brain, and body*. Penguin.

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

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

Griliches, Z. (1977). Estimating the returns to schooling: Some econometric problems. *Econometrica: Journal of the Econometric Society*, pages 1–22.

Gronqvist, E., Ockert, B., and Vlachos, J. (2017). The Intergenerational Transmission of Cognitive and Noncognitive Abilities. *Journal of Human Resources*, 52(4):887–918. DOI: 10.3368/jhr.52.4.0115-6882R1.

Hanushek, E. A. and Woessmann, L. (2008). The role of cognitive skills in economic development. *Journal of economic literature*, 46(3):607–68.

Hanushek, E. A. and Woessmann, L. (2012). Do better schools lead to more growth? cognitive skills, economic outcomes, and causation. *Journal of economic growth*, 17(4):267–321.

Heckman, J. J., Stixrud, J., and Urzua, S. (2006). The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior. *Journal of Labor Economics*, 24(3):411–482. DOI: 10.1086/504455.

Hermo, S., Paallysaho, M., Seim, D., and Shapiro, J. M. (2022). Labor Market Returns and the Evolution of Cognitive Skills: Theory and Evidence*. *The Quarterly Journal of Economics*. qjac022.

Hirshleifer, D., Levi, Y., Lourie, B., and Teoh, S. H. (2019). Decision fatigue and heuristic analyst forecasts. *Journal of Financial Economics*, 133(1):83–98.

Hopkins, K. D. and Bracht, G. H. (1975). Ten-year stability of verbal and nonverbal iq scores. *American Educational Research Journal*, 12(4):469–477.

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

Jackson, C. K. (2018). What do test scores miss? the importance of teacher effects on non–test score outcomes. *Journal of Political Economy*, 126(5):2072–2107.

Kautz, T., Heckman, J. J., Diris, R., Ter Weel, B., and Borghans, L. (2014). Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success. National Bureau of Economic Research.

Keynes, J. M. (1956). Newton the man. In Newman, J., editor, *The world of mathematics*, volume I of *The world of mathematics*.

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

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., and Barbuti, S. M. (2008). Validity of the sat® for predicting first-year college grade point average. research report no. 2008-5. *College Board*.

Kremer, M. (1993). The o-ring theory of economic development. *The Quarterly Journal of Economics*, 108(3):551–575.

Lampkin, C. (2012). Dietary Habits - An AARP Bulletin Poll. page 17.

Levy, D. M., Wobbrock, J. O., Kaszniak, A. W., and Ostergren, M. (2012). The effects of mindfulness meditation training on multitasking in a high-stress information environment. In *Proceedings of Graphics Interface 2012*, pages 45–52.

Lim, J. and Dinges, D. F. (2008). Sleep deprivation and vigilant attention. *Annals of the New York Academy of Sciences*, 1129:305–322. DOI: 10.1196/annals.1417.002.

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

Lindqvist, E. and Vestman, R. (2011). The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment. *American Economic Journal: Applied Economics*, 3(1):101–128. DOI: 10.1257/app.3.1.101.

Lira, B., O'Brien, J. M., Peña, P. A., Galla, B. M., D'Mello, S., Yeager, D. S., Defnet, A., Kautz, T., Munkacsy, K., and Duckworth, A. L. (2022). Large studies reveal how reference bias limits policy applications of self-report measures. *Scientific Reports*, 12(1):1–12.

Lykken, D. T. (2005). Mental energy. *Intelligence*, 33(4):331–335.

Machado, C. and Szerman, C. (2021). Centralized college admissions and student composition. *Economics of Education Review*, 85:102184.

Mandinach, E. B., Bridgeman, B., Cahalan-Laitusis, C., and Trapani, C. (2005). The impact of extended time on sat® test performance. *ETS Research Report Series*, 2005(2):i–35.

Mata, R., Frey, R., Richter, D., Schupp, J., and Hertwig, R. (2018). Risk preference: A view from psychology. *Journal of Economic Perspectives*, 32(2):155–72.

Meier, S. and Sprenger, C. D. (2015). Temporal stability of time preferences. *Review of Economics and Statistics*, 97(2):273–286.

Miller, M. D., Linn, R. L., and Gronlund, N. E. (2009). *Measurement and Assessment in Teaching*. Merrill/Pearson. Google-Books-ID: pMJTPgAACAAJ.

Mincer, J. (1958). Investment in human capital and personal income distribution. *Journal of political economy*, 66(4):281–302.

Mullainathan, S. and Shafir, E. (2013). *Scarcity: Why having too little means so much*. Macmillan.

Newport, C. (2016). *Deep work: Rules for focused success in a distracted world*. Hachette UK.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204.

Otero, S., Barahona, N., and Dobbin, C. (2021). Affirmative action in centralized college admission systems: Evidence from Brazil. Working paper.

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

Riehl, E. (2022). Fairness in college admission exams: From test score gaps to earnings inequality. Working Paper.

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

Schuerger, J. M. and Witt, A. C. (1989). The temporal stability of individually tested intelligence. *Journal of Clinical Psychology*, 45(2):294–302.

Shachtman, N. (2013). In silicon valley, meditation is no fad. it could make your career. *Wired, June*, 18.

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

Soares, J. A. (2015). *SAT wars: The case for test-optional college admissions*. Teachers College Press.

Stango, V. and Zinman, J. (2020). Behavioral biases are temporally stable. Technical report, National Bureau of Economic Research.

Steiny Wellsjo, A. (2022). Simple actions, complex habits: Lessons from hospital hand hygiene. Working Paper.

Thornton, B., Faires, A., Robbins, M., and Rollins, E. (2014). The mere presence of a cell phone may be distracting: Implications for attention and task performance. *Social Psychology*, 45(6):479.

UNESCO (2012). International standard classification of education: Isced 2011. *Comparative Social Research*, 30.

Wooden, M. (2012). The stability of personality traits. *R. Wilkins and D. Warren, Families, Incomes and Jobs*, 7.

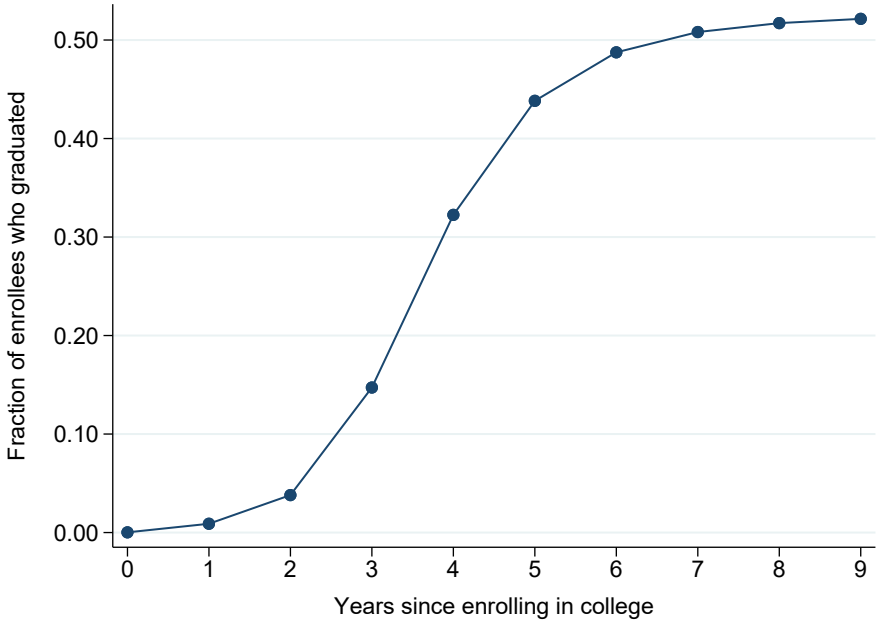
Zamarro, G., Cheng, A., Shakeel, M. D., and Hitt, C. (2018). Comparing and validating measures of non-cognitive traits: Performance task measures and self-reports from a nationally representative internet panel. *Journal of Behavioral and Experimental Economics*, 72:51–60.

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

Appendix

A Appendix Figures and Tables

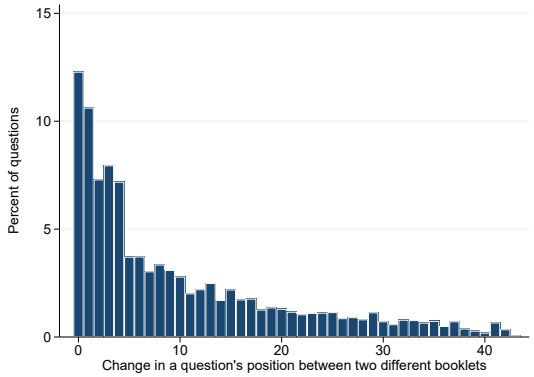
Figure A1: Fraction of students who graduate from college by years since enrollment



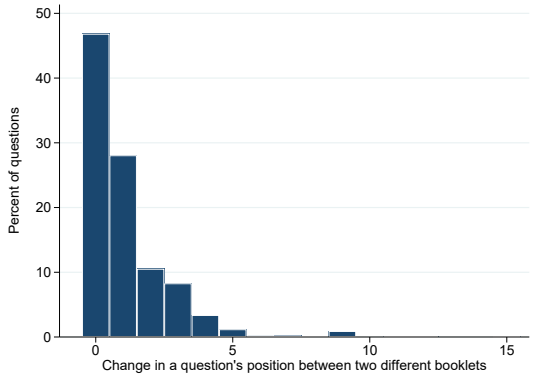
Notes: This figure shows the empirical cumulative distribution function of the graduation rate of individuals in the high-school-students sample.

Figure A2: Histogram of the change in a question's position across exam booklets

Panel A. All years (2009–2016)

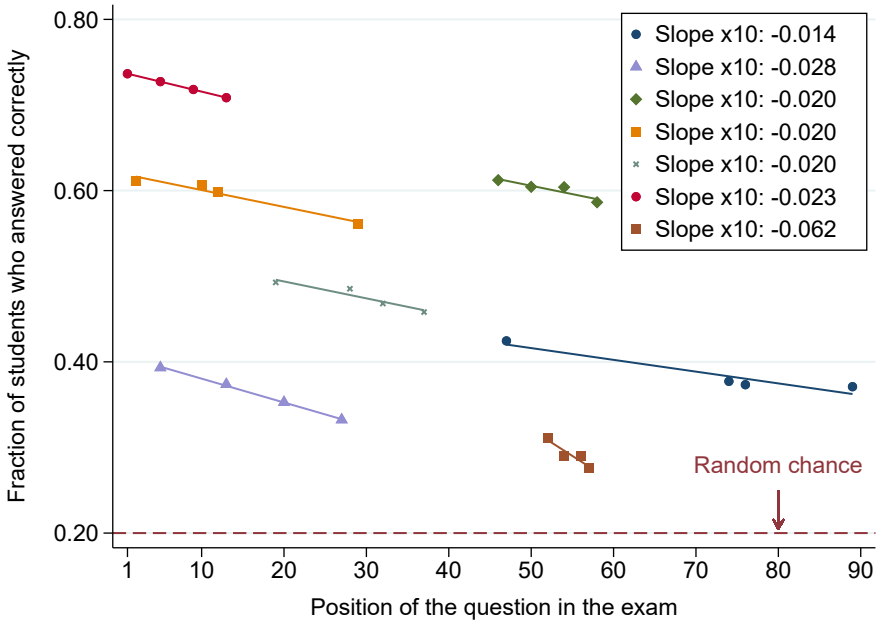


Panel B. First two cohorts (2009–2010)



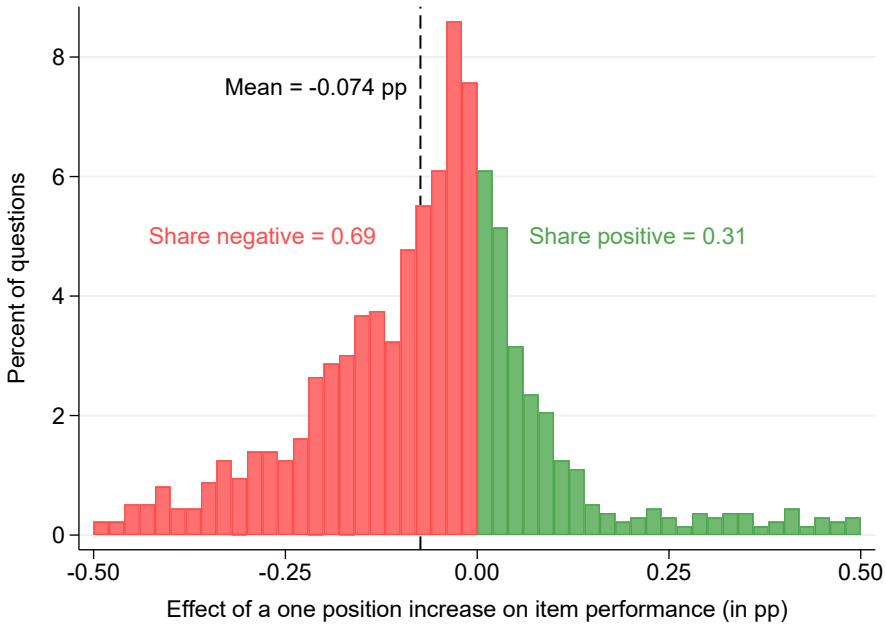
Notes: This figure shows the amount of variation available in a given question's position between different exam booklets. To construct this figure, I first calculate the difference (in absolute value) in a question's position in two exam booklets. This difference ranges from zero (if a question is in the same position in two different booklets) to 44 (if a question is in the first position of a section in one booklet and the last position of a section in another booklet). I repeat this process for each question and each possible booklet pair. The figure plots the resulting histogram of position differences.

Figure A3: Average student performance on selected questions by question position



Notes: This figure plots the fraction of correct responses on seven selected exam questions as a function of their position on the four different exam booklets. Solid lines denote predicted values from linear regressions estimated on the plotted points.

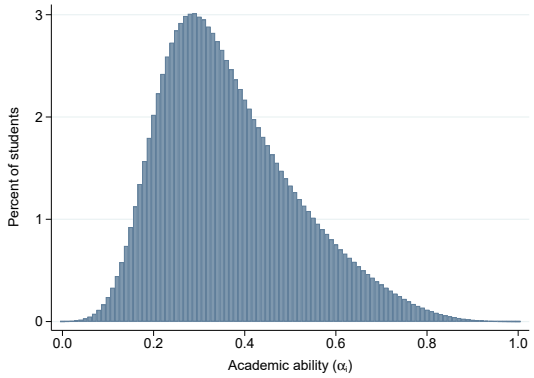
Figure A4: Histogram of question-level position effects



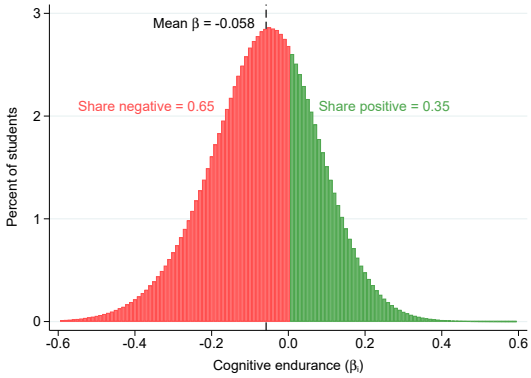
Notes: This figure plots the distribution of item-level position effects. To construct this figure, I estimate the impact of an increase in the position of a given question on student performance separately for each question. The figure displays the distribution of estimated β's (one for each item). The figure excludes outliers (i.e., questions for which the effect is below -0.50 or above 0.50 percentage points).

Figure A5: Distribution of academic ability and cognitive endurance

Panel A. Academic ability (α_i)



Panel B. Cognitive endurance (β_i)



Notes: This figure shows the distribution of my estimates of academic ability (Panel A) and cognitive endurance (Panel B) among individuals in the high-school-students sample. The measure of an individual's ability is the estimated intercept (α_i) in equation (6). The measure of an individual's cognitive endurance is the estimated slope (β_i) in equation (6).

Table A1: Summary statistics of the high-school-student sample by booklet color

		Day 1 booklet color			
	All	Yellow	Blue	Pink	White
	(1)	(2)	(3)	(4)	(5)
Panel A. Demographic characteristics and race					
Age	18.204	18.201	18.210	18.209	18.196
Female	0.598	0.595	0.600	0.595	0.599
White	0.476	0.478	0.476	0.476	0.474
Black/Brown	0.505	0.503	0.505	0.505	0.507
Panel B. Household characteristics					
Attends a private HS	0.222	0.225	0.220	0.223	0.220
Mom completed high school	0.534	0.538	0.532	0.535	0.531
Mom completed college	0.205	0.208	0.203	0.207	0.203
Family earns above 2x M.W.	0.388	0.392	0.386	0.390	0.385
Family earns above 5x M.W.	0.062	0.064	0.061	0.063	0.061
Panel C. Exam preparation					
Took a foreign lang. course	0.241	0.241	0.241	0.240	0.241
Took a test prep course	0.119	0.121	0.119	0.119	0.118
Panel D. Fraction of correct responses					
Natural Science	0.283	0.284	0.283	0.283	0.283
Social Science	0.398	0.398	0.398	0.398	0.398
Language	0.408	0.410	0.408	0.408	0.407
Math	0.283	0.284	0.283	0.283	0.282
Average	0.343	0.344	0.343	0.343	0.343
Panel E. Geographical location					
Lives in the North	0.089	0.089	0.089	0.090	0.089
Lives in the Northeast	0.305	0.305	0.303	0.306	0.305
Lives in the Southeast	0.389	0.388	0.390	0.388	0.389
Lives in the South	0.131	0.131	0.132	0.130	0.131
Lives in the Midwest	0.086	0.086	0.086	0.086	0.086
F-statistic	–	0.875	1.051	0.857	0.887
p-value F-statistic	–	0.599	0.452	0.614	0.588
Number of test-takers	14,941,156	3,655,807	3,903,653	3,590,977	3,790,719

Notes: This table shows summary statistics on all test-takers in the high-school-students sample (column 1) and based on the booklet color they received on the first day of testing (columns 2–5). The last panel reports the F-statistics and p-values from F-tests that the coefficients on all pre-determined covariates (Panels A, B, C, and E) are jointly equal across booklet colors.

Table A2: Examples of reliability estimates in economics and psychology

Construct (1)	Reliability estimate (2)	Reference (3)
IQ	0.80	Schuerger and Witt (1989)
Risk aversion	0.20–0.40	Mata et al. (2018)
Big 5 personality traits	0.60–0.73	Wooden (2012)
Present bias	0.36	Meier and Sprenger (2015)
Loss aversion	0.88	Stango and Zinman (2020)
Teacher value added	0.23–0.47	Chetty et al. (2014a)
Life satisfaction	0.67	Anusic and Schimmack (2016)
Self-esteem	0.71	Anusic and Schimmack (2016)
Academic ability	0.61–0.77	This paper
Cognitive endurance	0.14–0.30	This paper

Notes: This table displays examples of reliability estimates from the economics and psychology literature. The last two rows show the test-retest reliability of the measures of academic ability and cognitive endurance estimated in Section 5.

Table A3: IV estimates of the relationship between ability/endurance and college outcomes

	Dependent variable					
	Enrolled college (1)	College quality (2)	Degree quality (3)	1st-year credits (4)	Grad. on time (5)	Time to grad. (6)
Panel A. OLS estimates on retakers sample						
Endurance	0.048***	0.057***	0.110***	0.010***	0.026***	−0.082***
	(0.001)	(0.001)	(0.002)	(0.000)	(0.002)	(0.006)
Ability	0.110***	0.129***	0.217***	0.018***	0.049***	−0.154***
	(0.002)	(0.001)	(0.002)	(0.000)	(0.003)	(0.009)
Ratio coef.	0.441***	0.443***	0.509***	0.571***	0.543***	0.533***
	(0.011)	(0.006)	(0.007)	(0.008)	(0.037)	(0.032)
Mean DV	0.367	3.420	3.390	0.146	0.808	4.191
N	132,634	111,409	109,390	339,727	51,066	51,066
Panel B. IV estimates on retakers sample						
Endurance	0.046***	0.067***	0.141***	0.016***	0.041***	−0.120***
	(0.006)	(0.004)	(0.007)	(0.001)	(0.010)	(0.028)
Ability	0.107***	0.140***	0.239***	0.022***	0.057***	−0.169***
	(0.002)	(0.001)	(0.003)	(0.000)	(0.005)	(0.013)
Ratio coef.	0.430***	0.479***	0.590***	0.738***	0.722***	0.711***
	(0.053)	(0.025)	(0.028)	(0.026)	(0.156)	(0.145)
Mean DV	0.367	3.420	3.390	0.146	0.808	4.191
N	132,614	111,394	109,375	339,725	51,056	51,056

Notes: This table displays OLS and IV estimates of the relationship between ability/endurance and college outcomes.

The OLS estimates are analogous to Table 3 but estimated on the sample of retakers. See notes to Table 3 for details. The IV estimates instrument the year t measure of ability and cognitive endurance with the $t-1$ measures of these skills.

Heteroskedasticity-robust standard errors clustered at the individual level in parentheses. ***, ** and * denote significance at 10%, 5% and 1% levels, respectively.

Table A4: IV estimates of the relationship between ability/endurance and labor-market outcomes

	Dependent variable					
	Formal sector (1)	Hourly wage (2)	Monthly earnings (3)	Firm wage (4)	Occup. wage (5)	Industry wage (6)
Panel A. OLS estimates on retakers sample						
Endurance	0.001***	0.121***	0.124***	0.081***	0.030***	0.008***
	(0.000)	(0.005)	(0.005)	(0.005)	(0.003)	(0.001)
Ability	0.002***	0.231***	0.201***	0.163***	0.057***	0.018***
	(0.000)	(0.006)	(0.006)	(0.005)	(0.003)	(0.002)
Ratio coef.	0.518***	0.525***	0.615***	0.496***	0.518***	0.413***
	(0.072)	(0.018)	(0.020)	(0.024)	(0.040)	(0.057)
Mean DV	0.286	4.049	7.702	4.014	3.992	3.875
N	133,904	37,814	37,814	32,908	37,798	37,814
Panel B. IV estimates on retakers sample						
Endurance	0.003***	0.188***	0.232***	0.120***	0.052***	0.001
	(0.001)	(0.018)	(0.017)	(0.015)	(0.009)	(0.004)
Ability	0.002***	0.250***	0.215***	0.180***	0.061***	0.018***
	(0.000)	(0.008)	(0.007)	(0.007)	(0.004)	(0.002)
Ratio coef.	1.171***	0.753***	1.077***	0.670***	0.845***	0.050
	(0.323)	(0.069)	(0.079)	(0.081)	(0.150)	(0.241)
Mean DV	0.286	4.049	7.702	4.014	3.992	3.875
N	133,884	37,801	37,801	32,902	37,785	37,801

Notes: This table displays OLS and IV estimates of the relationship between ability/endurance and labor-market outcomes.

The OLS estimates are analogous to Table 3 but estimated on the sample of retakers. See notes to Table 3 for details. The IV estimates instrument the year t measure of ability and cognitive endurance with the $t-1$ measures of these skills.

Heteroskedasticity-robust standard errors clustered at the individual level in parentheses. ***, ** and * denote significance at 10%, 5% and 1% levels, respectively.

Table A5: Robustness of baseline test-score-gaps decomposition to measuring variables in percentiles

	Gap between				
	Male / Female (1)	White / Non-white (2)	Priv HS / Public HS (3)	Mom coll / No coll (4)	High-inc / Low-inc (5)
Panel A. Difference in average test-score percentile					
Score pctil gap	5.871***	14.127***	27.826***	21.609***	39.838***
	(0.015)	(0.015)	(0.019)	(0.018)	(0.023)
Panel B. Contribution of gaps in ability and endurance percentiles to test-score gaps					
Ability pctil gap	5.599***	10.618***	21.952***	16.728***	31.775***
	(0.012)	(0.011)	(0.017)	(0.015)	(0.024)
Endurance pctil gap	3.205***	3.097***	6.628***	4.596***	10.381***
	(0.006)	(0.006)	(0.010)	(0.008)	(0.015)
Panel C. Impact of a reform that halves the exam length on test-score percentile gaps					
Pctil. change gap	-1.603***	-1.548***	-3.314***	-2.298***	-5.190***
	(0.003)	(0.003)	(0.005)	(0.004)	(0.008)
Pct. change gap	-0.546***	-0.219***	-0.238***	-0.213***	-0.261***
	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
N (Students)	14,941,097	14,565,550	9,924,652	14,290,759	9,996,959

Notes: This table is analogous to Table 5, but the variables and effects are measured in percentiles. I construct the percentiles separately for each cohort. See notes to Table 5 for details.

Table A6: Robustness of baseline test-score-gaps decomposition to alternative ways of measuring ability and endurance

	Gap between				
	Male / Female (1)	White / Non-white (2)	Priv HS / Public HS (3)	Mom coll / No coll (4)	High-inc / Low-inc (5)
Panel A. Estimating ability/endurance separately by day and using the average					
Ability gap	0.030***	0.056***	0.127***	0.095***	0.188***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Endurance gap	0.034***	0.032***	0.075***	0.051***	0.126***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Panel B. Estimating ability/endurance separately by subject and using the average					
Ability gap	0.027***	0.061***	0.140***	0.104***	0.206***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Endurance gap	0.004***	0.042***	0.099***	0.069***	0.163***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Panel C. Including day fixed effects					
Ability gap	0.030***	0.056***	0.128***	0.096***	0.189***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Endurance gap	0.034***	0.031***	0.075***	0.051***	0.125***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Panel D. Including subject fixed effects					
Ability gap	0.026***	0.061***	0.141***	0.104***	0.207***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Endurance gap	0.002***	0.039***	0.095***	0.066***	0.158***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Panel E. Using linear correlation as an alternative measure of endurance					
Ability gap	0.030***	0.057***	0.129***	0.097***	0.191***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Endurance gap	0.021***	0.020***	0.047***	0.032***	0.079***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
N (Students)	14,941,097	14,565,550	9,924,652	14,290,759	9,996,959

Notes: This table shows estimates of the contribution of gaps in ability and endurance to test-score gaps using alternative specifications to estimate ability and endurance.

Each column shows the result for a different test-score gap. Each panel shows the result from estimating ability and endurance with a different specification. In Panels A–B, I estimate a student's ability/endurance separately for each testing day (Panel A) and academic subject (Panel B) and then average the estimates across days or subjects. In Panels C–D, I estimate endurance in a regression that controls for day fixed effects (Panel C) or subject fixed effects (Panel D). Finally, in Panel E, I use the correlation between question position and a dummy for correctly answering a question as an alternative measure of endurance.

Heteroskedasticity-robust standard errors clustered at the question level in parentheses. ***, ** and * denote significance at 10%, 5% and 1% levels, respectively.

Table A7: Robustness of baseline test-score-gaps decomposition to alternative ways of controlling for question difficulty when estimating ability/endurance

	Gap between				
	Male / Female (1)	White / Non-white (2)	Priv HS / Public HS (3)	Mom coll / No coll (4)	High-inc / Low-inc (5)
Panel A. Not controlling for question difficulty					
Ability gap	0.032***	0.049***	0.118***	0.087***	0.175***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Endurance gap	0.033***	0.026***	0.074***	0.050***	0.128***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Panel B. Estimating difficulty without adjusting for average position					
Ability gap	0.028***	0.057***	0.130***	0.097***	0.191***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Endurance gap	0.034***	0.036***	0.080***	0.055***	0.131***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Panel C. Estimating difficulty using question-specific position effects					
Ability gap	0.032***	0.055***	0.126***	0.095***	0.186***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Endurance gap	0.036***	0.031***	0.082***	0.058***	0.135***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Panel D. Estimating difficulty using shrunk question-specific position effects					
Ability gap	0.031***	0.056***	0.127***	0.095***	0.187***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Endurance gap	0.036***	0.032***	0.080***	0.057***	0.131***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Panel E. Estimating position effects separately by fraction of correct responses					
Ability gap	0.030***	0.056***	0.128***	0.096***	0.189***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Endurance gap	0.035***	0.032***	0.078***	0.053***	0.130***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Panel F. Estimating position effects separately by subject					
Ability gap	0.029***	0.057***	0.128***	0.096***	0.190***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Endurance gap	0.034***	0.034***	0.077***	0.053***	0.128***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
N (Students)	14,941,097	14,565,550	9,924,652	14,290,759	9,996,959

Notes: This table shows estimates of the contribution of gaps in ability and endurance to test-score gaps using alternative measures of difficulty in the specification used to estimate ability and endurance.

Each column shows the result for a different test-score gap. Each panel shows the result from a different way of controlling for question difficulty in equation (6). In Panel A, I compute the estimate equation (6) without controlling for question difficulty. In Panel B, I measure question difficulty as the fraction of students who incorrectly answer to the question across all booklets. In Panels C–F, I adjust for average question position by estimating the positon effects with alternative specifications. In column C, I compute question-specific position effects. In Panel D, I compute a shrinkage estimator of the position effects. In Panel E, I compute the position effects separately for questions with a below/above fraction of correct responses. In Panel F, I compute the position effects separately by subject. See Appendix D for details on each measure of question difficulty.

Heteroskedasticity-robust standard errors clustered at the question level in parentheses. ***, ** and * denote significance at 10%, 5% and 1% levels, respectively.

Table A8: Robustness of baseline test-score-gaps decomposition to alternative sample restrictions

	Gap between				
	Male / Female (1)	White / Non-white (2)	Priv HS / Public HS (3)	Mom coll / No coll (4)	High-inc / Low-inc (5)
Panel A. Excluding students in the bottom or top 10% of the ability distribution					
Ability gap	0.018***	0.036***	0.076***	0.055***	0.116***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Endurance gap	0.032***	0.031***	0.067***	0.045***	0.107***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Panel B. Excluding students in the bottom or top 10% of the endurance distribution					
Ability gap	0.027***	0.054***	0.125***	0.094***	0.190***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Endurance gap	0.017***	0.017***	0.041***	0.028***	0.078***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Panel C. Excluding students in the bottom or top 10% of either distribution					
Ability gap	0.016***	0.036***	0.073***	0.053***	0.113***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Endurance gap	0.017***	0.018***	0.037***	0.025***	0.062***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Panel D. Excluding students in the bottom or top 20% of either distribution					
Ability gap	0.009***	0.023***	0.042***	0.030***	0.066***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Endurance gap	0.009***	0.011***	0.019***	0.013***	0.032***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Panel E. Excluding individuals with positive estimated endurance					
Ability gap	0.021***	0.050***	0.112***	0.084***	0.165***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Endurance gap	0.013***	0.017***	0.037***	0.025***	0.063***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)

Notes: This table shows estimates of the contribution of gaps in ability and endurance to test-score gaps using alternative sample restrictions.

Each column shows the result for a different test-score gap. Each panel shows the result for a different sample of students. In Panel A, I exclude students in the bottom and top deciles of the ability distribution. In Panel B, I exclude students in the bottom and top deciles of the endurance distribution. In Panel C, I exclude students in the bottom and top deciles of the distribution of either skill. In Panel D, I exclude students in the bottom and top quintiles of the distribution of either skill. In Panel E, I exclude students with positive estimated endurance. I construct the deciles and quintiles using all the students in the high-school-students sample.

Heteroskedasticity-robust standard errors clustered at the question level in parentheses. ***, ** and * denote significance at 10%, 5% and 1% levels, respectively.

Table A9: Robustness of baseline test-score-gaps decomposition to accounting for measurement error

	Gap between				
	Male / Female (1)	White / Non-white (2)	Priv HS / Public HS (3)	Mom coll / No coll (4)	High-inc / Low-inc (5)
Panel A. Weighting each observation by its precision					
Ability gap	0.030***	0.056***	0.127***	0.095***	0.188***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Endurance gap	0.034***	0.031***	0.075***	0.051***	0.126***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Panel B. Shrunk estimator of ability and endurance					
Ability gap	0.021***	0.040***	0.091***	0.067***	0.136***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Endurance gap	0.009***	0.010***	0.023***	0.016***	0.040***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
N (Students)	14,941,097	14,565,550	9,924,652	14,290,759	9,996,959

Notes: This table shows estimates of the contribution of gaps in ability and endurance to test-score gaps accounting for measurement error in the estimates of ability and endurance.

Each column shows the result for a different test-score gap. In Panel A, I weight each observation by the inverse of the standard error of the ability and endurance estimates. Specifically, the weight of each observation is $w = 1/(\text{SE}^2_{\hat{\alpha}_i} + \text{SE}^2_{\hat{\beta}_i})$, where $\text{SE}_{\hat{\alpha}_i}$ and $\text{SE}^2_{\hat{\beta}_i}$ are the standard errors of $\hat{\alpha}_i$ and $\hat{\beta}_i$. In Panel B, I estimate the baseline regression using a shrunk estimator of ability and endurance. I compute the shrunk estimator of endurance as $\beta^s_i = \omega_i \hat{\beta}_i + (1 - \omega_i)\bar{\beta}$, where $\bar{\beta}$ is the average cognitive endurance in my sample. The individual-specific weight is $\omega_i = \frac{\mathbb{V}\text{ar}[\beta_i] - \mathbb{E}[\text{SE}^2_{\hat{\beta}_i}]}{\mathbb{V}\text{ar}[\beta_i] - \mathbb{E}[\text{SE}^2_{\hat{\beta}_i}] + \text{SE}^2_{\hat{\beta}_i}}$. The shrunk estimator, β^s_i, puts more weight on estimates of β_i that are more precisely estimated, as measured by a low standard error. I compute the shrunk estimator of ability analgously.

Heteroskedasticity-robust standard errors clustered at the question level in parentheses. ***, ** and * denote significance at 10%, 5% and 1% levels, respectively.

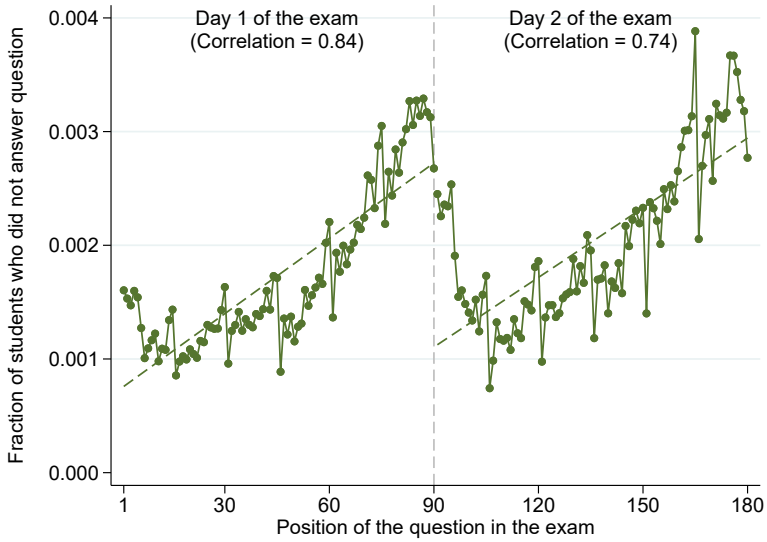
B Empirical Appendix

B.1 Limited Cognitive Endurance and Time Pressure

Is the causal effect of an increase in the order of a given question on student performance a manifestation of limited cognitive endurance or is it driven by students running out of time? Two pieces of evidence suggest that time pressure does not explain the estimated $\beta < 0$.

First, very few students leave responses unanswered. Appendix Figure B1 plots the fraction of students who left a question unanswered (possibly, because they ran out of time) against the question position. Questions that appear later in the test are more likely to be left unanswered. However, only a small fraction of students leave *any* questions unanswered. Thus, missing responses cannot account for the large change in performance observed throughout the exam.[25]

Figure B1: Fraction of question left unanswered throughout the ENEM



Notes: This figure shows the fraction of questions left unanswered over the course of each testing day. The y-axis displays the fraction of students who did not select any of the multiple-choice answers to a given question. The x-axis displays the position of each question in the exam. The dashed lines are predicted values from a linear regression estimated separately for each testing day.

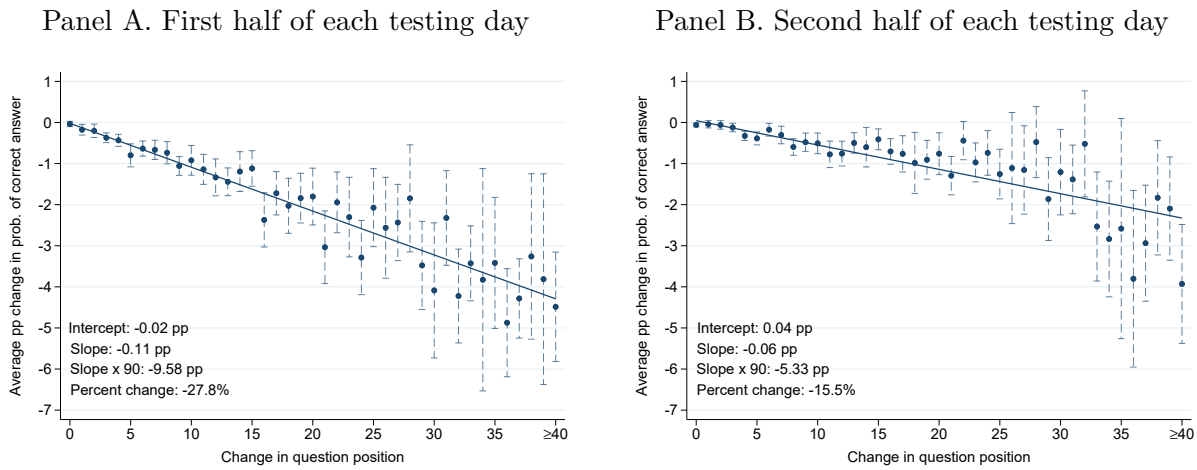
Second, student performance declines even in questions that students answer when they

[25]There is no penalty for incorrectly answering a question. Therefore, this evidence is only suggestive since leaving a question unanswered is a weakly dominated strategy.

are likely not time-pressured. Appendix Figure B2 shows the fatigue effect separately for questions that appear in the first half (Panel A) and the second half of each testing day (Panel B). Presumably, students should have plenty of time to answer the first half of the exam. Yet, I still find fatigue effects that are quantitatively similar—or even larger—to those estimated on the second half of each day or with all questions. This result is consistent with visual evidence in Figure 2, which shows that student performance tends to decline shortly after the exam starts and with the declines in performance exhibited by the example questions that appear at the beginning of the exam in Appendix Figure A3.

In summary, the evidence indicates that the effect of a question position on student performance is not driven by students running out of time.

Figure B2: The heterogeneous effect of fatigue on performance by question position



Notes: This figure shows heterogeneity in the effect of limited endurance on performance by question position. Panels A and B are analogous to Figure 3, but the effect is estimated separately for questions that appear on the first half of each testing day (Panel A) or the second half of each testing day (Panel B). The y-axis shows the average change (in percentage points) in the fraction of students correctly responding to a question. The x-axis plots the difference in the question position between each possible booklet pair. The dashed line denotes predicted values from a linear regression estimated on the plotted points, using the number of questions used to estimate each point as weights.

B.2 OLS Formulas of Academic Ability and Cognitive Endurance

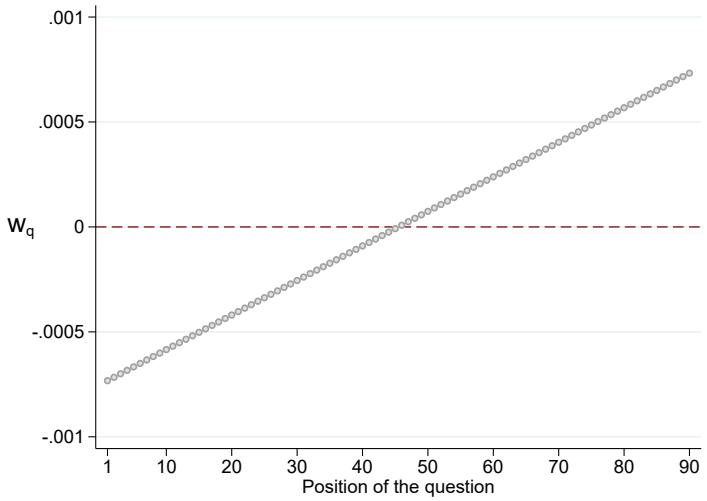
My measure of cognitive endurance is β_i in equation (6). Ignoring controls for question difficulty, the OLS estimator of β_i is

$$\begin{aligned}\hat{\beta}_i &= \frac{\sum_j (\text{Pos}_{ij} - \overline{\text{Pos}})(C_{ij} - \bar{C}_i)}{\sum_j (\text{Pos}_{ij} - \overline{\text{Pos}})^2} \\ &= \sum_j \underbrace{w_j}_{\substack{\text{Weight of} \\ \text{question } j}} \times \underbrace{(C_{ij} - \bar{C}_i)}_{\substack{\text{Performance on} \\ \text{question } j \text{ relative to} \\ i\text{'s average performance}}}, \end{aligned} \tag{B1}$$

where $\bar{C}_i$ is the fraction of questions correctly answered by student i, $\overline{\text{Pos}}$ is the average question position (which is constant across test-takers), and $w_j \equiv \frac{\text{Pos}_j - \overline{\text{Pos}}}{\sum_j (\text{Pos}_j - \overline{\text{Pos}})^2}$ is the weight of question j.

Equation (B1) shows that $\hat{\beta}_i$ is a weighted average of deviations from i's average score. The weight of each question depends on the location of the question on the test. Appendix Figure B3 plots the weight OLS places on each question. The questions with the largest weights (in absolute value) are the ones at the beginning and the end of the test.

Figure B3: Weight of each question in a test with 90 questions



Notes: This figure displays the weight put by the ordinary least squares (OLS) estimator of β_i (equation (6)) on each question of the test.

My measure of academic ability is α_i in equation (6). The OLS estimator of α_i is

$$\hat{\alpha}_i = \bar{C}_i - \hat{\beta}_i \overline{\text{Pos}} \tag{B2}$$

$$= \bar{C}_i - \sum_j w_j C_{ij} \overline{\text{Pos}}. \tag{B3}$$

Equation (B2) shows that α_i can be estimated by the difference between i's test score ($\bar{C}_i$) and the part of her test score that is explained by limited endurance, $\hat{\beta}_i \overline{\text{Pos}}$.

B.3 Estimating the Standard Deviation of Ability and Endurance

The estimate of cognitive endurance, $\hat{\beta}_i$, can be decomposed into latent cognitive endurance, β_i, and a sampling error e_i independent of β_i and with variance σ_e^2:

$$\hat{\beta}_i = \beta_i + e_i \tag{B4}$$

Calculating the variance on each side of equation (B4) yields:

$$\sigma_{\hat{\beta}}^2 = \sigma_\beta^2 + \sigma_e^2, \tag{B5}$$

where $\sigma_{\hat{\beta}}^2$ and σ_β^2 are the variances of $\hat{\beta}$ and β, respectively. Equation (B5) shows that the raw standard deviation of $\hat{\beta}$ overstates the variability of β since it includes variability in the sampling error. Let $\text{SE}_{\hat{\beta}}$ be the standard error of $\hat{\beta}$. The variance of the sampling error can be estimated as $\sigma_e^2 = \mathbb{E}[\text{SE}_{\hat{\beta}}^2]$. Thus, an estimate of the variance of β is given by

$$\hat{\sigma}_\beta^2 = \sigma_{\hat{\beta}}^2 - \mathbb{E}[\text{SE}_{\hat{\beta}}^2].$$

I use an analogous derivation to estimate the variance of latent ability, σ_α^2.

B.4 Robustness of the Relationship between Endurance and Long-Run Outcomes

Appendix Table B1 shows non-parametric estimates of the effect of ability and endurance on each outcome based on the slope of percentile changes on outcomes (Heckman et al., 2006). I estimate how a movement from the bottom decile to the top decile in the endurance

distribution affects a given outcome:

$$\mathbb{E}[Y_i|i \in \text{Top decile Endurance}] - \mathbb{E}[Y_i|i \in \text{Bottom decile Endurance}].$$

As a benchmark, I compare the size of a decile movement in the endurance distribution to an equivalent decile movement in the ability distribution. I compute these effects in a regression framework by estimating equations of the form:

$$Y_i = \phi + \lambda X_i + \sum_{d=2}^{10} \mathbb{1}\{i \in \text{TestScore decile } d\} + \zeta_i$$
$$Y_i = \tilde{\phi}_1 + \tilde{\lambda}_1 X_i + \sum_{d=2}^{10} \mathbb{1}\{i \in \text{Ability decile } d\} + \sum_{d=2}^{10} \mathbb{1}\{i \in \text{Endurance decile } d\} + \tilde{\zeta}_i,$$

where the omitted category is the bottom decile. The first row of each panel shows that moving higher in the distribution of test scores tends to improve college and labor-market outcomes. Subsequent rows show that both cognitive endurance and ability contribute to this effect. Depending on the outcome, the predicted effect of a movement from decile 1 to decile 10 in the endurance distribution represents 32.6%–53.0% of the corresponding effect of a movement in the ability distribution.

Appendix Tables B2–B3 show that the results are robust to estimating ability and endurance with alternative specifications. First, I compute the estimates of ability/endurance separately for each testing day and for each academic subject, and use the average estimate across days/subjects as regressors in equation (8). Second, I compute the estimates of endurance controlling for day fixed effects and subject fixed effects; thus accounting for possible differences in preparation across subjects. Finally, I use the correlation between question position and a dummy for correctly answering a question as an alternative measure of endurance. Across specifications, I find effects that are quantitatively similar and qualitatively identical to those of the baseline specification.

Appendix Tables B4–B5 show that the results are robust to controlling for question difficulty in alternative ways when estimating ability and endurance. First, I compute the estimates of ability and endurance in equation (6) without controlling for question difficulty. Second, I calculate question difficulty without adjusting for the average position of the question across booklets. Finally, I compute question difficulty adjusting for average question position in several alternative ways (see Appendix D). Consistent with the baseline results, I find that the estimates are remarkably robust across specifications.

Appendix Tables B6–B7 shows that the results are robust to different sample restrictions. Specifically, I estimate the baseline specification excluding students in the tails of the ability and the endurance distributions. These are students for whom floor and ceiling effects may be binding and, thus, for whom estimates may be biased. I also exclude students with a positive estimate of endurance. These are students who, for example, may answer the exam in reverse order. I find little impact of these sample restrictions on the estimates.

Appendix Tables B8–B9 show robustness of the baseline regressions to accounting for measurement error. First, I weight each observation by the inverse of the standard error of the ability and endurance estimates, thus giving more weight to students for which I estimate more precise measures. Second, I estimate the baseline regressions using shrunk estimates of ability and endurance. The shrunk estimators of ability and endurance put more weight on measures estimated with more precision, as measured by a low standard error. The results are very similar to the baseline results.

Table B1: The effect of a movement from decile 1 to decile 10 in the ability/endurance distribution on long-run outcomes

Panel A. College outcomes

	Dependent variable					
	Enrolled college (1)	College quality (2)	Degree quality (3)	1st-year credits (4)	Grad. on time (5)	Time to grad. (6)
Test score	0.300***	0.275***	0.389***	0.046***	0.215***	−0.397***
	(0.001)	(0.001)	(0.002)	(0.001)	(0.002)	(0.008)
Endurance	0.136***	0.139***	0.232***	0.030***	0.132***	−0.224***
	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)	(0.007)
Ability	0.319***	0.338***	0.488***	0.057***	0.272***	−0.486***
	(0.002)	(0.001)	(0.002)	(0.001)	(0.003)	(0.009)
Ratio coef.	0.426***	0.412***	0.474***	0.530***	0.485***	0.462***
	(0.005)	(0.003)	(0.003)	(0.009)	(0.007)	(0.012)
Mean DV	0.329	3.327	3.244	0.158	0.418	3.842
N	1,850,938	1,711,475	1,681,214	1,124,972	1,471,569	786,391

Panel B. Labor-market outcomes

	Dependent variable					
	Formal sector (1)	Hourly wage (2)	Monthly earnings (3)	Firm wage (4)	Occup. wage (5)	Industry wage (6)
Test score	0.005***	0.457***	0.395***	0.320***	0.141***	0.046***
	(0.000)	(0.004)	(0.004)	(0.004)	(0.002)	(0.001)
Endurance	0.002***	0.241***	0.240***	0.159***	0.084***	0.018***
	(0.000)	(0.005)	(0.004)	(0.004)	(0.003)	(0.001)
Ability	0.006***	0.543***	0.475***	0.387***	0.177***	0.055***
	(0.000)	(0.005)	(0.005)	(0.005)	(0.003)	(0.001)
Ratio coef.	0.337***	0.443***	0.506***	0.410***	0.472***	0.326***
	(0.034)	(0.007)	(0.007)	(0.009)	(0.012)	(0.020)
Mean DV	0.326	3.865	7.551	3.885	3.886	3.858
N	2,523,032	818,590	818,590	692,880	818,374	818,590

Notes: This table displays estimates of the relationship between ability/endurance and college outcomes (Panel A) and labor-market outcomes (Panel B).

The first row of each panel shows estimates of the mean outcome difference between individuals in the tenth and first decile of the test score distribution (the coefficient on the decile ten dummy in equation (**??**)). The following rows show estimates of the mean outcome difference between individuals in the tenth and first decile of the ability/endurance distribution (the coefficients on the decile ten dummies in equation (**??**)). See Section 5 for a description of the measures of ability and endurance. See Section 2.4 for outcome definitions. Heteroskedasticity-robust standard errors clustered at the individual level in parentheses.

The third-to-last row in each panel shows the ratio between the effect of ability and endurance on a given outcome. Standard errors estimated through the delta method in parentheses. ***, ** and * denote significance at 10%, 5% and 1% levels, respectively.

Table B2: Robustness of baseline regressions to alternative ways of measuring ability and endurance: College outcomes

	Dependent variable:					
	Enrolled college (1)	College quality (2)	Degree quality (3)	1st-year credits (4)	Grad. on time (5)	Time to grad. (6)
Panel A. Estimating ability/endurance separately by day and using the average						
Endurance	0.053***	0.050***	0.084***	0.010***	0.043***	−0.079***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.002)
Ability	0.114***	0.107***	0.157***	0.018***	0.081***	−0.158***
	(0.000)	(0.000)	(0.001)	(0.000)	(0.001)	(0.002)
Panel B. Estimating ability/endurance separately by subject and using the average						
Endurance	0.060***	0.055***	0.080***	0.009***	0.043***	−0.079***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.002)
Ability	0.103***	0.097***	0.140***	0.016***	0.067***	−0.130***
	(0.000)	(0.000)	(0.001)	(0.000)	(0.001)	(0.002)
Panel C. Including day fixed effects						
Endurance	0.031***	0.030***	0.050***	0.006***	0.025***	−0.047***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Ability	0.110***	0.104***	0.152***	0.018***	0.077***	−0.151***
	(0.000)	(0.000)	(0.001)	(0.000)	(0.001)	(0.002)
Panel D. Including subject fixed effects						
Endurance	0.019***	0.018***	0.026***	0.003***	0.014***	−0.025***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Ability	0.097***	0.092***	0.133***	0.015***	0.062***	−0.119***
	(0.000)	(0.000)	(0.001)	(0.000)	(0.001)	(0.002)
Panel E. Using linear correlation as an alternative measure of endurance						
Endurance	0.030***	0.030***	0.049***	0.006***	0.025***	−0.047***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Ability	0.101***	0.095***	0.138***	0.016***	0.072***	−0.139***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.002)
Mean DV	0.244	3.326	3.244	0.158	0.418	3.817
N	2,501,519	1,800,546	1,768,707	1,124,972	1,472,916	793,822

Notes: This table shows estimates of the relationship between ability/endurance and college outcomes using alternative specifications to estimate ability and endurance.

Each column shows the result for a different dependent variable. Each panel shows the result from estimating ability and endurance with a different specification. In Panels A–B, I estimate a student's ability/endurance separately for each testing day (Panel A) and academic subject (Panel B) and then average the estimates across days or subjects. In Panels C–D, I estimate endurance in a regression that controls for day fixed effects (Panel C) or subject fixed effects (Panel D). Finally, in Panel E, I use the correlation between question position and a dummy for correctly answering a question as an alternative measure of endurance.

Heteroskedasticity-robust standard errors clustered at the individual level in parentheses.***, ** and * denote significance at 10%, 5% and 1% levels, respectively.

Table B3: Robustness of baseline regressions to alternative ways of measuring ability and endurance: Labor-market outcomes

	Dependent variable:					
	Formal sector (1)	Hourly wage (2)	Monthly earnings (3)	Firm wage (4)	Occup. wage (5)	Industry wage (6)
Panel A. Estimating ability/endurance separately by day and using the average						
Endurance	0.001***	0.088***	0.085***	0.058***	0.028***	0.006***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
Ability	0.002***	0.172***	0.152***	0.121***	0.055***	0.016***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
Panel B. Estimating ability/endurance separately by subject and using the average						
Endurance	0.001***	0.088***	0.076***	0.061***	0.026***	0.008***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
Ability	0.002***	0.156***	0.135***	0.110***	0.048***	0.014***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
Panel C. Including day fixed effects						
Endurance	0.000***	0.053***	0.051***	0.035***	0.017***	0.004***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)
Ability	0.002***	0.167***	0.147***	0.117***	0.053***	0.016***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
Panel D. Including subject fixed effects						
Endurance	0.000***	0.029***	0.025***	0.020***	0.008***	0.003***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Ability	0.002***	0.147***	0.127***	0.104***	0.045***	0.013***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
Panel E. Using linear correlation as an alternative measure of endurance						
Endurance	0.000***	0.052***	0.050***	0.035***	0.016***	0.004***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)
Ability	0.002***	0.151***	0.132***	0.107***	0.048***	0.014***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
Mean DV	0.326	3.865	7.551	3.885	3.886	3.858
N	2,523,029	818,590	818,590	692,880	818,374	818,590

Notes: This table shows estimates of the relationship between ability/endurance and labor-market outcomes using alternative specifications to estimate ability and endurance.

Each column shows the result for a different dependent variable. Each panel shows the result from estimating ability and endurance with a different specification. In Panels A–B, I estimate a student's ability/endurance separately for each testing day (Panel A) and academic subject (Panel B) and then average the estimates across days or subjects. In Panels C–D, I estimate endurance in a regression that controls for day fixed effects (Panel C) or subject fixed effects (Panel D). Finally, in Panel E, I use the correlation between question position and a dummy for correctly answering a question as an alternative measure of endurance.

Heteroskedasticity-robust standard errors clustered at the individual level in parentheses. ***, ** and * denote significance at 10%, 5% and 1% levels, respectively.

Table B4: Robustness of the baseline regressions to alternative ways of controlling for question difficulty when estimating ability/endurance: College outcomes

	Dependent variable:					
	Enrolled college (1)	College quality (2)	Degree quality (3)	1st-year credits (4)	Grad. on time (5)	Time to grad. (6)
Panel A. Not controlling for question difficulty						
Endurance	0.040***	0.045***	0.080***	0.007***	0.028***	−0.061***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.002)
Ability	0.114***	0.112***	0.169***	0.018***	0.079***	−0.159***
	(0.000)	(0.000)	(0.001)	(0.000)	(0.001)	(0.002)
Panel B. Estimating difficulty without adjusting for average position						
Endurance	0.030***	0.028***	0.045***	0.006***	0.026***	−0.046***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Ability	0.097***	0.090***	0.130***	0.016***	0.068***	−0.133***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.002)
Panel C. Estimating difficulty using question-specific position effects						
Endurance	0.035***	0.038***	0.065***	0.007***	0.027***	−0.054***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Ability	0.106***	0.102***	0.152***	0.017***	0.074***	−0.147***
	(0.000)	(0.000)	(0.001)	(0.000)	(0.001)	(0.002)
Panel D. Estimating difficulty using shrunk question-specific position effects						
Endurance	0.034***	0.035***	0.059***	0.006***	0.026***	−0.052***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Ability	0.104***	0.098***	0.146***	0.017***	0.073***	−0.144***
	(0.000)	(0.000)	(0.001)	(0.000)	(0.001)	(0.002)
Panel E. Estimating position effects separately by fraction of correct responses						
Endurance	0.031***	0.030***	0.049***	0.006***	0.026***	−0.047***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Ability	0.101***	0.094***	0.138***	0.016***	0.071***	−0.139***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.002)
Panel F. Estimating position effects separately by subject						
Endurance	0.031***	0.029***	0.048***	0.006***	0.026***	−0.047***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Ability	0.099***	0.093***	0.135***	0.016***	0.071***	−0.137***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.002)
Mean DV	0.244	3.326	3.244	0.158	0.418	3.817
N	2,501,519	1,800,546	1,768,707	1,124,972	1,472,916	793,822

Notes: This table shows estimates of the relationship between ability/endurance and college outcomes using alternative measures of difficulty in the specification used to estimate ability and endurance.

Each panel shows the result using a different measure of question difficulty in equation (6). In Panel A, I estimate equation (6) without controlling for question difficulty. In Panel B, I measure question difficulty as the fraction of students who incorrectly answer the question across all booklets. In Panels C–F, I adjust for differences in average position across questions by estimating the position effects with alternative specifications. In column C, I compute question-specific position effects. In Panel D, I compute a shrinkage estimator of the position effects. In Panel E, I compute the position effects separately for questions with a below/above fraction of correct responses. In Panel F, I compute the position effects separately by subject. See Appendix D for details on each measure of question difficulty. Heteroskedasticity-robust standard errors clustered at the individual level in parentheses.***, ** and * denote significance at 10%, 5% and 1% levels, respectively.

Table B5: Robustness of the baseline regressions to alternative ways of controlling for question difficulty when estimating ability/endurance: Labor-market outcomes

	Dependent variable:					
	Formal sector (1)	Hourly wage (2)	Monthly earnings (3)	Firm wage (4)	Occup. wage (5)	Industry wage (6)
Panel A. Not controlling for question difficulty						
Endurance	0.001***	0.080***	0.077***	0.053***	0.022***	0.004***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
Ability	0.002***	0.183***	0.163***	0.127***	0.056***	0.015***
	(0.000)	(0.002)	(0.001)	(0.001)	(0.001)	(0.000)
Panel B. Estimating difficulty without adjusting for average position						
Endurance	0.000***	0.048***	0.047***	0.032***	0.016***	0.004***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)
Ability	0.001***	0.143***	0.125***	0.101***	0.046***	0.014***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
Panel C. Estimating difficulty using question-specific position effects						
Endurance	0.001***	0.066***	0.064***	0.044***	0.019***	0.004***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)
Ability	0.002***	0.165***	0.146***	0.115***	0.051***	0.015***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
Panel D. Estimating difficulty using shrunk question-specific position effects						
Endurance	0.000***	0.061***	0.059***	0.040***	0.018***	0.004***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)
Ability	0.002***	0.159***	0.140***	0.111***	0.050***	0.014***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
Panel E. Estimating position effects separately by fraction of correct responses						
Endurance	0.000***	0.053***	0.051***	0.035***	0.017***	0.004***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)
Ability	0.002***	0.152***	0.133***	0.107***	0.048***	0.014***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
Panel F. Estimating position effects separately by subject						
Endurance	0.000***	0.051***	0.049***	0.034***	0.016***	0.004***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)
Ability	0.002***	0.149***	0.130***	0.105***	0.048***	0.014***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
Mean DV	0.326	3.865	7.551	3.885	3.886	3.858
N	2,523,029	818,590	818,590	692,880	818,374	818,590

Notes: This table shows estimates of the relationship between ability/endurance and labor-market outcomes using alternative measures of difficulty in the specification used to estimate ability and endurance.

Each panel shows the result using a different measure of question difficulty in equation (6). In Panel A, I estimate equation (6) without controlling for question difficulty. In Panel B, I measure question difficulty as the fraction of students who incorrectly answer the question across all booklets. In Panels C–F, I adjust for differences in average position across questions by estimating the position effects with alternative specifications. In column C, I compute question-specific position effects. In Panel D, I compute a shrinkage estimator of the position effects. In Panel E, I compute the position effects separately for questions with a below/above fraction of correct responses. In Panel F, I compute the position effects separately by subject. See Appendix D for details on each measure of question difficulty. Heteroskedasticity-robust standard errors clustered at the individual level in parentheses.***, ** and * denote significance at 10%, 5% and 1% levels, respectively.

Table B6: Robustness of the baseline regressions to sample selection: College outcomes

	Dependent variable:					
	Enrolled college (1)	College quality (2)	Degree quality (3)	1st-year credits (4)	Grad. on time (5)	Time to grad. (6)
Panel A. Excluding students in the bottom or top 10% of the ability distribution						
Endurance	0.032***	0.027***	0.039***	0.007***	0.030***	−0.047***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Ability	0.102***	0.085***	0.107***	0.019***	0.087***	−0.146***
	(0.000)	(0.000)	(0.001)	(0.000)	(0.001)	(0.003)
Panel B. Excluding students in the bottom or top 10% of the endurance distribution						
Endurance	0.030***	0.030***	0.050***	0.006***	0.025***	−0.044***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.002)
Ability	0.100***	0.093***	0.135***	0.016***	0.075***	−0.138***
	(0.000)	(0.000)	(0.001)	(0.000)	(0.001)	(0.002)
Panel C. Excluding students in the bottom or top 10% of either distribution						
Endurance	0.031***	0.026***	0.037***	0.007***	0.030***	−0.046***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.002)
Ability	0.102***	0.083***	0.102***	0.020***	0.089***	−0.146***
	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)	(0.003)
Panel D. Excluding students in the bottom or top 20% of either distribution						
Endurance	0.030***	0.023***	0.030***	0.008***	0.034***	−0.044***
	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)	(0.003)
Ability	0.099***	0.074***	0.086***	0.022***	0.100***	−0.152***
	(0.001)	(0.001)	(0.001)	(0.000)	(0.001)	(0.004)
Panel E. Excluding individuals with positive estimated endurance						
Endurance	0.033***	0.028***	0.048***	0.005***	0.026***	−0.047***
	(0.000)	(0.000)	(0.001)	(0.000)	(0.001)	(0.002)
Ability	0.107***	0.097***	0.143***	0.016***	0.065***	−0.133***
	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)	(0.003)

Notes: This table shows estimates of the relationship between ability/endurance and college outcomes using alternative sample restrictions.

Each column shows the result for a different dependent variable. Each panel shows the result for a different sample of students. In Panel A, I exclude students in the bottom and top deciles of the ability distribution. In Panel B, I exclude students in the bottom and top deciles of the endurance distribution. In Panel C, I exclude students in the bottom and top deciles of the distribution of either skill. In Panel D, I exclude students in the bottom and top quintiles of the distribution of either skill. In Panel E, I exclude students with a positive estimate of endurance ($\hat{\beta} > 0$). I construct the deciles and quintiles using all the students in the high-school-students sample.

Heteroskedasticity-robust standard errors clustered at the individual level in parentheses.***, ** and * denote significance at 10%, 5% and 1% levels, respectively.

Table B7: Robustness of the baseline regressions to sample selection: Labor-market outcomes

	Dependent variable:					
	Formal sector (1)	Hourly wage (2)	Monthly earnings (3)	Firm wage (4)	Occup. wage (5)	Industry wage (6)
Panel A. Excluding students in the bottom or top 10% of the ability distribution						
Endurance	0.000***	0.046***	0.044***	0.031***	0.017***	0.004***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)
Ability	0.001***	0.130***	0.113***	0.092***	0.051***	0.016***
	(0.000)	(0.002)	(0.001)	(0.001)	(0.001)	(0.000)
Panel B. Excluding students in the bottom or top 10% of the endurance distribution						
Endurance	0.000***	0.053***	0.050***	0.035***	0.016***	0.004***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
Ability	0.001***	0.149***	0.131***	0.103***	0.049***	0.014***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
Panel C. Excluding students in the bottom or top 10% of either distribution						
Endurance	0.000***	0.044***	0.042***	0.029***	0.017***	0.004***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
Ability	0.001***	0.127***	0.112***	0.089***	0.050***	0.015***
	(0.000)	(0.002)	(0.001)	(0.001)	(0.001)	(0.000)
Panel D. Excluding students in the bottom or top 20% of either distribution						
Endurance	0.000***	0.039***	0.037***	0.025***	0.016***	0.004***
	(0.000)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
Ability	0.001***	0.117***	0.105***	0.083***	0.052***	0.016***
	(0.000)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
Panel E. Excluding individuals with positive estimated endurance						
Endurance	0.001***	0.054***	0.055***	0.033***	0.017***	0.003***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
Ability	0.002***	0.158***	0.137***	0.112***	0.049***	0.014***
	(0.000)	(0.002)	(0.002)	(0.002)	(0.001)	(0.000)

Notes: This table shows estimates of the relationship between ability/endurance and labor-market outcomes using alternative sample restrictions.

Each column shows the result for a different dependent variable. Each panel shows the result for a different sample of students. In Panel A, I exclude students in the bottom and top deciles of the ability distribution. In Panel B, I exclude students in the bottom and top deciles of the endurance distribution. In Panel C, I exclude students in the bottom and top deciles of the distribution of either skill. In Panel D, I exclude students in the bottom and top quintiles of the distribution of either skill. In Panel E, I exclude students with a positive estimate of endurance ($\hat{\beta} > 0$). I construct the deciles and quintiles using all the students in the high-school-students sample.

Heteroskedasticity-robust standard errors clustered at the individual level in parentheses.***, ** and * denote significance at 10%, 5% and 1% levels, respectively.

Table B8: Robustness of the baseline regressions to accounting for measurement error: College outcomes

	Dependent variable					
	Enrolled college (1)	College quality (2)	Degree quality (3)	1st-year credits (4)	Grad. on time (5)	Time to grad. (6)
Panel A. Weighting each observation by its precision						
Endurance	0.031***	0.030***	0.051***	0.006***	0.026***	−0.048***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Ability	0.100***	0.094***	0.139***	0.016***	0.073***	−0.140***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.002)
Panel B. Shrunk estimator of ability and endurance						
Endurance	0.045***	0.043***	0.073***	0.012***	0.037***	−0.067***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.002)
Ability	0.105***	0.099***	0.145***	0.023***	0.074***	−0.143***
	(0.000)	(0.000)	(0.001)	(0.000)	(0.001)	(0.002)
Mean DV	0.244	3.326	3.244	0.158	0.418	3.817
N	2,501,519	1,800,546	1,768,707	1,124,972	1,472,916	793,822

Notes: This table displays estimates of the relationship between ability/endurance and college outcomes accounting for measurement error in the estimates of ability and endurance.

Each column shows the result for a different dependent variable. In Panel A, I weight each observation by the inverse of the standard error of the ability and endurance estimates. Specifically, the weight of each observation is $w = 1/(\text{SE}^2_{\hat{\alpha}_i} + \text{SE}^2_{\hat{\beta}_i})$, where $\text{SE}_{\hat{\alpha}_i}$ and $\text{SE}^2_{\hat{\beta}_i}$ are the standard errors of $\hat{\alpha}_i$ and $\hat{\beta}_i$. In Panel B, I estimate the baseline regression using a shrunk estimator of ability and endurance. I compute the shrunk estimator of endurance as $\beta^s_i = \omega_i \hat{\beta}_i + (1 - \omega_j)\bar{\beta}$, where $\bar{\beta}$ is the average cognitive endurance in my sample. The individual-specific weight is $\omega_i = \frac{\mathbb{V}\text{ar}[\beta_i] - \mathbb{E}[\text{SE}^2_{\hat{\beta}_i}]}{\mathbb{V}\text{ar}[\beta_i] - \mathbb{E}[\text{SE}^2_{\hat{\beta}_i}] + \text{SE}^2_{\hat{\beta}_i}}$. The shrunk estimator, β^s_i, puts more weight on estimates of β_i that are more precisely estimated, as measured by a low standard error. I compute the shrunk estimator of ability analogously.

Heteroskedasticity-robust standard errors clustered at the individual level in parentheses.***, ** and * denote significance at 10%, 5% and 1% levels, respectively.

Table B9: Robustness of the baseline regressions to accounting for measurement error: Labor-market outcomes

	Dependent variable					
	Formal sector (1)	Hourly wage (2)	Monthly earnings (3)	Firm wage (4)	Occup. wage (5)	Industry wage (6)
Panel A. Weighting each observation by its precision						
Endurance	0.000***	0.053***	0.051***	0.035***	0.017***	0.004***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)
Ability	0.001***	0.151***	0.133***	0.107***	0.048***	0.014***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
Panel B. Shrunk estimator of ability and endurance						
Endurance	0.001***	0.076***	0.074***	0.050***	0.024***	0.005***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
Ability	0.002***	0.157***	0.138***	0.111***	0.050***	0.015***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
Mean DV	0.326	3.865	7.551	3.885	3.886	3.858
N	2,523,029	818,590	818,590	692,880	818,374	818,590

Notes: This table displays estimates of the relationship between ability/endurance and labor-market outcomes accounting for measurement error in the estimates of ability and endurance.

Each column shows the result for a different dependent variable. In Panel A, I weight each observation by the inverse of the standard error of the ability and endurance estimates. Specifically, the weight of each observation is $w = 1/(\text{SE}^2_{\hat{\alpha}_i} + \text{SE}^2_{\hat{\beta}_i})$, where $\text{SE}_{\hat{\alpha}_i}$ and $\text{SE}^2_{\hat{\beta}_i}$ are the standard errors of $\hat{\alpha}_i$ and $\hat{\beta}_i$. In Panel B, I estimate the baseline regression using a shrunk estimator of ability and endurance. I compute the shrunk estimator of endurance as $\beta^s_i = \omega_i \hat{\beta}_i + (1 - \omega_j)\bar{\beta}$, where $\bar{\beta}$ is the average cognitive endurance in my sample. The individual-specific weight is $\omega_i = \frac{\mathbb{V}\text{ar}[\beta_i] - \mathbb{E}[\text{SE}^2_{\hat{\beta}_i}]}{\mathbb{V}\text{ar}[\beta_i] - \mathbb{E}[\text{SE}^2_{\hat{\beta}_i}] + \text{SE}^2_{\hat{\beta}_i}}$. The shrunk estimator, β^s_i, puts more weight on estimates of β_i that are more precisely estimated, as measured by a low standard error. I compute the shrunk estimator of ability analogously.

Heteroskedasticity-robust standard errors clustered at the individual level in parentheses.***, ** and * denote significance at 10%, 5% and 1% levels, respectively.

C The ENEM

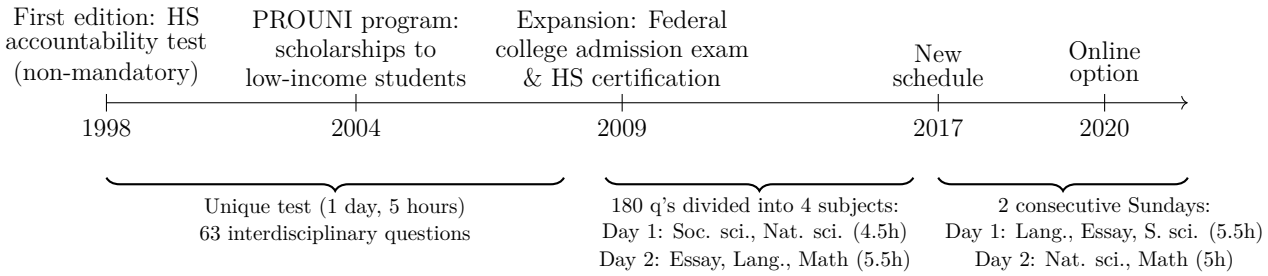
In this Appendix, I describe the changing role of the ENEM in the higher-education system over time, compare the ENEM to the US SAT and ACT exams, and describe the IRT grading system used by the Ministry of Education to generate ENEM test scores.

C.1 The Role of the ENEM in the Higher-education System

The ENEM was created in 1998 by the National Institute of Educational Studies (INEP), a unit of the Brazilian Ministry of Education, with the goal of evaluating student performance at the end of high school (Appendix Figure C1). The ENEM is an achievement test, that is, it was designed to test for mastery of material individuals should learn by the end of high school.[26]

The first ENEM contained 63 multiple-choice interdisciplinary questions and was conducted over a five-hour testing block. The test score was calculated as the fraction of correct responses. In its first edition, fewer than 200,000 individuals enrolled to take the ENEM.

Figure C1: Timeline of the ENEM



In 2004, the government created a college scholarship program for low-income students called ProUni (*Programa Universidade para Todos*). ProUni used ENEM scores to allocate the scholarships to applicants, with program-specific score cutoffs based on the number of seats available in each program. After ProUni was implemented, the number of individuals who signed up to take the ENEM doubled from 1.5 million in 2004 to 3.0 million in 2005.

[26] Researchers often divide standardized tests into two types: reasoning tests and achievement tests. Reasoning tests measure a student's verbal reasoning, critical reading, and skills. Achievement tests measure a student's mastery of specific subjects, like biology or physics. In practice, performance on both types of tests is highly correlated (Soares, 2015).

In 2009, the Ministry of Education reformed the ENEM with the aim of encouraging colleges to use it as an admission exam. The new ENEM consists of 180 multiple-choice questions conducted over two consecutive days of testing during a weekend. The new exam contains questions in four subjects: mathematics, natural sciences (which includes biology, physics, and chemistry questions), social sciences (which includes history, geography, philosophy, and sociology questions), and language arts (which includes questions on Portuguese language, literature, foreign language, arts, physical education, and information and communication technologies). On the first day of testing, individuals had five and a half hours to take the social science test, the natural science test, and the essay. On the second day of testing, individuals had five hours to take math and language arts tests. The new ENEM is graded according to Item Response Theory (IRT), which enables colleges to compare test scores over time (see Appendix C.4).

In 2010, the Ministry of Education introduced a centralized admission system called SISU (*Sistema de Seleção Unificada*) with the goal of simplifying the college application process for federal universities. The centralized system used ENEM scores to allocate students to participating colleges. All federal universities are part of the system, but other universities (including state and municipal universities) are not mandated to be part of it. Also in 2010, the Government started using ENEM scores to allocate student loans through a program called FIES (*Fundo de Financiamento ao Estudante do Ensino Superior*). In addition, starting in 2010 (and finishing in 2016), ENEM scores could be used to certify the attainment of high-school-level skills (analogously to the GED in the US). By 2010, over 4.6 million individuals enrolled to take the ENEM.

In 2017, INEP changed the schedule of the ENEM. The exam started being conducted over two consecutive Sundays. On the first Sunday, individuals have five and a half hours to answer the language arts test, the social science test, and the essay. On the second Sunday, individuals have five hours to answer the natural science and math tests. The other features of the exam remained constant.

In 2020, individuals had the option to take the ENEM through a computer without internet access. Over 5.7 million individuals enrolled to take the ENEM this year.

C.2 ENEM Sample Questions

Appendix Figures C2–C5 present sample questions from the natural science, social science, language arts, and math components of the ENEM. The questions come from the 2016 ENEM. The questions are average in terms of their difficulty.

Figure C2: Natural Science sample question (item #11898)

Panel A. Original (in portuguese)

Portadores de diabetes *insipidus* reclamam da confusão feita pelos profissionais da saúde quanto aos dois tipos de diabetes: *mellitus* e *insipidus*. Enquanto o primeiro tipo está associado aos níveis ou à ação da insulina, o segundo não está ligado à deficiência desse hormônio. O diabetes *insipidus* é caracterizado por um distúrbio na produção ou no funcionamento do hormônio antidiurético (na sigla em inglés, ADH), secretado pela neuro-hipófise para controlar a reabsorção de água pelos túbulos renais.

Tendo em vista o papel funcional do ADH, qual é um sintoma clássico de um paciente acometido por diabetes *insipidus*?

(A) Alta taxa de glicose no sangue.

(B) Aumento da pressão arterial.

(C) Ganho de massa corporal.

(D) Anemia crônica.

(E) Desidratação.

Panel B. Translation

Patients with diabetes *insipidus* complain about the confusion made by health professionals about the two types of diabetes: *mellitus* and *insipidus*. While the first type is associated with insulin levels or action, the second is not linked to insulin deficiency. Diabetes insipidus is characterized by a disturbance in the production or functioning of the antidiuretic hormone (ADH), secreted by the neurohypophysis to control the reabsorption of water by the renal tubules.

In view of the functional role of ADH, what is a classic symptom of a patient with diabetes *insipidus*?

(A) High blood glucose.

(B) Increase in blood pressure.

(C) Body mass gain.

(D) Chronic anemia.

(E) Dehydration.

Notes: The correct answer is underlined.

Figure C3: Social Science sample question (item #97290)

Panel A. Original (in portuguese)

Parceria Transpacífica

Dentro das atuais redes produtivas, o referido bloco apresenta composição estratégica por se tratar de um conjunto de países com

Ⓐ Elevado padrão social.

Ⓑ Sistema monetário integrado.

Ⓒ Alto desenvolvimento tecnológico.

Ⓓ Identidades culturais semelhantes.

Ⓔ Vantagens locacionais complementares.

Panel B. Translation

Trans-Pacific Partnership

Within the current production networks, the aforementioned bloc has a strategic composition because it is a group of countries with:

Ⓐ High social standard.

Ⓑ Integrated monetary system.

Ⓒ High technological development.

Ⓓ Similar cultural identities.

Ⓔ Complementary locational advantages.

Notes: The correct answer is underlined.

Figure C4: Language Arts sample question (item #86509)

Panel A. Original (in portuguese)

O último longa de Carlão acompanha a operária Silmara, que vive com o pai, um ex-presidiário, numa casa da periferia paulistana. Ciente de sua beleza, o que lhe dá certa soberba, a jovem acredita que terá um destino diferente do de suas colegas. Cruza o caminho de dois cantores por quem é apaixonada. E constata, na prática, que o romantismo dos contos de fada tem perna curta.

VOMERO, M. F. Romantismo de araque. **Vida Simples**, n. 121, ago. 2012.

Reconhece-se, nesse trecho, uma posição crítica aos ideais de amor e felicidade encontrados nos contos de fada. Essa crítica é traduzida

Ⓐ Pela descrição da dura realidade da vida das operárias.

Ⓑ Pelas decepções semelhantes às encontradas nos contos de fada.

Ⓒ Pela ilusão de que a beleza garantiria melhor sorte na vida e no amor.

Ⓓ Pelas fantasias existentes apenas na imaginação de pessoas apaixonadas.

Ⓔ Pelos sentimentos intensos dos apaixonados enquanto vivem o romantismo.

Panel B. Translation

Carlão's latest feature follows the worker Silmara, who lives with her father, an ex-convict, in a house on the outskirts of São Paulo. Aware of her beauty, which gives her a certain arrogance, the young woman believes that she will have a different destiny from her colleagues. She crosses paths with two singers she is in love with. And she finds, in practice, that the romanticism of fairy tales has short legs.

VOMERO, M. F. Romanticism of arak. **Simple Life**, n. 121, Aug. 2012.

This passage recognizes a critical position on the ideals of love and happiness found in fairy tales. This criticism is translated

Ⓐ For the description of the harsh reality of the workers' lives.

Ⓑ For disappointments similar to those found in fairy tales.

Ⓒ For the illusion that beauty would guarantee better luck in life and in love.

Ⓓ For the fantasies that exist only in the imagination of people in love.

Ⓔ For the intense feelings of those in love while living romanticism.

Notes: The correct answer is underlined.

Figure C5: Math sample question (item #37515)

Panel A. Original (in portuguese)

Para evitar uma epidemia, a Secretaria de Saúde de uma cidade dedetizou todos os bairros, de modo a evitar a proliferação do mosquito da dengue. Sabe-se que o número f de infectados é dado pela função $f(t) = -2t^2 + 120t$ (em que t é expresso em dia e $t = 0$ é o dia anterior à primeira infecção) e que tal expressão é válida para os 60 primeiros dias da epidemia.

A Secretaria de Saúde decidiu que uma segunda dedetização deveria ser feita no dia em que o número de infectados chegasse à marca de 1600 pessoas, e uma segunda dedetização precisou acontecer.

A segunda dedetização começou no

(A) 19° dia.

(B) 20° dia.

(C) 29° dia.

(D) 30° dia.

(E) 60° dia.

Panel B. Translation

To prevent an epidemic, the Health Department of a city sprayed all neighborhoods, in order to prevent the proliferation of the dengue mosquito. It is known that the number f of infected people is given by the function $f(t) = -2t^2 + 120t$ (where t is expressed in day and $t = 0$ is the day before the first infection) and that this expression is valid for the first 60 days of the epidemic.

The Health Department decided that a second extermination should be carried out on the day when the number of infected people reached the mark of 1,600 people, and a second extermination had to take place.

The second extermination started in

(A) 19th day.

(B) 20th day.

(C) 29th day.

(D) 30th day.

(E) 60th day.

Notes: The correct answer is underlined.

C.3 Comparison of the ENEM to the ACT and SAT exams

Appendix Table C1 compares important features of the SAT, ACT, and ENEM. The SAT contains 154 multiple-choice questions divided into three sections: reading, writing and language, and math, plus an optional essay. Including the essay, individuals have 3 hours and 50 minutes to take the test. On average across sections, test-takers have about 1 minute and 10 seconds to answer each question. Raw scores are converted into scaled scores through a score conversion table.

The ACT contains 215 multiple-choice questions divided into four sections: English, math, reading, and science, plus an optional essay. Including the essay, individuals have 3 hours and 35 minutes to take the test. On average across sections, test-takers have less than 1 minute to answer each question. Raw scores are converted into scaled scores through a score conversion table.

There are some notable differences between the SAT/ACT and the ENEM. First, the ENEM is conducted over two days of testing. Second, individuals in the ENEM have no assigned breaks. Third, the booklet ENEM test-takers receive contains all the questions they have to answer during the testing day. Thus, they may allocate time disproportionally across sections. In contrast, in the SAT and ACT, each section has an assigned amount of time. Finally, in the ENEM, each question is associated with a different text passage or prompt (in some cases, two questions share a prompt or passage). In contrast, in the SAT and ACT, a given passage is associated with multiple questions. This partly explains why the time per question is higher in the ENEM than in the ACT/SAT.

Table C1: Comparison of the SAT, ACT, and ENEM college admission exams

	SAT	ACT	ENEM
Cost	∼$60	∼$88	∼$17
Grading	Score conversion chart using raw scores	Score conversion chart using raw scores	Item Response Theory (IRT)
Starting time	Between 8:30 and 9am	Between 8:30 and 9am	1pm Brasilia time
Number of items	154 questions	215 questions	180 questions
Total length	3 hours and 50 mins over 1 testing day	3 hours and 35 mins over 1 testing day	10 hours over 2 testing days
Time per question	1 minute and 10 seconds	50 seconds	3 minutes
Breaks	10 mins break after reading section 5 min break between math sections 2 min break before the essay	10 min break after math section 5 min break before essay	N/A
Sections	Reading (65 mins, 52 items) Writing and Language (35 mins, 44 items) Math w/o calculator (25 mins, 20 items) Math w/ calculator (55 mins, 38 items) Optional essay (50 mins)	English (45 mins, 75 items) Math (60 mins, 60 items) Reading (35 mins, 40 items) Science (35 mins, 40 items) Optional essay (40 mins)	Social science (day 1, 45 items) Natural science (day 1, 45 items) Language arts (day 2, 45 items) Math (day 2, 45 items) Mandatory essay (day 2)

Notes: The SAT refers to the post-2016 version of the SAT, which includes an optional essay. This optional essay was eliminated in 2021. The ENEM refers to the 2009–2016 version of the exam (see Section C.1 for information on the pre-2009 and post-2016 versions). The exam length was computed excluding breaks and including the essay. The time per question does not account for the essay.

C.4 IRT Grading

The Brazilian Testing Agency grades the ENEM exam based on the three-parameter item response theory (IRT). According to IRT, the probability that an individual i with ability θ_i correctly answers question j is:

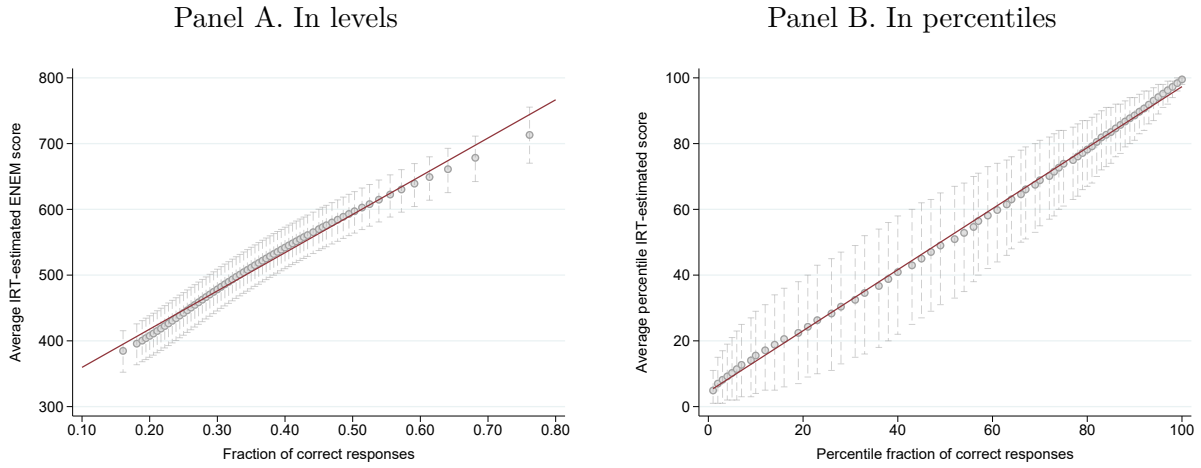
$$\Pr(C_{ij} = 1|\theta_i) = p_j(\theta_i) = c_j + \frac{1 - c_j}{1 + e^{-a_j(\theta_i - b_i)}}, \tag{C1}$$

where a_j, b_j, and c_j are three question-level parameters that represent, respectively, a question's "discrimination," "difficulty," and "pseudo-guess." A question's discrimination refers to its ability to discriminate between low- and high-ability individuals; the difficulty represents the value of θ at which $p_j(\theta_i)$ has the maximum slope, and the pseudo-guess parameter indicates the likelihood that a student with an infinitely negative ability has to correctly respond to the question. Notice that in equation (C1), the probability of correctly answering a question does not depend on its position. Thus, the type of position effects documented above suggests that the IRT estimates of individual-level ability are biased. Modern IRT approaches (e.g., Debeer and Janssen, 2013) include item position into the framework.

Each question's parameters are known from pre-testing. The testing agency estimates the θ_i that maximizes the empirical likelihood of the entire sequence of responses. They do this separately for each student and academic subject. ENEM scores are normalized to have a mean of 500 and a standard deviation of 100.

Despite its complexity, most of the variation in IRT-estimated ENEM scores is driven by variation in the fraction of correct responses in the exam. A regression of IRT-estimated ENEM scores on the fraction of correct responses yields an R-squared of 0.88 (the rank correlation between the two variables is 0.93). Consistent with this, Appendix Figure C6 shows that the relationship between these two variables is linear in both levels (Panel A) and percentiles (Panel B). The strong relationship between IRT-estimated scores and the fraction of correct responses holds not just for the overall score but also for the score in each academic subject (Appendix Table C2).

Figure C6: Comparison of IRT-estimated ENEM score and fraction of correct responses



Notes: This figure shows binned scatterplots plotting the average IRT-estimated ENEM score across all four academic subjects (y-axis) against the fraction of correct responses on the exam (x-axis). Panel A shows the results in levels and Panel B in percentiles. I first group students into 100 equally-sized bins based on their fraction of correct responses. Then, I calculate the average IRT-estimated ENEM score or score percentile in each bin. The vertical lines denote the 10th and 90th percentiles of the ENEM score distribution. The solid red line shows the predicted values from a linear regression on the plotted points.

Table C2: Correlation between IRT-estimated ENEM score and fraction of correct responses on each subject

	Academic subject				
	Social science	Natural science	Language arts	Math	Average score
	(1)	(2)	(3)	(4)	(5)
Panel A. Variables measured in levels					
Fraction correct resp.	0.892***	0.880***	0.907***	0.885***	0.937***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
N	14,936,699	14,936,699	14,936,699	14,936,699	14,936,699
R−squared	0.79	0.77	0.82	0.78	0.88
Panel B. Variables measured in percentiles					
Fraction correct resp.	0.904***	0.858***	0.917***	0.845***	0.931***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
N	14,936,699	14,936,699	14,936,699	14,936,699	14,936,699
R−squared	0.82	0.74	0.84	0.71	0.87

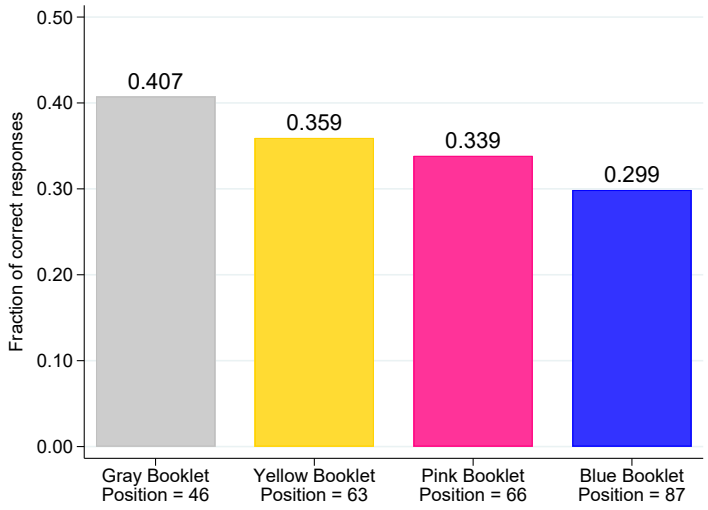
Notes: This table displays the correlation between the IRT-estimated ENEM score and the fraction of correct responses. Columns 1–4 present the correlations separately for each academic subject. Column 5 presents the correlation between the average score across all subjects and the fraction of correct responses in the entire exam. Heteroskedasticity-robust standard errors clustered at the question level in parentheses. ***, ** and * denote significance at 10%, 5% and 1% levels, respectively.

D Measuring Position-Adjusted Question Difficulty

In this Appendix, I describe my measures of position-adjusted question difficulty. Instead of taking a strong stance on what the right measure of position-adjusted question difficulty is, I show that the results are robust to measuring this variable in several ways.

An intuitive measure of a question's difficulty is the fraction of students who correctly answer the question. This measure is problematic in the presence of fatigue effects since a given question has a different fraction of correct responses depending on its location. To illustrate this problem, Appendix Figure D1 plots students' performance on a natural science question in each booklet. The position of this item ranged from position 46 in the gray booklet to position 87 in the blue booklet. Correspondingly, student performance varied from 40.7% in the gray booklet to 29.9% in the blue booklet.

Figure D1: Performance on a natural science question (item #11898)



Notes: This figure shows the fraction of individuals who correctly responded to item #11898 in each of the four booklets. See Appendix Figure C2 for the question's text.

The fact that performance on a question varies according to its position raises an important challenge for measuring question difficulty. It is hard to know whether questions that appear later in the exam are less likely to be correctly answered because they test more difficult material or because students are more fatigued by the time they get to these questions.

To account for fatigue effects, I estimate measures of question difficulty that represent

the fraction of students who would correctly answer a question if the question appeared in the first position of the exam. To estimate this fraction, I follow a three-step process. First, I compute the average position of each question across all booklets. Second, I estimate the effect of a one-position increase of a question position on performance on the question ("position effect"). Third, I multiply the average question position calculated in the first step by the position effect estimated in the second step and subtract this figure from the fraction of correct responses across all booklets. This yields a position-adjusted estimate of question difficulty. Appendix Table D1 illustrates these steps in calculating the difficulty of item #11898.

The measures of question difficulty differ in how I estimate the position effect in the second step. My baseline measure of question difficulty uses the position effect estimated by pooling all questions (Table 2, column 3). This measure assumes that the effect of a one-position increase on performance is homogeneous across questions.

The second measure of question difficulty uses a question-specific position effect. I estimate equation (3) separately for each question and use the intercept from the regression as the measure of difficulty. This does not assume homogeneity in position effects; however, for some questions the position effect is imprecisely estimated.

The third measure of question difficulty combines the first two by shrinking the question-specific position effect to the average effect by its signal-to-noise ratio. Specifically, let β_j be the position effect estimating using data only from question j and $\bar{\beta}$ be the average position effect across all questions. The shrunk position effect of question j, β_j^s, is a convex combination of β_j and $\bar{\beta}$:

$$\beta_j^s = \omega_j \beta_j + (1 - \omega_j)\bar{\beta},$$

where the question-specific weight, ω_j, is

$$\omega_j = \frac{\mathbb{V}\text{ar}[\hat{\beta}_j] - \mathbb{E}[\text{SE}^2_{\hat{\beta}_j}]}{\mathbb{V}\text{ar}[\hat{\beta}_j] - \mathbb{E}[\text{SE}^2_{\hat{\beta}_j}] + \text{SE}^2_{\hat{\beta}_j}}.$$

The shrunk estimator puts more weight on position effects that are more precisely estimated, as measured by a low standard error of $\hat{\beta}_j$, $\text{SE}^2_{\hat{\beta}_j}$.

The fourth measure estimates the position effect separately for questions with a below/above median fraction of correct responses. The fifth measure estimates the effect separately for each academic subject. These measures assume that the effect of a one-

position increase on performance is homogeneous within a type of question.

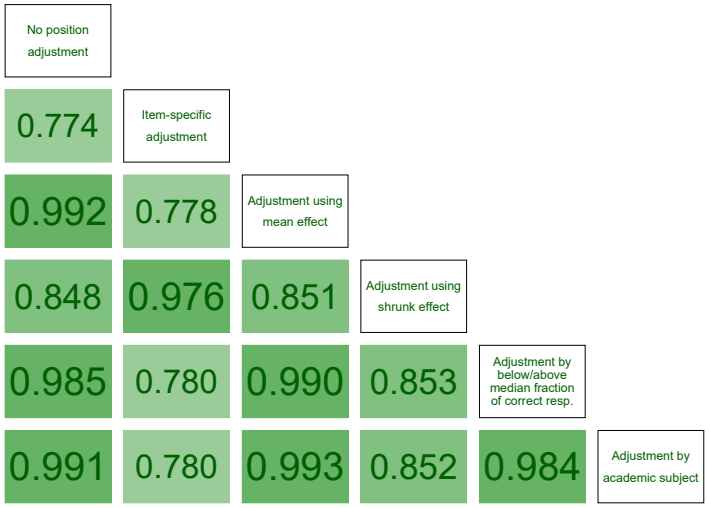
Table D1: Alternative measures of the difficulty of item #11898

Position effect estimation method (1)	Average fraction correct responses (2)	Fatigue effect (in pp) × average position (3)	Question difficulty (4)
None	0.36	0 × 64 = 0	0.36
Pooling all items	0.36	-0.08 × 64 = -5.1	0.41
Item-specific effect	0.36	-0.24 × 64 = -15.3	0.51
Shrinkage estimator	0.36	-0.24 × 64 = -15.3	0.51
By fraction corr. resp.	0.36	-0.15 × 64 = -9.6	0.45
By academic subject	0.36	-0.03 × 64 = -1.6	0.37

Notes: This table illustrates how the six measures of a question's difficulty are calculated. The average fraction of correct responses and the average question position are calculated using the number of students with each booklet as weights.

Appendix Figure D2 shows the cross-question correlation between the measures of question difficulty. Reassuringly, all difficulty measures are highly correlated, with coefficients ranging from 0.77 to 0.99.

Figure D2: Cross-question correlation matrix of item difficulty measures



Notes: This figure shows the relationship between the different measures of question difficulty. Each cell shows the cross-question linear correlation between two measures of question difficulty. The sample size is $N = 1,842$ across all cells.