Stakes and Signals: An Empirical Investigation of Muddled Information in Standardized Testing

Germán Reyes
Evan Riehl
Ruqing Xu

October 2023

Abstract. Muddled information models posit that higher stakes increase a signal's informativeness about individuals' *gaming ability* and decrease its informativeness for their *natural ability*. An important question is whether this muddling of abilities degrades the predictive value of a signal for long-run outcomes. We evaluate this question in the context of standardized testing by exploiting the introduction of Brazil's national college admission exam, the ENEM. The staggered adoption of the ENEM by universities meant that, depending on their location and cohort, students either took a low-stakes school accountability test or a high-stakes test that governed admission to the most selective colleges in their area. Using ENEM records linked to nationwide college and labor market data, we find that the increase in the stakes of the ENEM exam made scores *more* informative for students' longer-run outcomes. However, test score gaps between high- and lower-income students also expanded on the higher-stakes ENEM exam. Our results show that signals that include gaming ability can be more informative about individual productivity than signals that measure only natural ability, but they can also exacerbate socioeconomic inequality.

Reyes: briq Institute on Behavior & Inequality, Schaumburg-Lippe-Str. 5-9, 53113 Bonn, Germany (email: german.reyes@briq-institute.org). Riehl (corresponding author): Department of Economics, Cornell University, 266 Ives Hall, Ithaca, NY 14853 (email: eriehl@cornell.edu). Xu: Department of Economics, Cornell University, 449B Uris Hall, Ithaca, NY 14853 (email: rx24@cornell.edu).

1. Introduction

Selecting the right talent is crucial to the success of organizations. A fundamental challenge faced by recruiters is incomplete information on candidate quality. To address this challenge, recruiters often rely on signals of candidate quality from mechanisms like standardized tests or case interviews. Yet candidates have an incentive to manipulate their signals through preparation or even outright cheating, particularly when the positions they are applying for are highly desirable.[1]

An important question for the design of talent selection mechanisms is whether the potential for manipulation degrades the quality of the signal. Theoretical models of "muddled information" (Frankel and Kartik, 2019) show that, as the stakes of a mechanism increase, signals become more informative about an individual's *gaming ability* and less informative about the *natural action* individuals would take absent signaling concerns.[2] Yet, as Frankel and Kartik note, it is unclear whether recruiters would prefer to observe signals of candidate quality with or without gaming. Gaming ability may reflect a candidate's knowledge of the recruiting mechanism or manipulation skills that are unrelated to productivity, but it could also reflect work ethic, interest, or other desirable attributes.

This paper conducts an empirical evaluation of the relationship between stakes and signal quality in the context of admission exams for elite universities. We exploit a unique natural experiment in Brazil in which a low-stakes high school accountability test was repurposed into a high-stakes admission exam for the country's most selective universities. Our empirical strategy holds the structure of the exam and the composition of exam takers fixed and asks how the increase in exam stakes impacted two important outcomes: 1) test score gaps between advantaged and disadvantaged students; and 2) the predictive power of scores for college and labor market success.

It is *ex ante* unclear how an increase in the stakes of an exam affects inequality in test scores and their informativeness for student outcomes. Absent signaling concerns, the students who perform better on exams may be those with high intrinsic motivation, conscientiousness, and aptitude (Kreps, 1997; Bénabou and Tirole, 2003). This is consistent with a widely-held belief in education research that low-stakes tests are a better measure of student learning than high-stakes tests because there is less of an incentive to manipulate performance (e.g., Amrein and Berliner, 2002). Further, it is often said that the use of high-stakes tests in college admissions helps wealthy students "game the system" through expensive test prep

[1] See, for example, "Actresses, Business Leaders and Other Wealthy Parents Charged in U.S. College Entry Fraud," Jennifer Medina, Katie Benner and Kate Taylor, *The New York Times*, March 12, 2019.

[2] This phenomenon is broadly known in the social sciences as "Campbell's Law," named after psychologist Donald T. Campbell, who noted that "the more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor" (Campbell, 1979).

and other score-boosting strategies (Buchmann et al., 2010; Soares, 2015).[3] Thus an increase in exam stakes could both widen test score gaps between high- and lower-income students and degrade the quality of the test score signal. On the other hand, gaming ability could reflect skills that are important for college success like work ethic, grit, or the capacity to learn new material. The distribution of these skills may be less related to family background than the distribution of low-stakes test scores. In this case, higher exam stakes could make test scores more informative about college potential and also reduce socioeconomic inequality in test scores.

To provide empirical evidence on these relationships, we examine the rollout of a national standardized admission exam for elite Brazilian universities. From 2009–2017, Brazil's system of highly selective federal universities transitioned from their own university-specific admission exams to a common test called the ENEM (*Exame Nacional do Ensino Médio*). Federal universities in different states varied in the timing at which they adopted the ENEM in admissions (Machado and Szerman, 2021; Mello, 2022). Importantly, the ENEM was also used as a high school accountability test, and so many high school seniors took the exam regardless of its role in college admissions. Since most Brazilian students attend college close to home, this meant that ENEM participants were either taking a low-stakes (from their perspective) school accountability test or a high-stakes exam that governed admission to the most selective universities in their area, depending on their location and cohort. We define a sample of high school seniors who took the ENEM in 2009–2017 regardless of its role in college admissions, and exploit the variation in exam stakes across states and cohorts in a difference-in-differences design.

To implement our design, we link administrative records from the ENEM exam to nationwide college and labor market data. The ENEM data include individuals' responses to each exam question, which allows us to ask how the increase in exam stakes affected students' overall scores in each subject as well as their performance on different types of exam questions. We measure college enrollment, persistence, and graduation outcomes using the 2010–2019 waves of Brazil's higher education census. Lastly, we measure labor market wages using Brazil's national employer-employee data for the years 2016–2018. Using these data,

[3] Goodman et al. (2020) show that higher-income students are more likely to retake the SAT, which raises their admission-relevant scores. Affluent students are also more likely to hire tutors and procure special test accommodations like extended time. See: "SAT/ACT tutoring: $1500 for 90 minutes. And 14 sessions are required. Really," Valerie Strauss, *The Wall Street Journal*, August 31, 2014; and "Many More Students, Especially the Affluent, Get Extra Time to Take the SAT," Douglas Belkin, Jennifer Levitz and Melissa Korn, *The Wall Street Journal*, May 21, 2019.

we ask how the increase in the stakes of the ENEM impacted test score gaps between advantaged and disadvantaged students and the informativeness of scores as measured by the correlation coefficient between test scores and outcomes (Rothstein, 2004).[4]

We have two main findings. First, test score gaps between high- and lower-income students widened on the higher stakes exam. Specifically, gaps in average ENEM scores between private and public high school students expanded by roughly 10 percent when federal universities adopted the ENEM in admissions (relative to the mean gaps in pre-adoption cohorts). This increase was driven by private school students earning higher scores on the high-stakes exam, with the largest impacts on the math subject test. Racial and other socioeconomic test score gaps expanded by a similar percentage. Our estimates imply a significant increase in the selectivity of college programs that private school students could gain admission to.

Second, the increase in exam stakes caused ENEM scores to become *more* informative for students' academic potential. The adoption of the ENEM exam by federal universities increased the correlation coefficients between ENEM scores and students' college persistence and graduation outcomes by roughly 10–30 percent, depending on the outcome measure.[5] We find some evidence that the high-stakes ENEM scores were also more informative for earnings outcomes, although our labor market data is measured too early in students' careers to be conclusive. Overall, our results show that the high-stakes ENEM exam improved colleges' ability to identify students who were likely to succeed.

To shed light on mechanisms, we show that the higher-stakes test led to an improvement in private school students' performance across a wide range of exam skills. The adoption of the ENEM by federal universities led to especially large increases in the scores of students who attended private high schools with test prep oriented curriculum, suggesting that our results are partly driven by ENEM-specific exam preparation. Yet our analysis of question-level responses shows that private students performed better across a wide range of exam skills that are designed to reflect high school and college curricula. Further, exam skills in which we observe larger improvements in private students' performance also tend to be *more* predictive of college and labor market outcomes relative to other skills. This suggests that test prep for the higher-stakes ENEM was not confined to narrowly-targeted skills that merely raise exam scores; rather, it covered a broad set of skills that are informative about academic potential.

[4] An alternative approach to measuring the information content is to "anchor" them to an economic outcome of interest (Cawley et al., 1999; Jacob and Rothstein, 2016; Nielsen, 2023, e.g.,). We focus on the correlation between scale scores and outcomes because scale scores are what colleges use to admit students.

[5] The predictive power of scores increased both overall and measured among students who attended the same college programs. This shows that our findings are not driven by a causal impact of the scores on the programs students attended, but rather by an increase in the informativeness of scores.

To our knowledge, our paper provides the first empirical test of how an increase in the stakes of a talent selection mechanism impacts that quality of the signal from that mechanism, which informs theoretical work on muddled information (Fischer and Verrecchia, 2000; Frankel and Kartik, 2019, 2022). A key result in this literature is that as stakes rise, test scores become more informative about students' *gaming ability* (the marginal cost of improving test scores as stakes increase) and less informative for individuals' *natural action* (the test score without signaling incentives). However, these models are silent on whether gaming ability is informative for the individual's likelihood of succeeding in college, which is a key outcome that colleges care about in determining admissions. Our findings show that, at least in the case of college admissions, test scores that include gaming ability are more informative than those that reflect only the natural action. Our question-level results suggest that gaming ability may largely reflect broad learning capacity or non-cognitive skills that are beneficial in college, such as work ethic. This is consistent with work in behavioral economics that finds that mental effort and motivation can influence performance on cognitively demanding tasks (Wise and DeMars, 2005; Duckworth et al., 2011; Segal, 2012; Finn, 2015; Jalava et al., 2015; Gneezy et al., 2019; Schlosser et al., 2019).

Our paper is also unique in providing evidence that high-stakes tests can be more informative for outcomes that educators care about than low-stakes tests. A large literature finds that educators respond strategically to high-stakes testing by focusing on narrowly-defined skills (Jones et al., 1999; Jacob, 2005), manipulating the population of test takers (Figlio and Getzler, 2006; Cullen and Reback, 2006), focusing on specific subjects or students (Neal and Schanzenbach, 2010; Reback et al., 2014), or even outright cheating (Jacob and Levitt, 2003). Related work finds that gains in high-stakes test scores from accountability policies do not always translate to improved performance on low-stakes tests (Klein et al., 2000; Jacob, 2005; Corcoran et al., 2011). For this reason, many education researchers have argued that low-stakes exam scores are a better measure of student learning (e.g., Koretz and Barron, 1998; Amrein and Berliner, 2002). Yet these studies do not typically have the data to estimate how the informativeness of scores varies between high- and low-stakes exams. Our findings suggest that, at least in the context of college admission exams, exams with higher stakes can provide a better signal of students' future academic potential.

Lastly, our findings show that there is a tradeoff between equality and informativeness in the use of high-stakes college admission exams. Concerns about the fairness of admission tests have led many U.S. colleges to adopt test-optional or test-blind admission policies in recent years. There is research that finds that the design and implementation of admission exams can exacerbate inequality in college access (Hoxby and Turner, 2013; Bulman, 2015; Pallais, 2015; Goodman, 2016; Bhattacharya et al., 2017; Goodman et al., 2020; Reyes, 2023). Other work has found mixed results on the degree to which admission scores provide valuable information about academic potential as measured by the correlation between exams scores and student outcomes (Rothstein, 2004; Scott-Clayton, 2012; Bettinger et al., 2013; Riehl, 2023). But a limitation with papers in both of these literatures is that they focus on a static admissions environment, so it is unclear how inequality or informativeness would change if universities used alternative admission criteria. Our paper shows that changes in admission criteria can alter the quality of admission signals as prospective students respond to the new system. Consistent with a common criticism of admission exams, our findings show that test prep for high-stakes exams can shift the distribution of scores toward wealthier students. This suggests that the current movement away from high-stakes admission tests and toward lower-stakes signals like high school grades will help U.S. colleges admit more socioeconomically diverse classes.[6] But colleges also want to admit students who can succeed in their programs, and our findings show that colleges may have a harder time identifying such students in the absence of high-stakes testing.

2. Institutional background

2.1. **Colleges and high schools in Brazil.** The higher education system in Brazil is heavily privatized, but the most prestigious institutions tend to be in its system of *federal universities*. In 2009, there were 59 federal universities, with a presence in all of Brazil's 27 states. Together, federal universities account for about 11 percent of total college enrollment. Brazil also has a system of 40 *state universities* managed by the governments of each state. Federal and state universities are tuition-free, highly selective, and consistently top the national college rankings. The Brazilian higher education system additionally includes over 2,000 private universities and technical colleges that makeup roughly 80 percent of total enrollment. While a handful of these private institutions are elite and selective, the majority are moderately selective or follow an open enrollment policy.[7]

The situation is reversed at the secondary level, where private high schools represent a small but socioeconomically-advantaged share of enrollment. In 2009, 14 percent of secondary students attended a private high school, while 85 percent attended a public school managed by the state government.[8] Importantly, private high school students are vastly overrepresented in the higher education system. They accounted for 40 percent of all incoming college students in 2009 and made up 47 percent of federal university enrollees (Appendix Table A1).

[6] Related work shows that high-stakes exams can increase gender test score gaps due to competitive pressures (Ors et al., 2013; Azmat et al., 2016; Cai et al., 2019).

[7] Appendix Table A1 provides summary statistics on Brazilian high schools and colleges in 2009.

[8] Roughly 0.5 percent of Brazilian students attend a high school managed by the federal government (Appendix Table A1). We define "private high schools" to include both private and federal high schools since their students are similar in terms of socioeconomic status and achievement.

2.2. **Federal university admissions and the ENEM exam.** Admission to federal universities is highly competitive and relies exclusively on test scores from entrance exams. Before 2009, each federal university designed and administered its admission tests, known as *vestibular* exams. Thus, it was burdensome for students to apply to more than one university as they had to prepare for multiple tests and travel to each school on a specific date to sit for the exam.

To centralize federal university admissions, the Ministry of Education developed a national standardized college admission exam called the ENEM (*Exame Nacional do Ensino Médio*). The ENEM exam was initially created in 1998 for the purpose of high school accountability. Between 1998 and 2008, the ENEM exam consisted of 63 multi-disciplinary questions and an essay. In 2009, the Ministry redesigned and expanded the exam so that it could serve as a tool for college admissions. The post-2009 ENEM exam resembles the ACT exam in the United States; it contains 180 questions across four distinct subject areas (math, language arts, natural sciences, and social sciences) along with a written essay. The exam spans two days of testing and is taken by over five million students each November, making it the second-largest admission test globally. As part of the centralization effort, the Ministry also created a unified admission platform called SISU (*Sistema de Seleção Unificada*), which allocates students to colleges on the basis of their preferences and ENEM scores.

Although the college admission version of the ENEM exam began in 2009, federal universities varied in the timing at which they switched from their institution-specific tests to the ENEM exam. The Ministry of Education provided financial incentives to adopt the ENEM, but universities had unilateral control over their admission methods, and some were initially uncertain about the content of the new ENEM (Machado and Szerman, 2021).[9] Thus, some federal universities began using the ENEM immediately in 2009, while others adopted it five or more years later.[10] The variation in the timing at which federal universities adopted the ENEM exam is the basis of our empirical strategy, as we describe Section 3.

2.3. **The market for test prep.** Preparation for the ENEM or *vestibular* exams is a central part of the lives of many Brazilian students as they approach the end of high school. In Brazil, much of this exam preparation takes place at private high schools during school hours. Many students who hope to gain admission to elite universities choose to attend private high schools that use test-oriented curricula designed by for-profit companies (e.g., *Sistema Anglo* and *Sistema pH*). These schools offer multiple courses throughout the day that focus specifically on exam subjects and preparation strategies. Private schools frequently tout the successful

[9] See Machado and Szerman (2021) and Mello (2022) for details on the implementation of the ENEM/SISU system and its adoption by universities.

[10] Some state universities also adopted the ENEM as their admission test, but to this date, many still design and administer their own *vestibular* exams.

exam performance of prior cohorts in an effort to attract new students, and newspapers publish annual rankings of school-mean scores.[11] In addition, many Brazilian students study for the exams outside of school hours or after completing high school by taking for-profit prep courses known as *cursinhos* (Mitrulis and Penin, 2006; Fernandes, 2015).

The existence of this for-profit market for test prep raises a concern about inequality in access to Brazil's selective universities. Although there are a growing number of non-profit and free online services (e.g., *Descomplica*), test prep remains heavily concentrated in the private sector. Lower-income students often cannot afford to enroll in private high schools or *cursinhos*. In public school curricula, there is significantly less emphasis on preparing students for the ENEM exam. In this context, our empirical analysis examines whether the existence of high-stakes admission exams increases inequality in college access.

3. Data and identification

3.1. **Data.** Our base dataset includes administrative records on all individuals who took the ENEM exam in 2007–2017 (INEP, 2019a). This dataset is compiled by the National Institute of Educational Studies, or INEP (*Instituto Nacional de Estudos e Pesquisas Educacionais*). The data contains scores on each exam subject, demographic characteristics, and information on individuals' high schools. The data also contains individuals' responses to each exam question, which allows us to observe which questions individuals got right and wrong. Lastly, we observe information on the content of each question, including the learning objectives, the Item Response Theory (IRT) parameters, and the text of the question.

To measure longer-run outcomes, we link the ENEM data to two other administrative datasets at the individual level.[12] First, we measure college outcomes by linking to INEP's higher education census (*Censo da Educação Superior*) for the years 2010–2019 (INEP, 2022). This dataset contains information on the universe of students who were enrolled in the Brazilian higher education system in these years, including each student's university, major, admission method, enrollment year, and graduation/drop-out outcome.

Second, we measure labor market outcomes by linking to Brazil's employee-employer dataset, the RAIS (*Relação Anual de Informações Sociais*), for the years 2016–2018 (RAIS, 2022). The RAIS is maintained by the Ministry of Labor and covers the entire population of formal-sector workers in Brazil.

3.2. **Sample.** We begin by defining a sample with a consistent composition of ENEM exam takers over time. The total number of ENEM exam takers increased significantly after the

[11] See, for example, *Lista do ENEM 2015: Notas das escolas*, *Globo*, October 5, 2016, available at https://especiais.g1.globo.com/educacao/enem/2015/enem-2015-medias-por-escola/.

[12] We linked the three administrative datasets at the individual level in a secure data room at INEP's facilities in Brasília and extracted results for our analysis. See Appendix C.2 for details on the merge.

exam was converted into a college admissions test in 2009, as illustrated by the black bars in Panel A of Figure 1. Since our goal is to examine how the increase in the exam's stakes impacted the distribution of scores, we define a sample in which the number of test takers remained relatively constant over these years. For this, we take advantage of the fact that many high school students took the ENEM in their senior year regardless of its stakes because of the exam's legacy as a high school accountability test.

Our analysis sample includes high school seniors at schools that met the criteria to be included in the government's accountability reports in each year from 2005–2015. To define our sample, we use a dataset that contains school-level mean ENEM scores from 2005–2015, which were computed by INEP and distributed to federal and municipal agencies for publication (INEP, 2019b). Schools were included in the accountability reports if a large fraction of their students participated in the exam. Our analysis sample includes only ENEM exam takers who: 1) are in their last year of high school; and 2) attended a school that appears in the INEP school-level dataset in *each* year from 2005 to 2015.[13] The red bars in Panel A of Figure 1 show that our analysis sample contains a small subset of all ENEM participants, but the number of exam takers in our sample remains relatively constant between 2007 and 2017.[14] Section 3.5 presents formal tests for balance in our sample.

Table 1 shows that our analysis sample is positively selected on socioeconomic status and academic performance relative to other ENEM test takers. This table reports mean demographic characteristics (Panel A), ENEM scores (Panel B), and college and labor market outcomes (Panel C) for 2009–2017 ENEM participants. Columns (A)–(C) show statistics for all ENEM exam takers, all high school seniors, and high school seniors in our analysis sample, respectively. Our sample contains roughly 2.5 million high school seniors, which is six percent of all ENEM test takers and 22 percent of all high school seniors. On average, students in our sample are four years younger than the typical ENEM participant, and they are roughly ten percentage points (pp) more likely to be white and to have a college-educated parent. Relative to the average test taker and the average high school senior, students in our sample score about 0.2–0.3 standard deviations (SD) higher on each ENEM subject.[15]

[13] At the schools in our sample, the mean ENEM participation rate from 2005–2015 was 70 percent. The criterion for inclusion in the reports changed over this time period, as we describe in Appendix C.3.

[14] Appendix Table A5 shows that our main results are robust to different sample selection criteria.

[15] ENEM scores, as reported to the public, are scaled to have a mean of 500 and a SD of 100 in the population of 2009 high school seniors who took the exam. Throughout the paper, we report ENEM scores in SD units relative to this population. For ENEM subject scores, our transformation is: Transformed score = (Scale score − 500)/100. Our transformation is different for writing and overall scores since they are on different scales. In all cases, a score of zero in our paper is equivalent to the performance of the average high school senior who took the ENEM in 2009, and a score of one is 1 SD higher within this population. These transformations preserve the comparability of test scores across cohorts. See Appendix C.1 for more details.

Despite this positive selection, there is substantial inequality between private and public high school students in our sample. Columns (D)–(F) of Table 1 report statistics for private school students, public school students, and the private/public gap. 32 percent of students in our sample attended a private high school. Relative to public school students, private school students were 26pp more likely to be white, 44pp more likely to have a college-educated mother, and 52pp more likely to come from a high-income family. Mean ENEM score gaps are on the order of 1 SD; the test score gap is largest in math, with private school students scoring 1.4 SDs higher than public school students on average.[16] There is also substantial inequality in college and labor market outcomes. Private school students were 27pp more likely to go to college, 15pp more likely to attend a federal university, and had hourly wages 68 percent higher than public school students.

3.3. **ENEM exam stakes.** Our identification strategy exploits the gradual adoption of the ENEM exam by federal universities. The solid red line in Panel B of Figure 1 plots the proportion of all federal university enrollees in each year who were admitted using the ENEM exam. Although the college admissions version of the ENEM exam was first administered in November 2009, only 28 percent of federal university students nationwide were admitted using the ENEM in the following year.[17] The proportion of federal university seats that were allocated using the ENEM grew over subsequent years as more institutions switched from their own tests to the ENEM, reaching a peak of 72 percent in 2016.

This gradual adoption created geographic variation in the exam stakes because Brazilian students typically attend college in their home state. The black dashed line in Panel B of Figure 1 plots the proportion of federal university enrollees who attended college in the state where they were born. On average, 81 percent of federal university students are from in-state. Although there is evidence that the ENEM exam increased geographic mobility (Machado and Szerman, 2021), these effects were modest; the proportion of in-state students at federal universities remained above 80 percent throughout 2010–2018. Thus, the stakes of the ENEM exam varied across states and cohorts for students who wished to attend a federal university in their home state.

We use this variation to define two measures of ENEM stakes at the state $\times$ year level. Our benchmark measure, which we denote by $\text{ProportionENEM}_{st}$, is a continuous variable that equals the proportion of federal university enrollees in state s and year t who were admitted using the ENEM exam. This continuous treatment variable has two advantages: it is straightforward to define, and it is more-powered because it includes all variation in ENEM adoption timing. For example, $\text{ProportionENEM}_{st}$ incorporates variation in ENEM

[16] For reference, the white/Black gap in the 2017 U.S. SAT math exam was 0.85 SDs (College Board, 2017).

[17] Since the ENEM is administered in November, scores are used for admission to university cohorts that begin in the following calendar year.

adoption across federal universities within the same state, as well as variation in the use of the ENEM across programs within the same university.

Second, we define a binary treatment variable that equals one in years after each state "adopted" the ENEM exam. For this, we follow research on tipping points (e.g., Card et al., 2008) in identifying structural breaks in the time series of federal universities' use of the ENEM. For each state s, we regress an annual time series of the proportion of federal university enrollees who were admitted using the ENEM on a linear trend break function for each possible candidate adoption year τ_s. We define the state's *ENEM adoption year* as the value τ_s^* that yields the highest R^2 across these regressions.[18] Our binary measure, which we denote by HighStakes_{st}, is an indicator for years equal to or after the state's ENEM adoption year, τ_s^*.[19] Our binary treatment variable allows us to present our results using event study graphs, and it helps to address potential concerns about two-way fixed effects models with treatment effect heterogeneity (discussed below).

Figure 2 shows the relationship between our continuous and binary measures of ENEM stakes. In Panel A, we categorize Brazil's 27 states into ten groups based on their year of ENEM adoption, τ_s^*. The graph plots the mean of $\text{ProportionENEM}_{st}$ in these groups (y-axis) for each ENEM exam year (x-axis). In each group, the proportion of federal university students who were admitted through the ENEM increases sharply in the state's ENEM adoption year. Panel B presents an event-study version of Panel A, in which the x-axis denotes years relative to the state's ENEM adoption year. On average, the share of a state's federal university admissions that used the ENEM exam increased by 56 percent in the adoption year, and this share remains at a high level in subsequent years. Appendix Table A2 shows the values of $\text{ProportionENEM}_{st}$ and HighStakes_{st} in each state and exam year.

3.4. **Regression models.** Our benchmark regression model is a two-way fixed effects specification estimated at the high school × year level:

$$Y_{ht} = \gamma_{s(h)} + \gamma_t + \beta \text{ProportionENEM}_{s(h)t} + \epsilon_{ht}. \tag{1}$$

[18] Specifically, we estimate the following regression for each state s:

$$\text{ProportionENEM}_{st} = \delta_s^0 + \delta_s^1 \mathbb{1}\{t \geq \tau_s\} + \delta_s^2 \mathbb{1}\{t \geq \tau_s\}(t - \tau_s) + \delta_s^3 \mathbb{1}\{t < \tau_s\}(t - \tau_s) + \epsilon_{st},$$

where $\text{ProportionENEM}_{st}$ is our continuous treatment variable. We estimate this regression for all candidate adoption years $\tau_s \in \{2008, ..., 2016\}$ and pick the value τ_s^* that yields the highest R^2 value. Lastly, we define our binary treatment variable to be $\text{HighStakes}_{st} = \mathbb{1}\{t \geq \tau_s^*\}$. We define one state (Sergipe) as a "never adopter" since the value of $\text{ProportionENEM}_{st}$ never exceeds 0.06.

[19] Throughout the paper, we refer to cohorts prior to each state's ENEM adoption year as "low stakes" cohorts for brevity, although the ENEM was consequential for students in these cohorts who wished to attend a federal university in other states that had already adopted the ENEM. The ENEM was also used to determine ProUni scholarships and eligibility for the federal FIES financial aid system (OECD, 2021); these incentives mattered mainly for students who wished to attend private universities, as public universities in Brazil are tuition-free. Yet the adoption of the ENEM by federal universities significantly increased the stakes of the exam for students who wished to attend a selective federal university close to home.

Y_{ht} is an average outcome for students who attended high school h and took the ENEM exam in year t. We include fixed effects for years, γ_t, and for the states in which each high school is located, $\gamma_{s(h)}$. The variable of interest is our continuous treatment variable, $\text{ProportionENEM}_{s(h)t}$, which measures the stakes of the ENEM exam in state $s(h)$ and cohort t. In alternate specifications, we replace $\text{ProportionENEM}_{s(h)t}$ with our binary treatment variable, $\text{HighStakes}_{s(h)t}$. We weight our regressions by the number of individuals in each ht cell to recover population estimates within our sample. Our benchmark regressions include high school seniors who took the college admissions version of the ENEM exam in 2009–2017, which holds the structure of the ENEM exam fixed over time.[20] We cluster standard errors at the state level.

The coefficient of interest, β, measures how outcomes changed in a school when the stakes of the ENEM exam increased. We estimate equation (1) separately for public and private high school students to examine how the increase in exam stakes affected scores in these two populations. In addition, we estimate regressions that fully interact the covariates in equation (1) with an indicator for private high schools, Private_h:

$$Y_{ht} = \gamma_{s(h)} + \gamma_t + \beta \text{ProportionENEM}_{s(h)t} + \left[\tilde{\gamma}_{s(h)} + \tilde{\gamma}_t + \beta^{\text{gap}} \text{ProportionENEM}_{s(h)t}\right] \text{Private}_h + \nu_{ht}. \tag{2}$$

The β^{gap} coefficient in equation (2) shows how the increase in exam stakes impacted the private/public gap in ENEM scores.

To examine the robustness of our results to concerns about two-way fixed effects models with treatment effect heterogeneity (De Chaisemartin and d'Haultfoeuille, 2020), we use a specification that restricts identification to clean comparisons based on states' ENEM adoption years. Our approach follows Callaway and Sant'Anna (2021) in estimating separate treatment effects for each pair of ENEM adoption years, τ_s^* and $\tau_{s'}^*$, and then averaging the pairwise treatment effects to recover a single point estimate. For example, one of our pairs contains states that adopted the ENEM in 2010 and 2011, and we restrict the sample to students who took the exam in 2009–2010. In this pair, the 2010 adopters are our treated group since ENEM adoption "switches on" in 2010. The 2011 adopters are our control group since these states had not yet adopted the ENEM in these years. We define groups for all pairwise combinations of ENEM adoption years, and in each pair, we restrict the sample to students who took the ENEM before the control group's adoption year. We create a stacked dataset of these pairwise samples and estimate a version of equation (2) that uses our binary treatment variable, HighStakes_{st}, and includes interactions with dummies for the pairwise

[20] In robustness analyses, we add in students from the same high schools who took the old 63-question version of the ENEM in 2007–2008.

groups.[21] The resulting β^{gap} coefficients are regression-weighted averages of the pairwise treatment effects. Appendix Table A4 shows the pairwise groups and the structure of our stacked dataset.

3.5. **Identification assumptions and balance tests.** Our identification relies on an assumption of parallel trends across Brazilian states. This assumption requires that the timing of federal universities' switch to the ENEM exam is unrelated to state-level trends in potential test score outcomes. There are two main ways in which this assumption could be violated. First, the increase in ENEM stakes may have affected the characteristics of students who took the exam within our high school senior sample. Second, the timing at which federal universities adopted the ENEM could be related to trends in student achievement.

Table 2 presents balance tests for the composition of our analysis sample. The dependent variables are the number of exam takers per high school (Panel A), the demographic characteristics of exam takers (Panel B), and an individual's *predicted* overall score based on demographic characteristics (Panel C). Column (A) shows the mean of each dependent variable in cohorts prior to the state's ENEM adoption year. Columns (B)–(D) present the β coefficients from equation (1), which we estimate separately for all schools, private schools, and public schools. Column (E) reports β^{gap} coefficients from equation (2), which are equivalent to the difference between the β coefficients in columns (C) and (D).

Our balance tests suggest that the stakes of the ENEM exam are not related to the composition of our sample. In Panel A of Table 2, we do not find significant effects on the number of ENEM takers in our full sample or in the private school subsample. We find that a 100 percentage point increase in the proportion of federal university students who were admitted using the ENEM exam is associated with a roughly 10 percent increase in the number of public school exam takers in our sample (column D); this effect is marginally significant in levels but not in logs. In Panel B, we find no systematic relationship between ENEM stakes and the age, race, parental education, or family income of exam takers in our sample. The increase in exam stakes is associated with a statistically significant but small decrease in the proportion of female exam takers. We cannot reject the hypothesis that the coefficients on all demographic characteristics are jointly equal to zero in any subsample (last row of Panel B). Similarly, we find small and insignificant effects on the demographic-based index of predicted test scores (Panel C). Overall, these tests suggest that the composition of our high school senior sample did not change significantly when the ENEM stakes increased.

[21] Our stacked regression specification is:

$$Y_{htg} = \gamma_{s(h)g} + \gamma_{tg} + \beta \text{HighStakes}_{s(h)t} + \left[\tilde{\gamma}_{s(h)g} + \tilde{\gamma}_{tg} + \beta^{\text{gap}} \text{HighStakes}_{s(h)t}\right] \text{Private}_h + \epsilon_{htg}. \tag{3}$$

This specification differs from equation (2) in three ways: 1) the dataset is at the high school (h) $\times$ year (t) $\times$ pairwise group (g) level; 2) we include state $\times$ group dummies, $\gamma_{s(h)g}$ and $\tilde{\gamma}_{s(h)g}$, and year $\times$ group dummies, γ_{tg} and $\tilde{\gamma}_{tg}$; and 3) we replace $\text{ProportionENEM}_{s(h)t}$ with $\text{HighStakes}_{s(h)t}$.

Appendix Table A3 shows that there are no systematic differences between states with federal universities that were early- and late-adopters of the ENEM exam, which lends further support to the parallel trends assumption. For example, universities in the most populous state, São Paulo, adopted the ENEM immediately in 2009, while federal universities in the next two largest states, Minas Gerais and Rio de Janeiro, did not adopt it until 2013. There are no clear trends in the size, selectivity, or student body characteristics of federal universities in early- vs. late-adopting states. This evidence counters the hypothesis that ENEM adoption was correlated with concurrent trends in student achievement.[22]

4. Exam stakes and the distribution of scores

4.1. **Effects on test scores.** Table 3 presents our main results on how the stakes of the ENEM impacted mean test scores. Column (A) displays the mean private/public school gap in test scores in cohorts prior to the state's ENEM adoption year. Columns (B)–(D) present β coefficients from equation (1) estimated separately by high school type. Column (E) displays β^{gap} coefficients from equation (2). Our dependent variables are individuals' test scores in SD units averaged at the high school $\times$ cohort level. We examine scores on each of the four multiple-choice tests (math, language arts, natural science, social science), average scores across these four core subjects, and scores on the writing component.

We find that the increase in the stakes of the ENEM exam led to a widening of private/public test score gaps. Private school students' scores increased on the higher-stakes exam in each of the four core subjects (column C), with the largest effect in math (0.143 SDs). Public school students' scores did not change significantly on the core subjects (column D). Thus, test score gaps between private and public school students increased with the stakes of the exam (column E). Our point estimate implies that a 100 percentage point increase in the use of the ENEM by federal universities is associated with a 0.11 SD increase in the private/public test score gap on core subjects. This effect is nine percent of the mean test score gap in lower-stakes cohorts (column A). We also find that the increase in ENEM stakes widened the private/public gap in writing scores by 0.10 SDs.

Figure 3 shows that test score gaps typically widened in the first ENEM exam cohort after its adoption by federal universities. This figure presents estimates from an event study version of equation (2) using our binary treatment variable, HighStakes_{st}, and our stacked dataset of pairwise ENEM adoption years. This yields coefficients β_l^{gap} that show how the private/public score gap changed in each year l relative to the state's ENEM adoption year,

[22] It is likely that the first universities to adopt the ENEM were those that saw the greatest benefits to doing so, e.g., those that found it especially costly to administer their own tests. Still, it is ambiguous how variation in the ENEM's benefits across universities relates to trends in the achievement of local students.

τ_s^*.[23] In most subjects, we do not see significant pre-trends in the private/public score gap prior to the ENEM adoption year. In all subjects, we find increases in the private/public score gap in the first cohort after ENEM adoption on the order of 0.05 to 0.10 SDs. These wider gaps decline only slightly in subsequent cohorts; for example, the average gap between private and public students on core subjects increased by 0.08 SDs in the year of ENEM adoption, and it was still 0.05 SDs higher measured four years later (Panel E).

The magnitudes of these estimates represent meaningful increases in private students' chances of gaining admission to federal university programs. The effect of higher exam stakes on the private/public gap in average ENEM scores ($\beta^{\text{gap}} = 0.11$) is 21 percent of a standard deviation in the distribution of cutoff scores for admission to federal university programs.[24] To put this in perspective, consider a private school student whose low-stakes ENEM score would have made them barely eligible for admission to a program at the 50th percentile of their state's distribution of federal university programs. Our estimate of β^{gap} implies that this student's high-stakes ENEM score would instead make them eligible for a program at the 58th percentile.

Panel A of Figure 4 shows that test score gaps by race, family income, and mother's education also expanded on the higher stakes ENEM exam (see also Appendix Table A6). The white bars represent mean gaps in average ENEM scores in low-stakes cohorts for different demographic groups. The grey bars represent estimates of β^{gap} from a version of equation (2) that replaces Private_h with a dummy for the more advantaged group.[25] We find that the gap in average scores between white/non-white students expanded by 0.06 SDs on the higher-stakes test. Similarly, the average score gap between students with college/non-college-educated mothers expanded by 0.08 SDs, and the gap by family income expanded by 0.09 SDs. These point estimates are smaller than our estimate for the private/public high school gap, but they are similar as a percentage of the mean gap in low-stakes cohorts. We do not find a significant effect on the male/female test score gap.

[23] Figure 3 plots β_l^{gap} coefficients from the high school (h) $\times$ year (t) $\times$ pairwise group (g) level regression

$$(4) \quad Y_{htg} = \gamma_{s(h)g} + \tilde{\gamma}_{s(h)g}\text{Private}_h + \gamma_{tg} + \tilde{\gamma}_{tg}\text{Private}_h + \sum_{l=-7}^{7}\left[\beta_l + \beta_l^{\text{gap}}\text{Private}_h\right]\mathbb{1}\{t - \tau^*_{s(h)} = l\} + \epsilon_{htg},$$

where l denotes years relative to the state's ENEM adoption year, $\tau^*_{s(h)}$. We include state $\times$ group dummies, $\gamma_{s(h)g}$, year $\times$ group dummies, γ_{tg}, and dummies for years l, $\mathbb{1}\{t - \tau^*_{s(h)} = l\}$, omitting $l = -1$. We interact all covariates with a dummy for private schools, Private_h, and plot the β_l^{gap} coefficients from $l = -4$ to 4.

[24] In 2016 data from the SISU admission system, the within-state standard deviation of cutoff scores for federal university programs is 0.52 SDs (in the test score units of our sample). Thus $0.11/0.52 \approx 21$ percent.

[25] For Figure 4, we estimate equation (2) at the individual level rather than the high-school $\times$ year level.

4.2. **Robustness tests.** Table 4 examines the robustness of our results on private/public test score gaps. Column (A) reproduces our benchmark estimates of β^{gap} from column (E) of Table 3. Columns (B)–(F) present estimates of β^{gap} from alternative specifications.

Our results are robust to including demographic controls and using a binary measure of exam stakes. In column (B) of Table 4, we estimate equation (2) including high school $\times$ year averages of age, gender, and dummies for race, parental education, and family income bins. These demographic controls do not significantly alter our point estimates, which is consistent with the findings of our balance tests in Table 2. In column (C), we replace our continuous treatment variable, $\text{ProportionENEM}_{st}$, with our binary measure of ENEM stakes, HighStakes_{st}. This specification reduces the magnitudes of β^{gap} by about 50 percent in each subject, which is expected since $\text{ProportionENEM}_{st}$ increases by roughly 50 percent following a state's adoption of the ENEM (Figure 2, Panel B). Yet we continue to find that the increase in exam stakes widened private/public test score gaps in each subject, and the coefficient for the average score remains statistically significant at $p < 0.05$.

Next, we examine the robustness of our benchmark estimates to potential concerns about two-way fixed effects models with treatment effect heterogeneity (De Chaisemartin and d'Haultfoeuille, 2020). For this, we use three different samples from our stacked dataset as described in Section 3.3. Column (D) of Table 4 includes all pairwise combinations of ENEM adoption years that we can estimate using 2009–2017 exam takers. Column (E) focuses on a single pairwise comparison between the two most common ENEM adoption years, 2009 and 2013, which together account for 13 states (see Appendix Table A3). We require a pre-period for 2009 adopters to estimate a treatment effect in this pair, so this sample includes 2007–2012 test takers. In column (F), we include all 2007–2017 test takers and all pairwise combinations in our stacked dataset. Note that in columns (E)–(F), the sample includes two cohorts that took the old 63-question version of the ENEM exam (2007–2008), so these estimates may reflect the effects of the ENEM redesign in addition to the impacts of the exam's adoption by federal universities.[26] Appendix Table A4 shows the samples for each regression in columns (D)–(F) of Table 4.

The results in all of these specifications are similar to our benchmark estimates. The point estimates in column (D) are similar to those in column (C), which shows that our results are not impacted by restricting identification to clean pairwise comparisons. We continue to find positive and significant estimates of β^{gap} when we restrict to the simple

[26] The 2007–2008 ENEM reported only a single core-component score plus a writing score. To define scores for each subject, we categorized the multiple choice questions into math, language arts, natural science, and social science, and then computed a separate score for each subject using the IRT parameters. Since the reference populations differ for the 2007–2008 and 2009–2017 exams, our regressions in columns (E)–(F) of Table 4 standardize scores to have mean 0 and SD 1 within each year of our sample. See Appendix C.1 for details.

"2×2" difference-in-differences model that compares 2009 vs. 2013 adopters (column E). Lastly, our results are similar in the full stacked dataset with 2007–2017 test takers (column F). The consistency of estimates across specifications shows that our results are not the result of averaging oppositely-signed treatment effects with negative weights.

Our results are also robust to controls for the nationwide rollout of affirmative action during our sample period. Many federal and state universities implemented reserved quotas for disadvantaged students during the late 2000s and early 2010s (Mello, 2022), which could have impacted the achievement of high school seniors through a motivational channel (Akhtari et al., 2020). To examine this possibility, we use the higher education census to compute the fraction of new university students in each state × year who enrolled through reserved quotas and then add this variable as a control in our regressions. Appendix Table A8 shows that private/public test score gaps are not significantly related to the rollout of affirmative action and that our estimates for the impacts of ENEM stakes remain positive and significant with these controls.

4.3. **Mechanisms.** Our finding that private students earned higher scores on the high-stakes ENEM exam may be driven by several mechanisms. One possibility is that students may have exerted more effort while taking the exam. The typical private school student had a better chance of gaining admission to federal universities than the typical public student, and thus, private students had a stronger incentive to increase effort when the exam stakes increased. There is significant overlap between the distribution of private school ENEM scores and the distribution of admission cutoff scores for federal university programs, while the public school score distribution is shifted well to the left (Appendix Figure A1). Thus, moderate increases in ENEM scores were unlikely to significantly affect the admission chances of public students at most federal universities. Although we cannot observe student effort in our data, this may partly explain why we find significant increases in scores only for private students.[27]

The increase in private students' scores could also reflect additional test preparation prior to the exam. Both student-level and institutional-level responses could lead to additional test prep. Students, for instance, might have changed the amount of time spent studying for the ENEM or adjusted their study methods, for example, by increasing emphasis on mock tests or problem-solving strategies. Meanwhile, educational institutions like private high schools and test-prep companies may have re-oriented the materials taught. Several private high schools changed their curriculum to better align with the skills assessed by the ENEM. Similarly, test prep companies may have re-focused their sessions to cover solely the material

[27] Some individuals in our sample would have been eligible for affirmative action quotas that were reserved for public school students, but these quotas were not fully implemented at many federal universities until 2016, and they often included race and/or SES criteria in addition to a public school criterion.

of the ENEM rather than the content of multiple vestibular exams. Thus, the expansion of the private/public ENEM score gap may be driven by an increase in ENEM-specific test prep following the exam's adoption by federal universities.

We examine the role of test prep using two measures of students' engagement in test prep activities. First, we obtained lists of schools that use test-oriented curricula from the websites of four prominent test prep companies (*Sistema Anglo*, *Sistema pH*, *Elite Rede de Ensino*, and *Curso Objetivo*). We then merged these lists to our sample of high schools using geocoded addresses. This allows us to define a set of "prep schools" whose curriculum is specifically focused on preparation for college admission exams.[28] Second, we use a variable from the ENEM questionnaire that indicates whether individuals took an entrance exam preparation course. This question does not distinguish between courses that focused on the ENEM exam and courses that focused on other *vestibular* exams. Thus, we do not find that the proportion of students who took test prep courses increased when the ENEM was adopted by nearby federal universities. Yet, if these courses were more likely to focus on the ENEM exam after its adoption in admissions, this could impact the ENEM scores of students who took preparation courses. Appendix Table A7 provides details on these measures of test prep as well as the associated regression results.

Panel B of Figure 4 shows that the increase in ENEM stakes also expanded test score gaps between students who did/did not engage in test prep as defined by these measures. The gap in average ENEM scores between prep-focused private schools and public schools expanded by 0.18 SDs on the higher-stakes exam (second row in Panel B), which is more than 60 percent larger than our benchmark point estimate for the private/public gap (0.11 SDs, first row). The third row of Panel B shows that prep schools had slightly lower average scores than other private schools in our sample in the low-stakes cohorts, and the increase in exam stakes closed this gap by 0.08 SDs ($p < 0.05$). Lastly, we find that the increase in ENEM stakes led to a large increase in the average score gap between students who did/did not take a test prep course (fourth row), with a point estimate of 0.23 SDs. These heterogeneity results suggest that test prep is an important mechanism for the increase in private/public test score gaps.[29]

4.4. **Narrow vs. broad-based learning.** A common concern about high-stakes exams is that students who engage in test prep learn narrow skills that raise their exam scores but are not useful outside of the exam. As a first test of this concern, we ask whether the increase

[28] We focus on these four companies because they list the names and addresses of affiliate schools on their websites. There are other prominent test prep companies in Brazil for which we could not find lists of affiliate schools, so it is likely that many other private schools in our sample also have test-oriented curricula. Thus, the coefficients we estimate may be attenuated due to measurement error.

[29] Consistent with a role for test prep, Appendix Figure A2 shows that state-level Google searches for "ENEM" and the online prep company "*Descomplica*" increased when federal universities adopted the ENEM.

in private students' ENEM scores reflects narrowly-targeted or broad-based improvements in performance across different types of exam questions.[30]

Our data allows for this type of analysis because it includes students' responses to each exam question as well as information on the skills that the questions measure. The ENEM is designed to be closely related to high school curricula, and the questions are based on a "reference matrix" of skills that educators think are important for students to know by the end of high school.[31] For example, math questions are grouped into 7 *topic areas* that cover different branches of mathematics, such as algebra, geometry, and statistics. Math questions are further grouped into 30 *competencies* that measure specific abilities within each topic area, e.g., identifying concepts, solving problems, and constructing arguments. Questions from the other subject exams are categorized in a similar way.

We use our question-level data to estimate our regression model (2) separately for different topic areas and competencies with each exam subject. In these regressions, the dataset is at the high school (h) $\times$ year (t) $\times$ question (q) level, and the dependent variable is the proportion of correct answers in each htq cell. We focus on math performance in the main text because it is the subject with the largest increase in ENEM score gaps (Table 3) and because math exams are often thought to be more "preppable" (Riehl and Welch, 2023). Appendix Table A9 presents results for language arts, natural science, and social science.

Panel A of Table 5 shows that the increase in ENEM stakes expanded the private/public gap in the proportion of correct answers, consistent with our results on scale scores (Table 3). This panel shows results pooling across all 405 math questions (9 years $\times$ 45 questions/exam). Column (C) reports the mean proportion of correct answers for public school students in cohorts prior to the state's ENEM adoption year, and column (E) reports the mean private/public gap in these cohorts. The ENEM math exam is challenging for most students; the average public school student answered only 29.1 percent of the questions correctly. Private school students got 46.7 percent of the questions correct; thus, the private/public gap was 17.6pp. Columns (D) and (F) report the β and β^{gap} coefficients from equation (2). The β^{gap} coefficient suggests that a 100 percentage point increase in ENEM adoption by federal universities is associated with a 2.4pp increase in the private/public gap in correct responses. We do not find a significant change in the proportion of correct answers for public students.

Panels B–C of Table 5 show that we find increases in the private/public gap in correct responses across a wide range of math topic areas and competencies. These panels display

[30] Our heterogeneity analysis in this subsection is similar to Cohodes (2016)'s analysis of the impacts of charter school admission on different subscores of state standardized tests.

[31] See: https://download.inep.gov.br/download/enem/matriz_referencia.pdf (accessed in June 2023).

results from estimating equation (2) separately for each topic area and competency.[32] We find positive and statistically significant estimates of β^{gap} in all seven math topic areas, with estimates ranging from 1.3pp in algebra to 3.6pp in questions on proportions (Panel B). The estimates at the competency level are less powered since these regressions typically include only 10–15 questions across all years, but the β^{gap} coefficients are positive and greater than 0.8pp in 29 out of the 30 competencies (Panel C).

The results in Table 5 show that the increase in ENEM stakes induced a broad-based improvement in private students' performance on the math exam. In language arts, natural science, and social science, we also find increases in the private/public gap in correct responses across a range of topic areas (Appendix Table A9). The ENEM is specifically designed to test skills that educators think are important for high school graduates to know, and the exam features competencies such as problem-solving and interpreting that are intended to test reasoning more than memorization. Thus, the broad-based improvement in private students' performance runs counter to the concern that high-stakes exams induce students to focus on a narrow set of skills that merely raise their exam scores.

At the same time, there is significant variation in the β^{gap} coefficients in Table 5, which leaves open the possibility that students prepared for some types of exam questions more than others.[33] We reject equality of the β^{gap} coefficients for the math exam at both the topic area ($p = 0.015$) and competency ($p < 0.001$) levels, and the results are similar for other subjects (Appendix Table A9). Thus, it is possible that the exam skills that drove the increase in the private/public test score gap tended to also be skills that are less informative for future outcomes. Further, it is also possible that the broad-based improvement in students' performance reflects test-taking strategies, which could lead to better performance on many types of questions. It is difficult to draw strong conclusions about the value of test prep for high-stakes exams using data on exam performance alone. For this reason, we now turn to our analysis of the predictive power of the ENEM for college and labor market outcomes.

5. Exam stakes and the informativeness of scores

5.1. **Potential channels.** This section asks how the increase in the stakes of the ENEM impacted the informativeness of ENEM scores for students' college and labor market outcomes.

[32] For brevity, Panel C reports only competencies with the five largest and five smallest values of β^{gap}.

[33] In Appendix Table A10, we do not find systematic differences in the β^{gap} coefficients between math questions that are/are not related to topics in an ENEM study guide created by the test prep company *Me Salva!*. This provides suggestive evidence against the hypothesis that our findings are driven by certain types of questions that are more amenable to test prep.

Literature in economics, as well as popular debates about high-stakes testing, suggest that the answer to this question is ambiguous.[34]

On the one hand, high-stakes exams may reduce the informativeness of scores by distorting effort toward activities that raise exam performance rather that activities that promote beneficial learning. In their seminal paper on incentive contracts, Holmstrom and Milgrom (1991) highlight the possibility that teachers who are rewarded for student test performance may focus on "the narrowly defined basic skills that are tested on standardized exams." Frankel and Kartik (2019) show theoretically that higher stakes tests are relatively more informative about individuals' "gaming ability" in settings where signaling is important. Critics of standardized testing often argue that the ability to "game the system" may be unrelated—or even negatively related—to an individual's potential for academic success (e.g., Harris et al., 2011).

Yet it is also possible that high-stakes tests allow students to reveal or develop skills that are beneficial for college. As Frankel and Kartik (2019) note, an individual's gaming ability for high-stakes tests could reflect work ethic, the capacity to learn new material, or other attributes that colleges value. If high-stakes scores are more correlated with family income (as we found in Section 4), they may be more correlated with college success because family income also helps students succeed in college. In addition, it is also possible that the existence of high-stakes testing may compel students to accumulate new skills that benefit them beyond the exam. For example, students may learn useful academic material if the exams are well-aligned with high school and college curricula. Alternatively, high-stakes testing may promote the development of non-cognitive skills such as cognitive endurance (Brown et al., 2022; Reyes, 2023) or a growth mindset (Dweck, 2006).

The magnitude of many of these potential mechanisms is *ex ante* unclear, which motivates our analysis in this section.

5.2. **Outcome variables.** To examine the relationship between exam stakes and exam informativeness, we take advantage of our merged administrative data that connect ENEM participants to national higher education and labor market records (see Section 3). Our sample for this analysis includes the subset of students in our high school senior sample who took the ENEM exam in 2009–2014, excluding the 2011 cohort (for which there is a data issue).[35] Our linkage with Brazil's higher education census allows us to measure enrollment, persistence, and graduation outcomes at all Brazilian colleges in the years 2010–2019. Our

[34] Appendix B presents a theoretical framework that illustrates the potential channels through which exam stakes can impact test score gaps and exam informativeness. Here we just briefly describe the intuition.

[35] We exclude 2011 ENEM takers from our analysis of exam informativeness because the crosswalk variable that INEP created to match individuals across their different datasets is not correctly defined for this cohort. We also exclude 2015–2017 ENEM takers from this analysis because we cannot define many of our longer-run outcomes variables in these cohorts given the timing of our data.

linkage with Brazil's national employer-employee data (the RAIS) provides us with earnings outcomes for the years 2016–2018.

Using this data, we define outcomes variables that measure individuals' academic and labor market success. We categorize these outcomes into three groups based on the sample for which they are defined. First, we use outcomes that are defined for all individuals in our sample, which include an indicator for enrolling in any college, an indicator for completing a college degree during our data period, and an indicator for appearing in the RAIS dataset (a measure of formal employment). Second, we measure college persistence and graduation outcomes in the subsample of individuals who enrolled in college. These outcomes include indicators for persisting in college x years after enrolling and an indicator for completing the program within five years. Lastly, our main labor market outcome is an individual's mean hourly wage over the years 2016–2018 (measured in both logs and levels), which we observe only for individuals who appear in the RAIS dataset. Many of the 2009–2014 ENEM participants in our sample were still in college during the period of our RAIS data, and even those who had left college were still early in their careers. Thus an important caveat is that our earnings outcomes may not capture the long-run returns to their college investments.[36]

5.3. **Informativeness of scale scores.** We begin by examining how the increase in ENEM stakes impacted the informativeness of scale scores for student's college and labor market outcomes. Our main measure of exam informativeness is the correlation coefficient between ENEM scores and future outcomes, which follows testing agencies' standard practice in measuring the "validity" of their exams (e.g., Kobrin et al., 2008). In addition to the potential mechanisms discussed above, a student's ENEM score may be correlated with their longer-run outcomes through its direct impact on which college and/or major they gained admission to and attended. Further, federal universities that adopted the ENEM exam for admissions simultaneously joined the SISU centralized college admission system, which also impacted the matching of students to college programs (see Section 2.2). To reduce the influence of these direct impacts on student/college matches and isolate the predictive power of scores, we also estimate correlation coefficients using residuals from regressing both ENEM scores and outcomes on dummies for college × major pairs, which is also standard in studies of exam validity. The correlations between these residuals reflect only variation in the informativeness of scores among students who attended the same college programs.

Table 6 presents our main results on how the informativeness of ENEM scores for student outcomes changed when federal universities adopted the ENEM in admissions. For this table we compute correlation coefficients between students' average ENEM scores (across the four core subjects) and their longer-run outcomes separately for each state × year pair

[36] See Appendix C.1 for details on the definitions of our college and labor market outcome variables.

(st). We then use these correlation coefficients as dependent variables, Y_{st}, in a state × year version of our regression model (1).[37] Column (A) shows the mean correlation coefficient in cohorts prior to federal universities' adoption of the ENEM (HighStakes$_{st}$=0). Columns (B)–(C) presents our coefficients of interest, β, using our continuous treatment variable, ProportionENEM$_{st}$. Columns (D)–(E) present β coefficients using our binary treatment variable, HighStakes$_{st}$. We present results in which the dependent variables are both raw correlation coefficients (columns B and D) and correlation coefficients after residualizing ENEM scores and outcomes on college × major pairs (columns C and E).

We find that scores on the higher-stakes ENEM exam were *more* informative for students' college enrollment and degree attainment outcomes. Panel A of Table 6 shows that the correlation coefficients between average ENEM scores and both college enrollment and college degree attainment increased when federal universities adopted the ENEM in admissions. The point estimates in column (B) imply that a 100 percentage point increase in the adoption of the ENEM is associated with a 0.036 increase in the correlation between ENEM scores and an indicator for college enrollment, and a 0.033 increase in the correlation between ENEM scores and an indicator for completing a college degree by 2019. We find similar (albeit slightly less-powered) results using our binary treatment variable (column D). Our results are also similar we when compare degree attainment outcomes for students who attended the same college programs (columns C and E), which suggests that our findings are not driven by direct impacts of ENEM performance on the schools or majors that students attended.[38]

The increase in the informativeness of ENEM scores was even more pronounced for college persistence outcomes measured within the population of college enrollees. In Panel B of Table 6, our outcome variables include indicators for persisting in college one and three years after enrolling as well as an indicator for completing the program within five years. The mean correlations between these outcomes and ENEM scores tend to be lower than for the outcomes in Panel A, but our estimates of the β coefficients in columns (B)–(E) are broad similar in magnitude in the two panels. Thus as a percentage of the mean correlation coefficients in lower-stakes cohorts, the impact of the higher-stakes exam on the informativeness of scores was larger for college persistence outcomes. For example, the estimates in column (B) imply that the full adoption of the ENEM exam by federal universities is associated with a 23

[37] In other words, our regression model for Table 6 is:

$$Y_{st} = \gamma_s + \gamma_t + \beta \text{ProportionENEM}_{st} + \epsilon_{st}, \tag{5}$$

where Y_{st} is the correlation coefficient between ENEM scores and longer-run outcomes for students who attended high school in state s and took the ENEM in year t. All covariates are defined as in equation (1).

[38] We do not find a significant relationship between exam stakes and formal employment as measured by appearing in the RAIS dataset (last row of Panel A). ENEM performance is negatively correlated with formal employment in the lower-stakes cohorts (column A), which likely reflects the fact that many students who attended college had not yet entered the labor market during our data period.

percent increase in the correlation between ENEM scores and 3-year persistence rates, and a 62 percent increase in the correlation between scores and 5-year graduation rates.

We find less systematic evidence on the relationship between exam stakes and the informativeness of scores for hourly wages. Panel C of Table 6 show that scores on the higher-stakes exam also became more correlated with hourly wages measured in levels, but these results are not robust to using log wages. These inconclusive results likely reflect the fact that we observe wages for only one-quarter of students in our sample because many were still enrolled in college during our data period. Our prior is that the higher-stakes ENEM scores would likely be more informative for long-run wages since we found that they were more informative for educational attainment, but we cannot directly test this given our data constraints.

Figure 5 shows that the increase of informativeness was relatively uniform across four subjects. The translucent areas represent the mean correlation coefficients in lower-stakes cohorts (analogous to column A of Table 6), and the darker areas depict the β coefficients from our benchmark two-way fixed effect regressions (analogous to column B of Table 6). In each of the core subjects—math, language arts, natural science, and social science—we find that the increase in exam stakes caused ENEM scores to become more correlated with degree completion, three-year college persistence, and hourly wages (in levels). The point estimates are slightly larger in math than in other subjects, but we find mostly significant estimates at $p < 0.05$ in each of the four core subjects. We also find increases in the informativeness of writing scores, but these are smaller in magnitude and mostly insignificant.

5.4. **"Preppability" vs. informativeness of exam skills.** As noted above, critics often argue that high stakes exams distort effort toward skills that are not useful outside of the exam. To provide evidence on this concern, we next ask whether the exam skills that tend to be more "preppable" also tend to be less informative for student outcomes.

We define the "preppability" of exam skills using the β_{gap} coefficients from estimating equation (2) at the topic area or competency level, as in column (F) of Table 5.[39] These coefficients show how the private/public gap in the proportion of correct answers changed in each exam skill when federal universities adopted the ENEM in admissions. For brevity, we refer to these β_{gap} coefficients as the exam skill's preppability. But, more precisely, what we mean by preppability is the degree to which questions in a certain skill group contributed to the expansion of private/public gaps in test scores following the increase in ENEM stakes.

Our measure of the informativeness of each exam skill is the average *return to a correct answer* on questions in each topic area or competency. We define the return to a correct answer as the difference in the average outcomes of students who got the question correct and incorrect. This is a measure each question's informativeness because it indicates how well

[39] See Section 4.4 for details on the ENEM's categorization of exam skills into topic areas and competencies.

the question can discriminate between individuals who are more and less likely to succeed as defined by a future outcome. We average these returns at the topic area or competency level to measure the informativeness of different exam skills.

Figure 6 provides an illustration of our analysis of the informativeness and preppability of exam skills. The y-axes depict the return to a correct answer as measured by log wages, which is the difference in log wages between individuals who got the question correct and incorrect.[40] Panel A shows average returns to a correct answer at the topic area level, and Panel B shows averages at the competency level. The x-axes show our measure of preppability, which is the β_{gap} coefficients from equation (2) estimate at the topic area (Panel A) and competency (Panel B) levels. Figure 6 includes exam skills from all four core subjects, as illustrated by the marker colors and symbols. The dashed line shows the linear relationship between exam skill informativeness and preppability.

In Figure 6, we find that exam skills that are more preppable tend to be *more* informative for log wages. Correct answers on the ENEM are highly informative for individuals' future wages; the y-axes of Figure 6 show that the mean returns to a correct answer are positive for every exam skill, with an average value of 0.15 log points. Notably, exam skills that contributed more to the expansion of the private/public test score gap, as indicated by larger β_{gap} coefficients, also tend to have larger returns to a correct answer. At the topic area level, for example, math questions on proportions have the largest β_{gap} coefficient (3.6pp), and they also have one of the largest mean returns to a correct answer (0.19 log points). Conversely, foreign language questions on the language arts exam have the lowest value of β_{gap} (-0.5pp), and the mean difference in wages for individuals with correct/incorrect responses in this topic area is only 0.13 log points.

Table 7 generalizes the results in Figure 6 to other college and labor/market outcomes. Column (A) shows the mean return to a correct answer measured across all 120 exam competencies (4 subjects $\times$ 30 competencies per subject) using the same outcome variables as in Table 6. Columns (B)–(F) report the OLS coefficients from a bivariate regression of these mean returns to a correct answer on the competency-level β_{gap} coefficients. These OLS coefficients are analogous to the slope of the dashed line in Panel B of Figure 6, except we normalize the coefficients to represent a one percentage point increase in β_{gap}. Columns (B)–(E) show results for each of four core subjects separately, and column (F) shows results pooling across all subjects.

The results in Table 7 show that more preppable exam competencies tend to be more informative for most college and labor market outcomes. Column (A) shows that the average

[40] When we examine labor market outcomes in Figure 6 and Table 7, we restrict the sample to 2009–2012 ENEM participants so that our estimates are more likely to reflect individuals' post-college labor market returns.

return to a correct answer is positive for college enrollment and graduation (Panel A), college persistence (Panel B), and wages (Panel C). In columns (B)–(F), we find that exam competencies with larger β_{gap} coefficients also tend to be more informative for most outcomes. Measured using all subjects and competencies, a one percentage point increases in β_{gap} is associated with a 0.3pp increase in the likelihood of earning a college degree by 2019, and a 0.2pp increase in the probability of persisting in college for 3 years conditional on enrolling, and a 0.8 percent increase in hourly wages (column F). The positive relationship between preppability and informativeness also arises within competencies on the math and language arts exams, but we find no significant relationship in natural science or social science.[41]

5.5. **Discussion.** The above results show that high stakes tests can be more informative about students' potential to succeed than low stakes exams. The adoption of the ENEM exam by federal universities led to broad-based improvements in private students' ENEM performance (5), and it increased the predictive power of students' scale scores for their college outcomes (Tables 6). Further, we found that the exam skills that contributed most to the increase in students' performance also tend to be more informative for future success (7).

Our results do not necessarily imply that the skills that students learned while preparing for the ENEM benefited them in the future, although this is one possible mechanism for our results. One of the motivations for the redesign of the ENEM exam was to create a nationwide admission test that was better aligned with high school and college curriculum than many of the university-specific *vestibular* exams. Thus the adoption of the ENEM by federal universities may have redirected students' preparation efforts toward material that benefited them in college. Yet it is also possible that the higher stakes exam was simply more correlated with other individual characteristics that help students succeed in college. These characteristics could include socioeconomic factors as well as aptitudes such grit or the capacity to learn new material.

Regardless of the mechanism, our findings can explain why many colleges around the world use high stakes tests for admissions. Colleges care about their reputations (MacLeod et al., 2017), and thus they want to admit students who will graduate from their programs and go on to high paying jobs. It is often argued that test prep for high stakes exams creates "bias" in scores, and thus that low stakes exams provide a better measure of students' true potential (**ER:Add cites to low-stakes literature from the introduction**. Our findings

41 Appendix Table A11 shows that the positive correlations between the informativeness and preppability of ENEM skills persist when we add controls for question difficulty and other IRT parameters.

show that, by contrast, higher stakes exams make it easier for colleges to identify students who are likely to be successful in their programs.[42]

6. Conclusion

This paper exploited a natural experiment in Brazil to examine how the use of high-stakes standardized admission exams affects inequality in exam scores and their information content. From 2009–2017, Brazil's system of highly-selective federal universities transitioned from institution-specific admission exams to a national standardized test called the ENEM. Since the ENEM exam was also used for high school accountability, many students took the exam regardless of its role in college admissions. This setting allowed us to focus on a comparable population of students and ask how the distribution and informativeness of exam scores changed as the ENEM's role in college admission grew.

We found that socioeconomic test score gaps increased by roughly 10 percent when federal universities began to use the ENEM for college admissions. High-income students earned better scores on the higher-stakes exam, suggesting that they exerted more effort or engaged in test prep to boost their performance. This shows that high-stakes exams can give wealthy students a leg up in college admissions, consistent with a common criticism of these exams.

Yet we also found that the adoption of the ENEM exam by federal universities made the exam scores *more* informative for students' college outcomes. Consistent with this increase in informativeness, we found that the performance of high-income students improved across a wide range of subjects and question types, suggesting that any test prep was not narrowly-targeted. This shows that high-stakes exams provide information on individual characteristics that help students succeed academically, such as the capacity to learn new material.

Increasingly, U.S. colleges are reducing their reliance on high-stakes exams in favor of lower-stakes admission signals like high school grades. Our findings show that this change will help them to diversify their student bodies. But if these schools wish to maintain their average graduation rates, our results show that they must also find other ways of identifying students who are likely to succeed.

[42] Our findings also shed light on why many U.S. colleges use SAT or ACT "superscores"—which take the maximum of each subject score across all of the student's test attempts—rather than average scores in admissions. High-ability and advantaged students are more likely to retake these exams, and thus more likely to benefit from the practice of superscoring (Goodman et al., 2020). Thus superscores, relative to average scores, may be more correlated with characteristics that predict college success, such as socioeconomic background or the willingness to exert effort.

References

Akhtari, M., N. Bau, and J.-W. P. Laliberté (2020). Affirmative action and pre-college human capital. NBER Working Paper No. 27779.

Amrein, A. L. and D. C. Berliner (2002). High-stakes testing & student learning. *Education Policy Analysis archives 10*, 18–18.

Azmat, G., C. Calsamiglia, and N. Iriberri (2016). Gender differences in response to big stakes. *Journal of the European Economic Association 14*(6), 1372–1400.

Bénabou, R. and J. Tirole (2003). Intrinsic and extrinsic motivation. *The review of economic studies 70*(3), 489–520.

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

Bhattacharya, D., S. Kanaya, and M. Stevens (2017). Are university admissions academically fair? *Review of Economics and Statistics 99*(3), 449–464.

Brown, C. L., S. Kaur, G. Kingdon, and H. Schofield (2022). Cognitive endurance as human capital. NBER Working Paper No. 30133.

Buchmann, C., D. J. Condron, and V. J. Roscigno (2010). Shadow education, american style: Test preparation, the sat and college enrollment. *Social forces 89*(2), 435–461.

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

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

Callaway, B. and P. H. Sant'Anna (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics 225*(2), 200–230.

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and program planning 2*(1), 67–90.

Card, D., A. Mas, and J. Rothstein (2008). Tipping and the dynamics of segregation. *The Quarterly Journal of Economics 123*(1), 177–218.

Cawley, J., J. Heckman, and E. Vytlacil (1999). On policies to reward the value added by educators. *Review of Economics and Statistics 81*(4), 720–727.

Cohodes, S. R. (2016). Teaching to the student: Charter school effectiveness in spite of perverse incentives. *Education Finance and Policy 11*(1), 1–42.

College Board (2017). SAT Suite of Assessments Annual Report. https://reports.collegeboard.org/media/pdf/2017-total-group-sat-suite-assessments-annual-report.pdf (accessed July 2023).

Corcoran, S. P., J. L. Jennings, and A. A. Beveridge (2011). Teacher effectiveness on high-and low-stakes tests. Society for Research on Educational Effectiveness.

Cullen, J. B. and R. Reback (2006). Tinkering toward accolades: School gaming under a performance accountability system. In *Improving school accountability*, Volume 14, pp. 1–34. Emerald Group Publishing Limited.

De Chaisemartin, C. and X. d'Haultfoeuille (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review 110*(9), 2964–96.

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

Sciences 108(19), 7716–7720.

Dweck, C. S. (2006). *Mindset: The new psychology of success.* Random house.

Fernandes, S. (2015). Neoliberalization of education in Brazil: the impact of cursinhos and the private sector on pedagogical practices and access to university. *Canadian Journal of Latin American and Caribbean Studies/Revue canadienne des études latino-américaines et caraïbes 40*(3), 351–367.

Figlio, D. N. and L. S. Getzler (2006). Accountability, ability and disability: Gaming the system? In *Improving school accountability*, Volume 14, pp. 35–49. Emerald Group Publishing Limited.

Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Research Report Series 2015*(2), 1–17.

Fischer, P. E. and R. E. Verrecchia (2000). Reporting bias. *The Accounting Review 75*(2), 229–245.

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

Frankel, A. and N. Kartik (2022). Improving information from manipulable data. *Journal of the European Economic Association 20*(1), 79–115.

Gneezy, U., J. A. List, J. A. Livingston, X. Qin, S. Sadoff, and Y. Xu (2019). Measuring success in education: The role of effort on the test itself. *American Economic Review: Insights 1*(3), 291–308.

Goodman, J., O. Gurantz, and J. Smith (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.

Harris, P., B. M. Smith, J. Harris, et al. (2011). *The myths of standardized tests: Why they don't tell you what you think they do.* Rowman & Littlefield Publishers.

Holmstrom, B. and P. Milgrom (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics and Organization 7*(1), 24–52.

Hoxby, C. and S. Turner (2013). Expanding college opportunities for high-achieving, low income students. Stanford Institute for Economic Policy Research Discussion Paper.

INEP (2019a). Microdata for ENEM exam (Exame Nacional do Ensino Médio). Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, Brasília, Brazil. https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem (accessed September 2019).

INEP (2019b). School-level ENEM scores (Exame Nacional do Ensino Médio). Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, Brasília, Brazil. https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem-por-escola (accessed September 2019).

INEP (2022). Microdata for Higher Education Census (Censo da Educação Superior). Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, Brasília, Brazil. https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior (accessed April 2022).

Jacob, B. and J. Rothstein (2016). The measurement of student ability in modern assessment systems. *Journal of Economic Perspectives 30*(3), 85–108.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics 89*(5-6), 761–796.

Jacob, B. A. and S. Levitt (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics 118*(3), 843–877.

Jalava, N., J. S. Joensen, and E. Pellas (2015). Grades and rank: Impacts of non-financial incentives on test performance. *Journal of Economic Behavior & Organization 115*, 161–196.

Jones, M. G., B. D. Jones, B. Hardin, L. Chapman, T. Yarbrough, and M. Davis (1999). The impact of high-stakes testing on teachers and students in north carolina. *The Phi Delta Kappan 81*(3), 199–203.

Klein, S. P., L. Hamilton, D. F. McCaffrey, B. Stecher, et al. (2000). What do test scores in texas tell us? *Education policy analysis archives 8*, 49–49.

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

Koretz, D. M. and S. I. Barron (1998). *The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS).* ERIC.

Kreps, D. M. (1997). Intrinsic motivation and extrinsic incentives. *The American economic review 87*(2), 359–364.

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

MacLeod, W. B., E. Riehl, J. E. Saavedra, and M. Urquiola (2017). The big sort: College reputation and labor market outcomes. *American Economic Journal: Applied Economics 9*(3), 223–261.

Mello, U. (2022). Centralized admissions, affirmative action, and access of low-income students to higher education. *American Economic Journal: Economic Policy 14*(3), 166–97.

Mitrulis, E. and S. T. d. S. Penin (2006). Pré-vestibulares alternativos: da igualdade à eqüidade. *Cadernos de Pesquisa 36*, 269–298.

Neal, D. and D. W. Schanzenbach (2010). Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics 92*(2), 263–283.

Nielsen, E. (2023). Test questions, economic outcomes, and inequality. Working Paper.

OECD (2021). *Education in Brazil: An International Perspective.* OECD Publishing, Paris.

Ors, E., F. Palomino, and E. Peyrache (2013). Performance gender gap: does competition matter? *Journal of Labor Economics 31*(3), 443–499.

Pallais, A. (2015). Small differences that matter: Mistakes in applying to college. *Journal of Labor Economics 33*(2), 493–520.

RAIS (2022). Relação Anual de Informações Sociais. Ministério do Trabalho e Emprego, Brasília, Brazil. http://www.rais.gov.br/ (accessed April 2022).

Reback, R., J. Rockoff, and H. L. Schwartz (2014). Under pressure: Job security, resource allocation, and productivity in schools under No Child Left Behind. *American Economic Journal: Economic Policy 6*(3), 207–41.

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

Riehl, E. (2023). Do less informative college admission exams reduce earnings inequality? evidence from Colombia. *Journal of Labor Economics*. Forthcoming.

Riehl, E. and M. Welch (2023). Accountability, test prep incentives, and the design of math and English exams. *Journal of Policy Analysis and Management 42*(1), 60–96.

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

Schlosser, A., Z. Neeman, and Y. Attali (2019). Differential performance in high versus low stakes tests: evidence from the GRE test. *The Economic Journal 129*(623), 2916–2948.

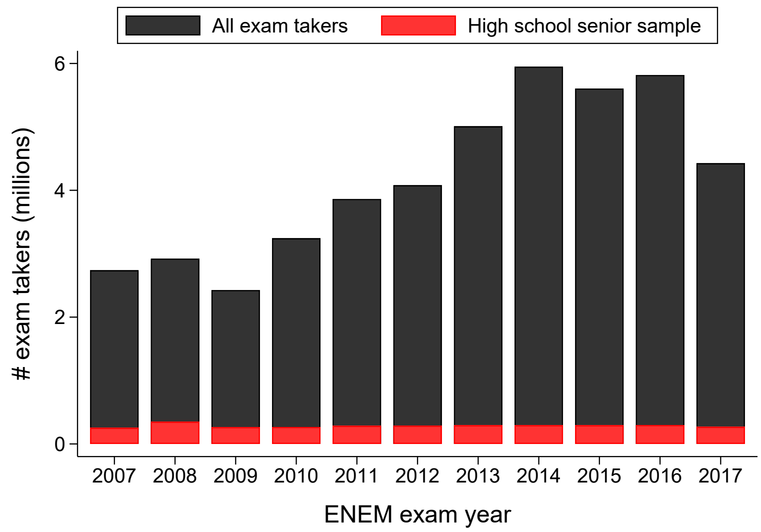
Scott-Clayton, J. (2012). Do high-stakes placement exams predict college success? ccrc working paper no. 41. *Community College Research Center, Columbia University*.

Segal, C. (2012). Working when no one is watching: Motivation, test scores, and economic success. *Management Science 58*(8), 1438–1457.

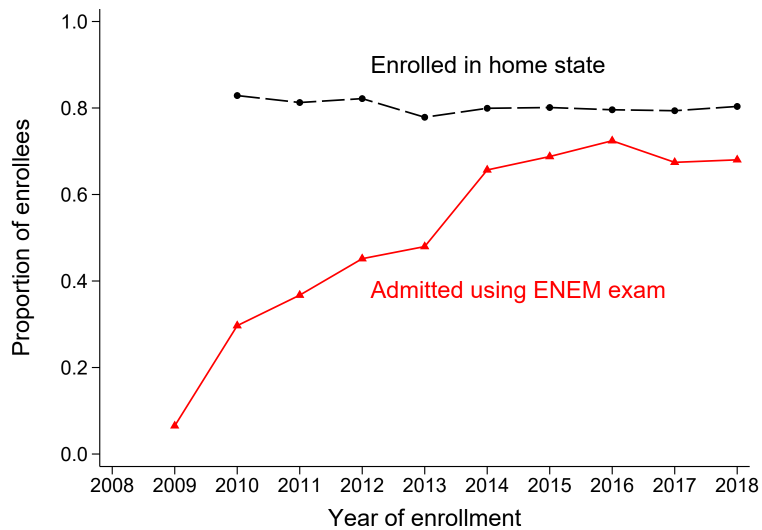
Soares, J. A. (2015). *SAT wars: The case for test-optional college admissions*. Teachers College Press.

Wise, S. L. and C. E. DeMars (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational assessment 10*(1), 1–17.

Figures and tables



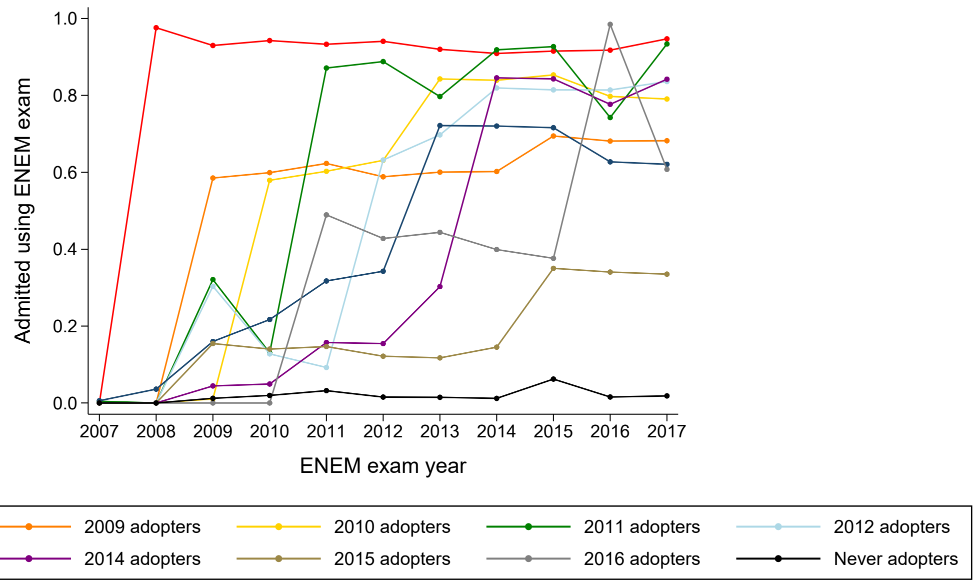
Panel A. Number of ENEM exam takers



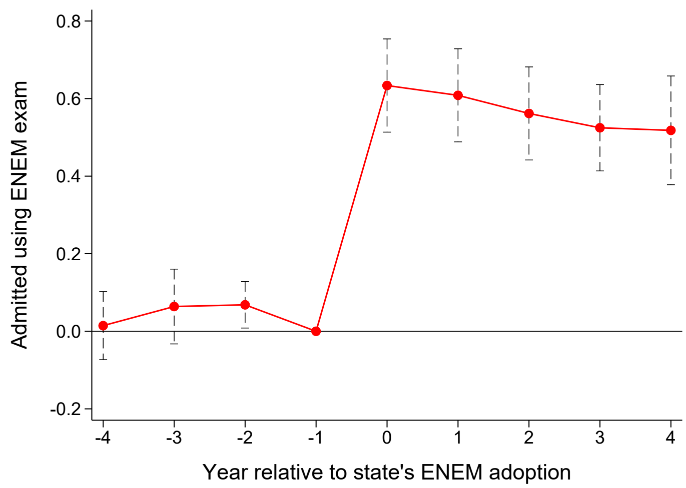
Panel B. Federal university enrollment

Figure 1. Adoption of ENEM exam by federal universities

Notes: This figure illustrates the impact of increased ENEM stakes on the number of exam takers and federal university enrollees through the ENEM. Panel A shows the total number of exam takers of the ENEM (including the pre-2009 version) over the 2007-2017 period. Each bar displays the overall number of ENEM exam takers (black area) and the number of exam takers in our analysis sample of high school graduates (red area). Panel B plots the fraction of new enrollees to federal universities admitted through the ENEM exam (red line) and that enrolled in the state where they were born (black dashed line) over the 2009-2018 period.



Panel A. Proportion of federal university enrollees admitted using ENEM exam by ENEM adoption year



Panel B. Event study for proportion of federal university enrollees admitted using ENEM exam

FIGURE 2. Variation in ENEM exam adoption by federal universities across states and years

Notes: This figure illustrates the staggered adoption of the ENEM exam by federal universities. The outcome in both panels is the proportion of new enrollees in federal universities in state s who were admitted using the ENEM exam administered in year t (the calendar year prior to enrollment), which we denote by $\text{ProportionENEM}_{st}$. Panel A plots the mean of $\text{ProportionENEM}_{st}$ in groups of state(s) based on their ENEM adoption year, τ_s^*, denoted by the legend. Panel B plots event-study coefficients, β_l, from the state (s) $\times$ year (t) $\times$ pairwise group (g) level regression

$$\text{ProportionENEM}_{stg} = \gamma_{sg} + \gamma_{tg} + \sum_{l=-7}^{7} \beta_l \mathbb{1}\{t - \tau_s^* = l\} + \epsilon_{stg}$$

where l denotes years relative to τ_s^*. We include state $\times$ group dummies, γ_{sg}, year $\times$ group dummies, γ_{tg}, and dummies for years l, $\mathbb{1}\{t - \tau_s^* = l\}$, omitting $l = -1$. The graph plots the β_l coefficients from $l = -4$ to 4. Dashed lines depict 95% confidence intervals using standard errors clustered at the state level. See Appendix Tables A2–A3 for details on $\text{ProportionENEM}_{st}$ and each state's ENEM adoption year, τ_s^*. See the text in Section 3.4 and Appendix Table A4 for details on the dataset of pairwise ENEM adoption years that we use for our event studies.

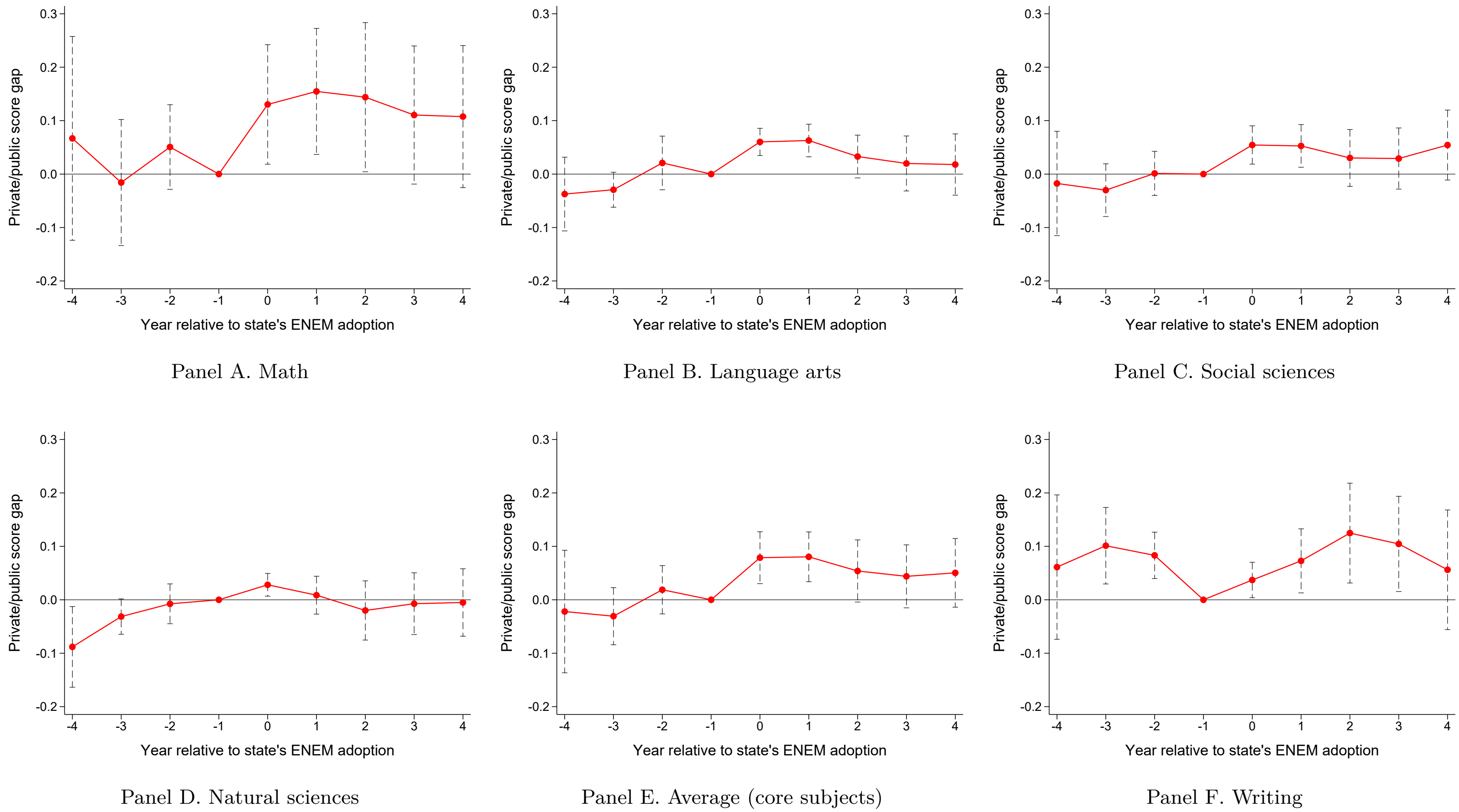
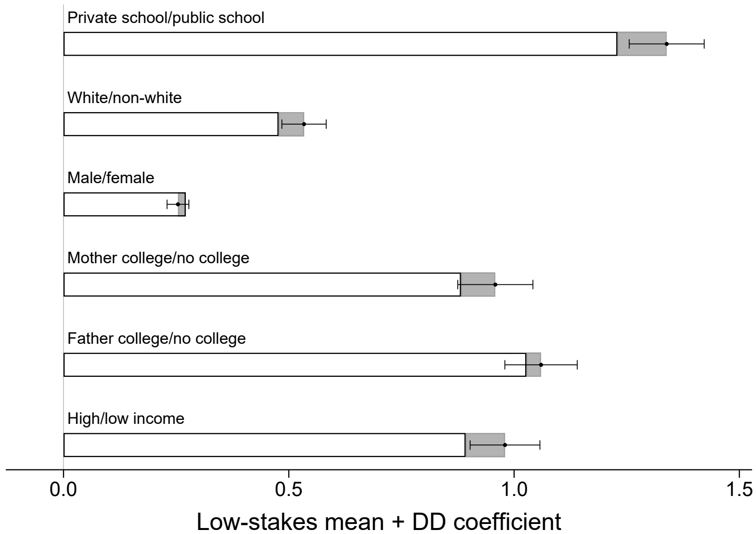
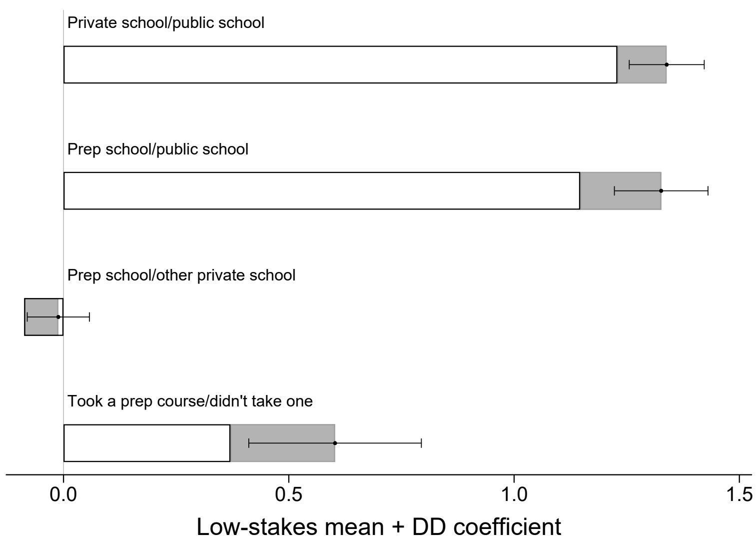


FIGURE 3. Event studies for effects of ENEM adoption on private/public test score gaps

Notes: This figure plots event study estimates of the effects of ENEM stakes on test score gaps between private and public school students. The sample includes all pairwise combinations of ENEM adoption years for which we can estimate treatment effects using 2009–2017 exam takers (the boxed cells in Appendix Table A4). The dependent variables are ENEM subject scores in SD units. "Average (core subjects)" is the average score across math, language arts, natural science, and social science. Each panel plots the β_l^{gap} coefficients (y-axis) from our event study regression (4) for years $l = -4$ to 4 relative to the state's ENEM adoption year, $\tau^*_{s(i)}$ (x-axis). Dashed lines depict 95% confidence intervals using standard errors clustered at the state level.



Panel A. Gaps by demographic characteristics



Panel B. Gaps by test prep activity

FIGURE 4. Effects of ENEM adoption on gaps in average (core subject) scores

Notes: This figure illustrates the effects of the ENEM stakes on different gaps in average (core subject) ENEM scores.

Panel A shows impacts on demographic test score gaps. "High-income" individuals are those with family income greater or equal to two times the minimum wage. Panel B shows impacts on test score gaps between students who did/did not engage in test prep activity as defined by two different measures. We define "prep schools" as private schools whose curriculum is specifically focused on preparation for college admission exams. For this, we obtained lists of schools that use test-oriented curricula from the websites of four prominent test prep companies and matched them to our sample of high schools using geocoded addresses. For the last bar in Panel B, we use a variable from the ENEM questionnaire that indicates whether individuals took an entrance exam preparation course.

In both panels, the white bars represent mean gaps in average ENEM scores in low-stakes cohorts for each demographic/test prep group. The grey bars represent estimates of β^{gap} from a version of equation (2) that replaces Private_h with a dummy for the first group listed in the heading (e.g., white, male, etc.) For this figure, we estimate equation (2) at the individual level rather than the high-school $\times$ year level. The black error bars represent 95% confidence intervals using standard errors clustered at the state level.

See Appendix Tables A6 and A7 for details on variable definitions and the point estimates from these regressions.

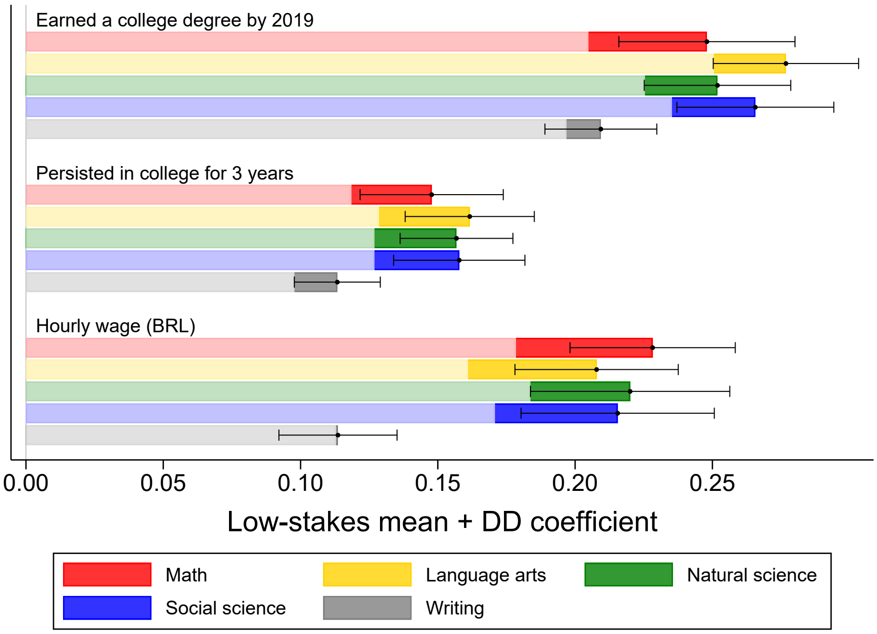
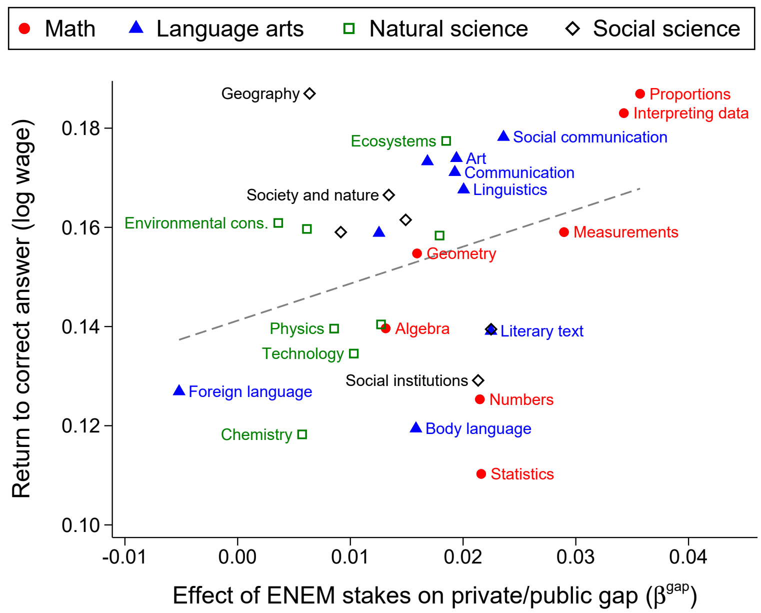
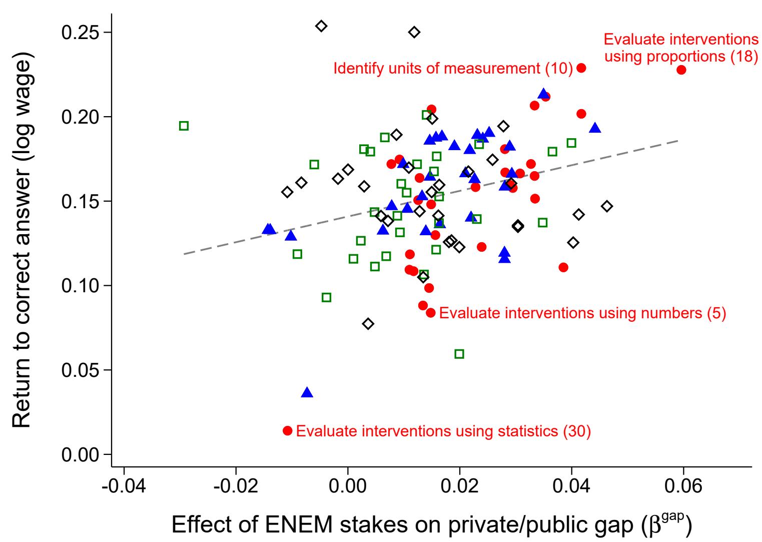


Figure 5. Effects of ENEM adoption on the correlation between outcomes and test scores by subject

Notes: This figure illustrates the effects of the ENEM stakes on the correlation between student outcomes and test scores by subject. The lighter-shaded areas depict the average correlation coefficients between subject scores and outcomes in low-stakes cohorts (i.e., cohorts with $\text{HighStakes}_{st} = 0$). The darker-shaded areas depict β coefficients from regression (5) in which the outcome variables, Y_{st}, are state $\times$ year correlation coefficients between subject scores and outcomes. The black error bars represent 95% confidence intervals using standard errors clustered at the state level.



Panel A. Topic areas



Panel B. Competencies

FIGURE 6. Preppability of ENEM skills and their informativeness for log wages

Notes: This figure shows how ENEM exam skills vary in their informativeness for individuals' log wages (y-axis) and their preppability (x-axis). The y-axis shows our measure of exam skill informativeness, which is the average log wage *return to a correct answer* for questions in different topic areas (Panel A) and competencies (Panel B). For this we compute the difference in the mean log wages for students who got each ENEM question correct and incorrect using 2009–2010 and 2012 ENEM participants in our high school senior sample, and then average these returns at the topic area/competency level. The x-axis displays our measure of exam skill preppability, which is the β^{gap} coefficient from estimating equation (2) separately for groups of ENEM questions in each topic area (Panel A) and competency (Panel B). The dependent variable in these regressions is an indicator for a correct answer, and the sample includes 2009–2017 ENEM participants in our high school senior sample. Marker colors and shapes indicate the exam subject, as described in the legend. The dashed line shows the OLS relationship between informativeness and preppability.

Table 1. Summary statistics for 2009–2017 ENEM exam takers

	(A)	(B)	(C)	(D)	(E)	(F)
			Analysis sample (high school seniors)			
	All exam takers	All HS seniors	All schools	Private schools	Public schools	Private/ public gap
Panel A. Exam taker characteristics						
Age at exam	22.14	18.55	17.91	17.44	18.13	−0.70
Female	0.58	0.59	0.58	0.55	0.60	−0.05
White	0.40	0.44	0.51	0.69	0.43	0.26
Black	0.12	0.11	0.08	0.04	0.10	−0.06
Brown	0.44	0.42	0.37	0.23	0.44	−0.21
Mother attended college	0.15	0.18	0.27	0.56	0.13	0.44
Father attended college	0.11	0.13	0.21	0.49	0.07	0.41
Family income > 2x min. wage	0.35	0.38	0.49	0.85	0.32	0.52
Private high school	0.24	0.24	0.32	1.00	0.00	1.00
Panel B. ENEM scores						
Math score	−0.03	−0.01	0.32	1.28	−0.13	1.42
Language arts score	0.08	0.04	0.24	0.78	−0.01	0.79
Natural science score	−0.17	−0.18	0.05	0.75	−0.28	1.03
Social science score	0.30	0.22	0.43	1.07	0.14	0.93
Average score (core subjects)	0.05	0.02	0.30	1.12	−0.09	1.20
Writing score	−0.41	−0.38	−0.13	0.50	−0.43	0.93
Panel C. College and labor market outcomes						
Ever enrolled in college			0.76	0.95	0.67	0.27
Enrolled in a federal university			0.16	0.26	0.11	0.15
Graduated college within 5 years			0.17	0.23	0.15	0.08
Ever graduated college			0.31	0.43	0.25	0.18
Persisted in college for 3 years			0.66	0.73	0.61	0.12
Fraction of college credits completed			0.69	0.75	0.64	0.11
Appears in RAIS			0.26	0.20	0.29	−0.08
Hourly wage (BRL)			48.89	70.03	41.59	28.44
Number of exam takers	40,391,604	11,626,416	2,512,214	807,293	1,704,921	2,512,214
Number of high schools	46,584	45,867	3,276	1,437	1,839	3,276

Notes: This table reports summary statistics on the ENEM exam takers. Column (A) includes all the ENEM exam takers with a valid score (i.e., non-zero, non-missing) in all four subjects of the ENEM. Column (B) includes all the exam takers who were high school seniors. Column (C) includes all the exam takers in our analysis sample. Columns (D) and (E) include the exam takers in our analysis sample that attended a private and public high school, respectively. Column (F) displays the difference between columns (D) and (E).

Panel (A) describes demographic characteristics of the exam takers, including age, gender, race, parental education, family income, and whether they attended a private high school. Panel (B) reports the average ENEM scores (in SD units) in the respective samples. "Average score (core subjects)" is the average score across math, language arts, natural science, and social science. Panel (C) displays the college and labor market outcomes for the exam takers in our analysis sample. The last two rows reports the number of exam takers and high schools in the respective samples. See Appendix C.1 for detailed variable definitions.

Table 2. Balance tests for exam taker characteristics

	(A)	(B)	(C)	(D)	(E)
	Low-stakes mean	DD coefficients			
Dependent variable	All schools	All schools	Private schools	Public schools	Private/ public gap
Panel A. Number of exam takers per school × year					
Number of exam takers	151.713	14.225	1.308	14.966*	−13.658
		(8.985)	(18.395)	(8.073)	(18.169)
Log number of exam takers	4.713	0.076	−0.004	0.089	−0.093
		(0.065)	(0.118)	(0.056)	(0.107)
Panel B. Demographic characteristics of exam takers					
Age at exam	18.190	0.030	0.010	0.073	−0.063
		(0.054)	(0.015)	(0.083)	(0.075)
Female	0.599	−0.014**	−0.013**	−0.012*	−0.000
		(0.006)	(0.005)	(0.006)	(0.006)
White	0.469	−0.007	−0.009	0.001	−0.010
		(0.010)	(0.008)	(0.010)	(0.012)
Mother attended college	0.258	0.001	0.009	0.000	0.009
		(0.008)	(0.010)	(0.005)	(0.010)
Father attended college	0.195	0.006	0.013	0.005	0.008
		(0.006)	(0.011)	(0.004)	(0.009)
Family income > 2x min. wage	0.476	0.003	0.016	−0.002	0.018
		(0.022)	(0.010)	(0.028)	(0.031)
Joint balance test (p value)		0.159	0.308	0.206	0.708
Panel C. Predicted score based on demographics					
Predicted ENEM score	0.181	0.004	0.017	−0.001	0.018
		(0.011)	(0.012)	(0.011)	(0.014)
N (# exam takers)	492,436	2,512,214	807,293	1,704,921	2,512,214

Notes: This table displays balance tests for the ENEM participants in our analysis sample. The sample includes 2009–2017 ENEM exam takers in our high school senior sample (column C of Table 1). The dependent variables are: the number of exam takers in levels and logs (Panel A); exam taker demographic characteristics (Panel B); and the predicted value from a regression of ENEM scores (averaged across math, language arts, natural science, and social science) on age, gender, and dummies for race, mother's education, father's education, and family income bins (Panel C). Our dependent variables are high school × year totals (Panel A) and averages (Panel B–C) of these variables.

Column (A) shows the mean of each dependent variable in exam cohorts prior to each state's ENEM adoption year, i.e., cohorts with $\text{HighStakes}_{st} = 0$. Columns (B)–(D) display β coefficients from equation (1) estimated using all students, private students, and public students, respectively. Column (E) displays β^{gap} coefficients from equation (2) estimated using all students. The last row of Panel B shows the p value from an F test that the coefficients in Panel B are jointly equal to zero.

Parentheses contain standard errors clustered at the state level.

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

Table 3. Effects of ENEM adoption on test scores in public and private high schools

	(A)	(B)	(C)	(D)	(E)
	Low-stakes mean	DD coefficients			
Dependent variable	Private/ public gap	All schools	Private schools	Public schools	Private/ public gap
Math score	1.358	0.022	0.143**	−0.015	0.158*
		(0.055)	(0.058)	(0.070)	(0.079)
Language arts score	0.837	0.035	0.068***	−0.008	0.076***
		(0.035)	(0.020)	(0.034)	(0.026)
Natural science score	1.059	0.026	0.062*	−0.003	0.065*
		(0.040)	(0.031)	(0.042)	(0.034)
Social science score	1.010	0.019	0.056*	−0.024	0.081***
		(0.034)	(0.029)	(0.029)	(0.023)
Average score (core subjects)	1.229	0.029	0.095**	−0.014	0.110**
		(0.043)	(0.036)	(0.045)	(0.040)
Writing score	0.784	0.049	0.165**	0.063*	0.102*
		(0.035)	(0.072)	(0.033)	(0.058)
N (# exam takers)	492,436	2,512,214	807,293	1,704,921	2,512,214

Notes: This table shows how the increase in the stakes of the ENEM exam impacted scores for private and public high school students. The sample includes 2009–2017 ENEM exam takers in our high school senior sample (column C of Table 1). The dependent variables are ENEM subject scores in SD units. "Average score (core subjects)" is the average score across math, language arts, natural science, and social science.

Column (A) shows the mean private/public score gap in exam cohorts prior to each state's ENEM adoption year, i.e., cohorts with $\text{HighStakes}_{st} = 0$. Columns (B)–(D) display β coefficients from equation (1) estimated using all students, private students, and public students, respectively. Column (E) displays β^{gap} coefficients from equation (2) estimated using all students.

Parentheses contain standard errors clustered at the state level.

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

Table 4. Robustness checks on the effects of ENEM adoption on private/public test score gaps

Dependent variable	(A) Benchmark model	(B) Demo-graphic controls	(C) Binary treatment	(D) Stacked regression	(E) 2009 vs 2013 adopters (2007–2012)	(F) Stacked regression (2007–2017)
Math score	0.158*	0.128*	0.105**	0.116**	0.086***	0.061**
	(0.079)	(0.067)	(0.050)	(0.055)	(0.019)	(0.024)
Language arts score	0.076***	0.072***	0.042**	0.055***	0.106***	0.062***
	(0.026)	(0.023)	(0.017)	(0.014)	(0.031)	(0.022)
Natural science score	0.065*	0.074**	0.032	0.025	0.059***	0.022
	(0.034)	(0.036)	(0.030)	(0.016)	(0.017)	(0.014)
Social science score	0.081***	0.053*	0.042**	0.053**	0.046	0.042
	(0.023)	(0.028)	(0.020)	(0.020)	(0.035)	(0.025)
Average score (core subjects)	0.110**	0.094**	0.064**	0.072***	0.088**	0.049**
	(0.040)	(0.040)	(0.030)	(0.023)	(0.029)	(0.024)
Writing score	0.102*	0.144**	0.035	0.023	0.058	0.064***
	(0.058)	(0.058)	(0.044)	(0.028)	(0.047)	(0.021)
N (# exam takers)	2,512,214	2,512,214	2,512,214	5,858,862	1,099,500	15,738,474
Treatment variable:	Continuous	Continuous	Binary	Binary	Binary	Binary
Demographic controls:		Yes				
Level of dataset:	HS × year	HS × year	HS × year	Stacked	HS × year	Stacked
Included exam cohorts:	2009–2017	2009–2017	2009–2017	2009–2017	2007–2012	2007–2017

Notes: This table examines the robustness of our estimates for the effects of ENEM stakes on the private/public school gap in test scores. Our main sample includes 2009–2017 ENEM exam takers in our analysis sample (column C of Table 1). In columns (E)–(F), we add in 2007–2008 ENEM exam takers from the same set of high schools. The dependent variables are ENEM subject scores in SD units. "Average score (core subjects)" is the average score across math, language arts, natural science, and social science. Columns (E)–(F) include scores from the 2007–2008 ENEM tests; in these columns, we standardize scores to have mean 0 and SD 1 within each year of our sample. For the 2007–2008 exams, "average score" is the reported core-component score, and we compute math, language arts, natural science, and social science scores by categorizing the multiple choice questions into these subjects and then estimating scores using the IRT parameters. See Appendix C.1 for details.

Column (A) replicates the estimates from column (E) of Table 3, which are the β^{gap} coefficients from equation (2). Column (B) estimates equation (2) including high school × year averages of age, gender, and dummies for race, mother's education, father's education, and family income bins. Column (C) estimates equation (2) replacing our continuous treatment variable, $\text{ProportionENEM}_{st}$, with our binary treatment variable, HighStakes_{st}. Columns (D)–(F) display estimates of β^{gap} from equation (3) using our stacked dataset, which contains pairwise combinations of ENEM adoption years (as described in Section 3.4). Column (D) includes all pairwise combinations for which we can estimate treatment effects using 2009–2017 exam takers (the boxed cells in Appendix Table A4). Column (E) includes 2007–2012 exam takers and a single pair of ENEM adoptions years: 2009 and 2013 (the bolded cells in Appendix Table A4). Column (F) includes all 2007–2017 exam takers and all pairwise combinations (all cells in Appendix Table A4).

Parentheses contain standard errors clustered at the state level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TABLE 5. Heterogeneity by topic area and competency — Math exam
Dependent variable: Proportion correct answers

(A)	(B)	(C)	(D)	(E)	(F)
		Public students		Private/public gap	
Question group	N_q	Mean	β (SE)	Mean	β^{gap} (SE)
Panel A. All questions					
All questions	405	0.291	−0.005 (0.010)	0.176	0.024 (0.009)**
Panel B. Topic area (and competency reference numbers)					
Numbers (1–5)	67	0.307	−0.010 (0.011)	0.159	0.021 (0.011)*
Geometry (6–9)	57	0.317	0.003 (0.007)	0.160	0.016 (0.008)*
Measurements (10–14)	62	0.257	−0.006 (0.010)	0.193	0.029 (0.009)***
Proportions (15–18)	51	0.336	−0.008 (0.019)	0.225	0.036 (0.015)**
Algebra (19–23)	66	0.264	−0.003 (0.005)	0.172	0.013 (0.007)*
Interpreting data (24–26)	47	0.325	−0.014 (0.017)	0.193	0.034 (0.015)**
Statistics (27–30)	55	0.241	−0.000 (0.007)	0.137	0.022 (0.009)**
All coefficients equal (p value)			0.299		0.015
Panel C. Competencies (top 5 and bottom 5 by β^{gap}/mean)					
Evaluate interventions using proportions (18)	12	0.293	−0.012 (0.023)	0.219	0.060 (0.017)***
Use tables/graphs to construct arguments (26)	14	0.363	0.001 (0.023)	0.233	0.042 (0.019)**
Identify units of measurement (10)	10	0.375	−0.001 (0.020)	0.313	0.042 (0.018)**
Calculate statistical quantities from data (27)	15	0.220	0.005 (0.010)	0.140	0.039 (0.022)*
Identify proportional relationships (15)	12	0.395	−0.013 (0.025)	0.274	0.035 (0.023)
. . .					
Use numbers to construct arguments (4)	15	0.266	−0.007 (0.010)	0.161	0.011 (0.012)
Use algebra to construct arguments (22)	9	0.211	0.006 (0.008)	0.107	0.011 (0.009)
Solve problems using geometry (8)	18	0.236	0.013 (0.003)***	0.147	0.009 (0.010)
Interpret Cartesian graphs (20)	11	0.541	−0.018 (0.027)	0.209	0.008 (0.039)
Evaluate interventions using statistics (30)	10	0.253	0.001 (0.011)	0.072	−0.011 (0.016)
21 coefficients equal (p value)			0.000		0.000

Notes: This table shows how the increase in ENEM stakes impacted students' performance on different topic areas and competencies of the math exam. The sample includes 2009–2017 ENEM exam takers in our high school senior sample (column C of Table 1). Regressions are at the high school (h) × year (t) × exam question (q) level. The dependent variable is the proportion of correct answers in each htq cell. We estimate regressions pooling across all math questions (Panel A) and separately for math questions in 7 topic areas (Panel B) and 30 competencies (Panel C) defined by ENEM test designers. Panel C reports only the top 5/bottom 5 competencies by the values in column (F). See Appendix C.4 for details on the categorization of ENEM math questions.

Column (A) defines the group of questions for each regression. Column (B) shows the number of questions in each group. Column (C) shows the mean proportion of correct answers for public school students in cohorts prior to each state's ENEM adoption year (i.e., cohorts with $\text{HighStakes}_{st} = 0$). Column (E) shows the mean private/public gap in the proportion of correct answers in those cohorts. Columns (D) and (F) display the β and β^{gap} coefficients from equation (2) estimated for each group of questions. In Panel B, the last row reports p values from F tests that the 7 topic area coefficients in columns (D) or (F) are equal. In Panel C, the last row reports p values from F tests that 21 competency coefficients (the first 3 in each topic area) are jointly equal.

Parentheses contain standard errors clustered at the state level.

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

Table 6. Effects of ENEM adoption on the correlation of average ENEM scores and outcomes

	(A)	(B)	(C)	(D)	(E)
	Low-stakes mean	Benchmark model DD coefficients		Binary treatment DD coefficients	
Dependent variable	Raw corr.	Raw corr.	Within-program	Raw corr.	Within program
Panel A. Outcomes for all exam takers					
Enrolled in any college by 2019	0.372	0.036***		0.022***	
		(0.006)		(0.004)	
Finished college within 5 years of ENEM	0.121	0.014*	0.026***	0.007	0.015***
		(0.008)	(0.007)	(0.004)	(0.005)
Earned a college degree by 2019	0.257	0.033**	0.038***	0.016	0.023***
		(0.015)	(0.010)	(0.011)	(0.007)
Appears in RAIS in 2016–2018	−0.112	0.056	0.020*	0.014	0.008
		(0.044)	(0.011)	(0.025)	(0.007)
N (# exam takers)	336,175	1,266,412	1,266,412	1,266,412	1,266,412
Panel B. Outcomes for college enrollees					
Persisted in college for 1 year	0.064	0.008	0.024***	0.007	0.011*
		(0.014)	(0.007)	(0.011)	(0.006)
Persisted in college for 3 years	0.142	0.033**	0.043***	0.018**	0.024***
		(0.013)	(0.008)	(0.009)	(0.006)
Completed program within 5 years	0.071	0.044**	0.035***	0.022*	0.018**
		(0.018)	(0.011)	(0.011)	(0.007)
Fraction of college credits completed	0.214	0.003	0.014	−0.011	0.013*
		(0.014)	(0.013)	(0.011)	(0.007)
N (# in higher ed.)	274,022	966,649	966,649	966,649	966,649
Panel C. Outcome for individuals in RAIS					
Hourly wage (BRL)	0.200	0.046**	0.027**	0.027***	0.012*
		(0.018)	(0.010)	(0.007)	(0.006)
Log hourly wage	0.362	−0.029**	−0.001	−0.017*	−0.001
		(0.014)	(0.010)	(0.008)	(0.006)
N (# in RAIS)	80,382	328,773	328,773	328,773	328,773

Notes: This table shows how the increase in the stakes of the ENEM exam impacted the correlation between average ENEM scores and student outcomes. For each state × year pair, we compute correlation coefficients between the outcomes in the column header and average (core subject) ENEM scores using 2009–2010 and 2012–2014 ENEM participants in our high school senior sample. We then estimate equation (5) using these correlation coefficients as dependent variables, weighting each state × year observation by the number of ENEM participants for which the outcome is defined. Column (A) shows the mean correlation coefficients in exam cohorts prior to each state's ENEM adoption year (i.e., cohorts with $\text{HighStakes}_{st} = 0$). Columns (B)–(E) β coefficients from equation (5). Columns (B) and (D) use raw correlation coefficients as dependent variables. In columns (C) and (E), the dependent variables are correlation coefficients computed after demeaning all variables within college × program cells. Regressions in columns (B)–(C) use our continuous treatment variable, $\text{ProportionENEM}_{st}$. Columns (D)–(E) use our binary treatment variable, HighStakes_{st}. Parentheses contain standard errors clustered at the state level.

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

Table 7. OLS relationships between the informativeness and preppability of ENEM competencies

	(A)	(B)	(C)	(D)	(E)	(F)
	Mean return to correct	Change in return to a correct answer from 1pp increase in β^{gap}				
Dependent variable	All subjs	Math	Lang. arts	Nat. science	Soc. science	All subjs
Panel A. Outcomes for all exam takers						
Enrolled in any college by 2019	0.101	0.010**	0.008**	0.000	−0.000	0.004**
		(0.005)	(0.004)	(0.004)	(0.002)	(0.002)
Finished college within 5 years of ENEM	0.029	0.003*	0.003***	0.000	0.000	0.001
		(0.002)	(0.001)	(0.002)	(0.001)	(0.001)
Earned a college degree by 2019	0.071	0.008**	0.007***	0.001	−0.000	0.003*
		(0.003)	(0.002)	(0.003)	(0.002)	(0.002)
Appears in RAIS in 2016–2018	−0.016	−0.000	0.006***	0.000	−0.001	0.001*
		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
N (# competencies)	120	30	30	30	30	120
Panel B. Outcomes for college enrollees						
Persisted in college for 1 year	0.012	0.001**	0.000	0.000	−0.000	0.000*
		(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
Persisted in college for 3 years	0.047	0.005***	0.002	0.000	−0.000	0.002**
		(0.002)	(0.001)	(0.002)	(0.001)	(0.001)
Completed program within 5 years	0.023	0.002	0.003***	0.000	0.001	0.001
		(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
Fraction of college credits completed	0.048	0.005***	0.004***	0.001	−0.001	0.002**
		(0.002)	(0.001)	(0.002)	(0.001)	(0.001)
N (# competencies)	120	30	30	30	30	120
Panel C. Outcome for individuals in RAIS						
Hourly wage (BRL)	13.285	2.103***	1.217***	0.027	−0.616	0.573**
		(0.500)	(0.416)	(0.503)	(0.370)	(0.285)
Log hourly wage	0.153	0.023***	0.014***	0.001	−0.005	0.008**
		(0.005)	(0.005)	(0.005)	(0.004)	(0.003)
N (# competencies)	120	30	30	30	30	120

Notes: This table presents OLS relationships between the informativeness and preppability of ENEM competencies. Our measure of informativeness is the average *return to a correct answer* at the competency level defined using the outcomes in the column header. This return is the difference in the mean outcome for students who got each ENEM question correct and incorrect, which we estimate using 2009–2010 and 2012–2014 ENEM participants in our high school senior sample (for outcomes in Panel C, the sample excludes 2013–2014 ENEM participants). Our measure of preppability is the β^{gap} coefficient from equation (2) estimated separately for groups of questions in each competency (as in Table 5, Panel C, column F). The dependent variable in these regressions is an indicator for a correct answer, and the sample includes 2009–2017 ENEM participants in our high school senior sample. Column (A) shows the mean return to a correct answer averaged across all subjects and competencies. Columns (B)–(F) show the OLS coefficients from bivariate regressions of the mean returns to a correct answer on the β^{gap} coefficients, estimated by subject and pooling across all subjects. We normalize the OLS coefficients to represent a 1pp increase in β^{gap}. Parentheses contain robust standard errors.

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

Appendix — For Online Publication

A. Appendix figures and tables

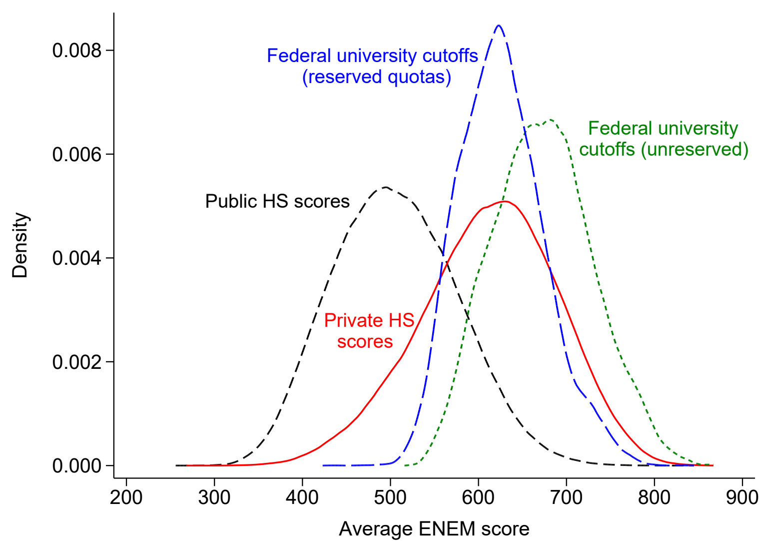
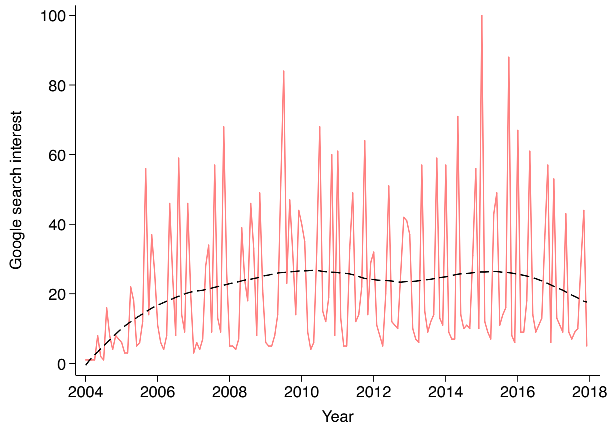


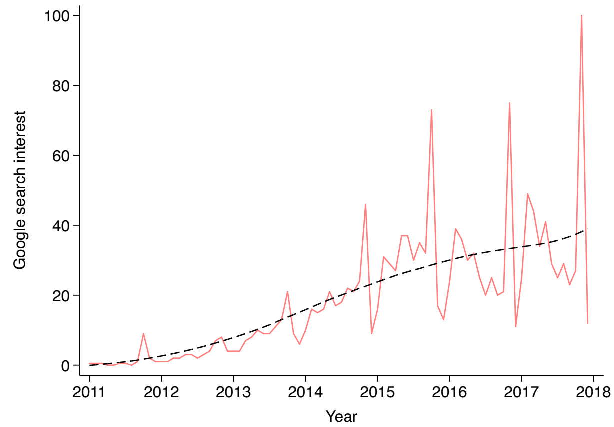
Figure A1. Distributions of average ENEM scores and federal university admission cutoffs

Notes: This figure compares the distributions of actual ENEM scores and federal univeristy cutoff scores. The solid red line shows the distribution of average ENEM scores for private school students in our sample who took the ENEM exam in exam cohorts prior to each state's ENEM adoption year, i.e., cohorts with $\text{HighStakes}_{st} = 0$. The dashed black line shows the same distribution for public school students. In both distributions, average ENEM scores are the average score across math, language arts, natural science, and social science.

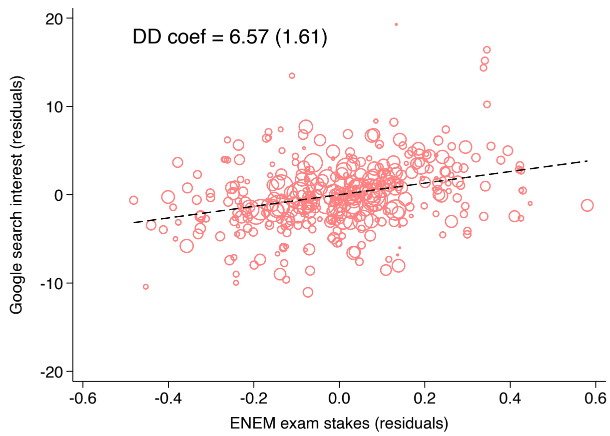
The green short-dashed line plots the distribution of cutoff scores for unreserved admissions to all federal university programs in 2016. The blue long-dashed line plots the same distribution for reserved quotas at federal university programs, which include quotas for public high students, low-SES students, and/or underrepresent minority students.



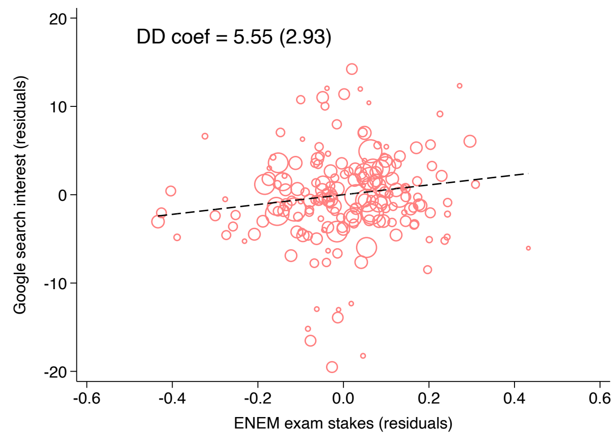
Panel A. National trend in
Google search for "ENEM"



Panel B. National trend in
Google search for "Descomplica"



Panel C. Effect of ENEM stakes on
Google search for "ENEM"



Panel D. Effect of ENEM stakes on
Google search for "Descomplica"

FIGURE A2. Google search trends for "ENEM" and the online test prep service "Descomplica"

Notes: This figure shows how Google search trends for "ENEM" and the online test prep service "Descomplica" varied over time at the national and state levels. Panels A and C show results for the search "ENEM" using data from 2004–2017. Panels B and D show results for the search "Descomplica" using data from 2011–2017; we do not include years prior to 2011 because search volume for "Descomplica" was low and state-level data are noisy.

Panels A–B plot monthly Google search interest for the entire country of Brazil (solid red lines) plus non-parametric predicted values (black dashed lines). A value of 100 is the peak popularity for the term. A value of 50 means that the term is half as popular. A score of 0 means there was not enough data for this term.

Panels C–D plot state × year level means of Google search interest (y-axis) against our conitnuous treatment variable, ProportionENEM$_{st}$ (x-axis). We residualize both variables on state and year dummies, and plot the residuals along with a linear regression line (black dashed line). We also display DD coefficients (and standard errors) from estimating equation (1) using Google search interest for each term as the dependent variable. These DD coefficients are equivalent to the slopes of the black dashed lines in Panels C–D.

Table A1. Summary of Brazilian high school and college markets

(A) Institution type	(B) # of schools	(C) Prop. of schools	(D) # of students	(E) Prop. of students	(F) # students per school	(G) Attended a private HS
Panel A. High school seniors in 2009						
Federal high schools	100	0.004	9,772	0.005	98	1.000
State high schools	16,583	0.702	1,823,524	0.849	110	0.000
Municipal high schools	373	0.016	23,156	0.011	62	0.000
Private high schools	6,567	0.278	290,366	0.135	44	1.000
All high schools	23,623	1.000	2,146,818	1.000	91	0.140
Panel B. New college enrollees in 2009						
Federal universities	59	0.025	225,112	0.108	3,815	0.471
State universities	40	0.017	119,489	0.057	2,987	0.370
Municipal universities	9	0.004	22,453	0.011	2,495	0.319
Private universities	225	0.094	1,018,698	0.489	4,528	0.458
Public technical colleges	168	0.070	55,609	0.027	331	0.259
Private technical colleges	1,888	0.790	640,021	0.307	339	0.331
All colleges	2,389	1.000	2,081,382	1.000	871	0.401

Notes: This table presents summary statistics on the Brazilian high school and college systems. Panel A presents statistics for students who were high school seniors in 2009 using data from a national primary and secondary school census (*Censo Escolar*). Panel B presents statistics for students who were new college enrollees in 2009 using data from a national higher education census (*Censo da Educação Superior*).

Column (A) categorizes high schools by ownership (federal, state, municipal, or private), and it categorizes colleges by both ownership and institution type (university or technical college). University includes both *Universidade* and *Centro Universitário* institutions. Technical colleges include *Faculdade*, *Instituto Federal de Educação Ciência e Tecnologia*, and *Centro Federal de Educação Tecnológica* institutions. Column (B) shows the number of schools in each category, and column (C) shows the proportion of schools. Column (D) shows the number of students who attended schools in each category, and column (E) shows the proportion of students. Column (F) shows the number of students per school (column D divided by column B). Column (G) shows the proportion of students at each school type who attended a private high school. Throughout the paper, we include the small number of federal high schools in the group of private high schools since both tend to enroll wealthier and higher-achieving students.

Table A2. Proportion of federal university enrollees admitted using the ENEM by state and year

(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)	(K)	(L)	(M)
		Proportion admitted using ENEM by exam year ($\text{ProportionENEM}_{st}$)										
State	# 2009 enrollees	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Pernambuco (PE)	7,375	0.00	**0.98**	**0.93**	**0.94**	**0.93**	**0.94**	**0.92**	**0.91**	**0.92**	**0.92**	**0.95**
Amazonas (AM)	2,821	0.00	0.00	**0.32**	**0.36**	**0.38**	**0.42**	**0.43**	**0.42**	**0.50**	**0.47**	**0.47**
Espirito Santo (ES)	3,302	0.00	0.00	**0.99**	**0.92**	**1.00**	**0.90**	**0.92**	**0.89**	**0.87**	**0.94**	**0.92**
Maranhão (MA)	2,359	0.00	0.00	**0.98**	**0.88**	**0.95**	**0.94**	**0.91**	**0.87**	**0.89**	**0.78**	**0.92**
Mato Grosso (MT)	3,582	0.00	0.00	**0.64**	**0.93**	**0.90**	**0.86**	**0.83**	**0.94**	**0.95**	**0.89**	**0.97**
Paraná (PR)	6,820	0.00	0.00	**0.53**	**0.51**	**0.51**	**0.47**	**0.58**	**0.60**	**0.63**	**0.57**	**0.57**
Rio Grande do Sul (RS)	12,723	0.00	0.00	**0.42**	**0.37**	**0.42**	**0.41**	**0.39**	**0.41**	**0.62**	**0.63**	**0.62**
São Paulo (SP)	5,774	0.00	0.00	**0.87**	**0.95**	**0.89**	**0.80**	**0.74**	**0.70**	**0.74**	**0.77**	**0.71**
Ceará (CE)	3,704	0.00	0.00	0.00	**0.85**	**0.85**	**0.82**	**0.85**	**0.83**	**0.87**	**0.77**	**0.81**
Mato Grosso do Sul (MS)	2,710	0.00	0.00	0.00	**0.66**	**0.63**	**0.54**	**0.65**	**0.72**	**0.71**	**0.62**	**0.56**
Paraíba (PB)	7,160	0.00	0.00	0.02	**0.40**	**0.45**	**0.55**	**0.94**	**0.90**	**0.92**	**0.93**	**0.90**
Acre (AC)	955	0.00	0.00	0.00	0.00	**0.96**	**0.76**	**0.94**	**0.89**	**0.82**	**0.86**	**0.83**
Alagoas (AL)	3,008	0.00	0.00	0.00	0.00	**0.90**	**0.83**	**0.87**	**0.88**	**0.94**	**0.95**	**0.93**
Piauí (PI)	3,592	0.01	0.00	0.55	0.29	**0.82**	**1.00**	**0.71**	**0.97**	**0.96**	**0.58**	**0.98**
Rio Grande do Norte (RN)	6,443	0.00	0.00	0.34	0.15	0.06	**0.61**	**0.70**	**0.82**	**0.82**	**0.79**	**0.82**
Rondônia (RO)	1,106	0.00	0.00	0.08	0.02	0.31	**0.77**	**0.69**	**0.79**	**0.81**	**0.95**	**0.96**
Bahia (BA)	6,251	0.00	0.00	0.16	0.16	0.31	0.36	**0.90**	**0.83**	**0.82**	**0.81**	**0.77**
Distrito Federal (DF)	5,296	0.00	0.00	0.00	0.01	0.10	0.04	**0.23**	**0.24**	**0.20**	**0.21**	**0.21**
Minas Gerais (MG)	20,918	0.00	0.08	0.19	0.20	0.29	0.36	**0.72**	**0.74**	**0.75**	**0.72**	**0.73**
Pará (PA)	2,937	0.00	0.00	0.00	0.00	0.03	0.09	**0.89**	**0.82**	**0.78**	**0.76**	**0.81**
Rio de Janeiro (RJ)	16,871	0.02	0.02	0.22	0.37	0.49	0.45	**0.77**	**0.77**	**0.78**	**0.50**	**0.48**
Roraima (RR)	842	0.00	0.00	0.00	0.07	0.11	0.17	**0.48**	**0.46**	**0.38**	**0.41**	**0.39**
Goiás (GO)	3,636	0.00	0.00	0.00	0.01	0.14	0.19	0.36	**0.85**	**0.83**	**0.77**	**0.84**
Tocantins (TO)	1,602	0.00	0.00	0.14	0.15	0.20	0.08	0.18	**0.83**	**0.88**	**0.78**	**0.84**
Santa Catarina (SC)	4,632	0.00	0.00	0.15	0.14	0.15	0.12	0.12	0.15	**0.35**	**0.34**	**0.34**
Amapá (AP)	309	0.00	0.00	0.00	0.00	0.49	0.43	0.44	0.40	0.38	**0.98**	**0.61**
Sergipe (SE)	3,309	0.00	0.00	0.01	0.02	0.03	0.02	0.01	0.01	0.06	0.02	0.02
All states	140,037	0.00	0.07	0.30	0.37	0.45	0.48	0.66	0.69	0.72	0.67	0.68

Notes: This table shows the proportion of federal university enrollees who were admitted using the ENEM by state and year. Column (A) lists the 27 states of Brazil. Column (B) shows the number of new federal university enrollees in each state in the 2009 calendar year. Columns (C)–(M) show the proportion of new federal university enrollees who were admitted using the ENEM based on the year students took the ENEM exam (the calendar year prior to enrollment). The sample for these statistics is new enrollees in bachelor's programs at federal universities using data from the Brazilian higher education census.

The numbers in columns (C)–(M) are the values we use for our continuous treatment variable, $\text{ProportionENEM}_{st}$, where s denotes states and t denotes years. Bolded numbers represent state × years that we classify as high stakes using our binary treatment variable, HighStakes_{st}. See Section 3.3 for details on the definition of $\text{ProportionENEM}_{st}$ and HighStakes_{st}.

Table A3. Characteristics of federal universities by state's ENEM adoption year

(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)	(K)	(L)
	Year of state's ENEM adoption (τ_s^*)										
Characteristic	2008	2009	2010	2011	2012	2013	2014	2015	2016	Never	All
Panel A. University characteristics											
# states	1	7	3	3	2	6	2	1	1	1	27
# federal universities	3	15	6	3	3	24	2	1	1	1	59
# 2009 enrollees	7,375	37,381	13,574	7,555	7,549	53,115	5,238	4,632	309	3,309	140,037
Mean university size	2,458	2,492	2,262	2,518	2,516	2,213	2,619	4,632	309	3,309	2,373
Mean cutoff score (2016)	670	682	669	656	657	711	660	707	670	655	688
Panel B. Characteristics of 2009 enrollees											
Age at enrollment	23.64	24.53	23.80	24.31	24.17	24.15	23.60	24.38	26.81	24.76	24.21
Female	0.49	0.46	0.48	0.50	0.45	0.49	0.48	0.44	0.51	0.48	0.48
White	0.42	0.67	0.33	0.46	0.40	0.58	0.46	0.85	0.40	0.30	0.56
Black	0.07	0.07	0.40	0.08	0.07	0.09	0.08	0.04	0.15	0.10	0.12
Brown	0.46	0.19	0.24	0.44	0.51	0.31	0.43	0.09	0.41	0.57	0.29
Private high school	0.40	0.43	0.60	0.26	0.74	0.52	0.58	0.79	0.87	0.62	0.52

Notes: This table display characteristics of federal universities and their student bodies by their state's ENEM adoption year. Columns (B)–(K) categorize the federal universities by the year in which their state adopted the ENEM exam, τ_s^*, as defined in Section 3.3. Column (L) includes all federal universities. The ENEM adoption years for each state are:

- 2008: Pernambuco.
- 2009: Amazonas, Espirito Santo, Maranhão, Mato Grosso, Paraná, Rio Grande do Sul, São Paulo.
- 2010: Ceará, Mato Grosso do Sul, Paraíba.
- 2011: Acre, Alagoas, Piauí.
- 2012: Rio Grande do Norte, Rondônia.
- 2013: Bahia, Distrito Federal, Minas Gerais, Pará, Rio de Janeiro, Roraima.
- 2014: Goiás, Tocantins.
- 2015: Santa Catarina.
- 2016: Amapá.
- Never: Sergipe.

Data on enrollment size and student characteristics are from the Brazilian higher education census. In Panel A, the number of universities, the number of enrollees, and the mean university size are defined using new 2009 enrollees in bachelor's programs at federal universities. In Panel B, some demographic variables are missing in the 2009 census year, so we compute student characteristics using students who enrolled in 2009 but appear in any census year between 2009–2018.

The mean cutoff score (2016) is from a public data request from the centralized admission platform SISU (*Sistema de Seleção Unificada*). These averages correspond to non-reserved quotas for bachelor's degree programs at federal universities in the year 2016. The cutoff scores are typically weighted averages of ENEM scores in up to five subjects (math, language arts, natural science, social science, and writing) and are presented in ENEM scale score units. We obtained the SISU data in March 2020 at:

http://www.consultaesic.cgu.gov.br/busca/dados/Lists/Pedido/Item/displayifs.aspx?List=0c839f31-47d7-4485-ab65-ab0cee9cf8fe&ID=518622&Web=88cc5f44-8cfe-4964-8ff4-376b5ebb3bef.

Table A4. Visualization of stacked dataset for event studies and robustness tests

(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)	(K)	(L)	(M)	(N)
ENEM adoption yr		Number of test takers in sample (1000s) by exam year											
Treated group	Control group	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	Total
2008	2009	114	123										237
2008	2010	29	30	31									90
2008	2011	11	12	13	14								50
2008	2012	13	15	15	16	18							78
2008	2013	68	78	70	81	88	89						474
2008	2014	18	21	20	21	25	25	25					154
2008	2015	12	13	14	15	16	17	17	17				120
2008	2016	7	8	9	9	11	11	11	11	10			86
2008	Never	8	9	10	10	12	12	12	11	11	11	10	115
2009	2010	131	139	121									391
2009	2011	113	121	104	120								456
2009	2012	115	124	105	122	128							594
2009	2013	**170**	**186**	**161**	**187**	**198**	**198**						1,100
2009	2014	120	130	110	126	135	134	136					890
2009	2015	114	122	104	121	126	126	127	129				969
2009	2016	109	116	99	114	121	120	121	123	124			1,048
2009	Never	110	117	100	115	121	121	122	124	125	127	117	1,299
2010	2011	28	28	28	41								125
2010	2012	31	31	29	43	46							180
2010	2013	86	94	84	108	115	118						606
2010	2014	35	37	34	48	52	55	56					318
2010	2015	29	29	28	42	44	46	48	47				314
2010	2016	24	24	23	36	38	40	42	41	40			309
2010	Never	25	25	24	37	39	41	43	42	41	40	36	393
2011	2012	12	13	12	14	16							67
2011	2013	67	76	67	79	85	86						460
2011	2014	16	19	17	19	22	22	22					138
2011	2015	10	11	11	13	14	14	14	14				101
2011	2016	5	6	6	7	8	8	8	8	8			63
2011	Never	6	7	7	8	9	9	9	9	8	9	8	89
2012	2013	70	79	69	81	87	88						474
2012	2014	19	22	19	21	24	25	24					154
2012	2015	13	14	13	15	16	16	16	16				119
2012	2016	8	9	7	9	10	10	10	10	10			84
2012	Never	9	10	8	10	11	11	11	11	11	11	10	112
2013	2014	74	85	74	86	94	95	97					605
2013	2015	68	77	68	80	85	87	89	90				644
2013	2016	63	72	63	74	80	81	83	84	83			682
2013	Never	64	73	63	75	80	82	84	84	84	85	80	855
2014	2015	17	20	17	20	22	23	23	23				167
2014	2016	12	15	12	14	17	17	17	17	18			140
2014	Never	13	16	13	15	17	18	18	18	19	19	18	184
2015	2016	6	7	6	8	8	9	9	9	9			72
2015	Never	7	8	7	9	9	10	10	10	10	11	7	97
2016	Never	2	3	2	3	3	4	4	4	4	4	4	37
Fig 3 & Tab 4, Col D sample				813	1,016	1,054	1,014	740	535	345	181	162	5,859
Tab 4, Col E sample		170	186	161	187	198	198						1,100
Tab 4, Col F sample		2,079	2,274	1,898	2,087	2,052	1,864	1,311	950	615	318	289	15,738

Notes: This table illustrates the stacked dataset for our event studies (Figure 3) and robustness tests (Table 4). Columns (A)–(B) show the pairwise combinations of ENEM adoption years, τ_s^* and $\tau_{s'}^*$. Columns (C)–(N) show the number of observations (in 1000s) in each pair by exam year. Boxed cells show the sample for both Figure 3 and column (D) of Table 4, which is defined by the pairwise treatment effects that we can estimate using 2009–2017 exam takers. Bold cells show the sample for column (E) of Table 4, which includes 2007–2012 exam takers in the 2009 vs. 2013 pair of ENEM adoptions years. The sample for column (F) of Table 4 includes all cells in this table. The bottom three rows show totals for each sample by exam year.

Table A5. Effects of ENEM adoption on private-public test score gaps in alternative samples

	(A)	(B)	(C)	(D)	(E)
Dependent variable	Main sample	Appear any year	Appear pre-ENEM	Partici-pation pre-ENEM	Partici-pation all years
Math score	0.158*	0.097	0.112*	0.158**	0.127*
	(0.079)	(0.058)	(0.060)	(0.070)	(0.069)
Language arts score	0.076***	0.048***	0.054***	0.088***	0.072***
	(0.026)	(0.017)	(0.013)	(0.028)	(0.024)
Natural science score	0.065*	0.047	0.046	0.102**	0.080**
	(0.034)	(0.032)	(0.030)	(0.036)	(0.035)
Social science score	0.081***	0.066**	0.069***	0.119**	0.103**
	(0.023)	(0.026)	(0.025)	(0.044)	(0.043)
Average score (core subjects)	0.110**	0.074**	0.081**	0.135***	0.110**
	(0.040)	(0.035)	(0.032)	(0.043)	(0.040)
Writing score	0.102*	0.124**	0.129**	0.037	0.006
	(0.058)	(0.048)	(0.049)	(0.105)	(0.119)
N (# exam takers)	2,512,214	10,991,098	6,774,892	856,165	718,748
Appear in INEP:	All years	Any year	At least 2005–2008	At least 2005–2008	All years
Participation rate:				$\geq$ 50% in 2005–2008	$\geq$ 50% in all years

Notes: This table examines the robustness of our results on private/public test score gaps in alternative high school samples. The dependent variables are ENEM subject scores in SD units. "Average score (core subjects)" is the average score across math, language arts, natural science, and social science. Each column estimates β^{gap} coefficients from equation (2) with a different underlying sample of high schools.

Column (A) replicates the estimates in the main sample, which consists of high schools that appeared in the INEP report in each year from 2005 to 2015. Column (B) relaxes the selection criterion and includes high schools that appeared at least once in the INEP report from 2005 to 2015. Column (C) includes high schools that appeared every year in the pre-ENEM period, i.e., from 2005 to 2008. In addition to the requirements in column (C), column (D) requires a participation rate of over 50% every year in the pre-ENEM period, meaning that at least 50% of the seniors in the high school took the ENEM in those years. Finally, column (E) requires that the high schools appeared in the INEP report each year from 2005 to 2015 and maintained a participation rate of over 50% in all years.

Note that since INEP requires a minimum participation rate for the high schools to be included in the report in certain years, columns (A)-(C) have this innate participation rate requirement. The requirements in columns (D) and (E) are in addition to the innate participation rate requirement of INEP. Details on the INEP requirement on participation rate can be found in Appendix C.3.

Parentheses contain standard errors clustered at the state level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A6. Effects of ENEM adoption on demographic test score gaps

Dependent variable	(A) Private/ public school	(B) White/ non-white	(C) Male/ female	(D) Mother college/ no college	(E) Father college/ no college	(F) High/low income
Math score	0.158*	0.069*	−0.017	0.111*	0.082	0.125*
	(0.079)	(0.038)	(0.016)	(0.058)	(0.060)	(0.067)
Language arts score	0.076***	0.027	−0.007	0.047	0.011	0.050*
	(0.026)	(0.021)	(0.012)	(0.034)	(0.034)	(0.025)
Natural science score	0.065*	0.041*	−0.019*	0.039	−0.001	0.055**
	(0.034)	(0.023)	(0.011)	(0.031)	(0.033)	(0.025)
Social science score	0.081***	0.060***	−0.017	0.067**	0.022	0.073***
	(0.023)	(0.020)	(0.011)	(0.030)	(0.028)	(0.024)
Average score (core subjects)	0.110**	0.057**	−0.017	0.076*	0.033	0.088**
	(0.040)	(0.024)	(0.012)	(0.041)	(0.039)	(0.038)
Writing score	0.102*	0.005	−0.014	0.052	0.066	0.061*
	(0.058)	(0.033)	(0.014)	(0.039)	(0.045)	(0.032)
N (# exam takers)	2,512,214	2,387,052	2,512,214	2,489,743	2,489,191	2,487,270
Low-stakes average score gap	1.229	0.477	0.271	0.882	1.027	0.892

Notes: This table examines the effects of ENEM adoption on test score gaps between students from different demographic groups, as listed in the column headers. Column (B) defines "non-white" to include black, brown, and indigenous students; this regression excludes Asian (*amarelo*) students. Column (D) and (E) defines "college" as having a college or post-graduate degree. Column (F) defines a student as "high-income" if his/her reported family income was greater or equal to two times the minimum wage in the year of the exam.

The dependent variables are ENEM subject scores in SD units. "Average score (core subjects)" is the average score across math, language arts, natural science, and social science. The sample includes ENEM test takers in our high school senior sample (column C of Table 1). Some columns have slightly smaller sample sizes due to missing values of demographic variables.

The table displays β^{gap} coefficients from versions of equation (2) estimated at the individual level. Column (A) replicates our benchmark β^{gap} coefficients from column (E) of Table 3. Columns (B)–(F) display β^{gap} coefficients from a version of equation (2) in which we replace the Private_h dummy with an indicator for being in the more advantaged demographic group. The bottom row shows the gap in the average score (core subjects) in exam cohorts prior to each state's ENEM adoption year, i.e., cohorts with $\text{HighStakes}_{st} = 0$.

Parentheses contain standard errors clustered at the state level.

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

Table A7. Effects of ENEM adoption on score gaps by test prep activity

Dependent variable	(A) Private/ public school	(B) Prep/ public school	(C) Prep/ other private school	(D) Took/didn't take a prep course
Math score	0.158*	0.202**	0.052	0.215**
	(0.079)	(0.094)	(0.052)	(0.080)
Language arts score	0.076***	0.115**	0.040	0.194**
	(0.026)	(0.045)	(0.031)	(0.083)
Natural science score	0.065*	0.134***	0.075**	0.185**
	(0.034)	(0.043)	(0.028)	(0.076)
Social science score	0.081***	0.172***	0.097***	0.212**
	(0.023)	(0.030)	(0.030)	(0.089)
Average score (core subjects)	0.110**	0.180***	0.076**	0.232**
	(0.040)	(0.051)	(0.034)	(0.093)
Writing score	0.102*	0.244	0.154	0.028
	(0.058)	(0.153)	(0.124)	(0.045)
N (# exam takers)	2,512,214	1,779,119	807,293	500,938
Low-stakes average score gap	1.229	1.147	−0.087	0.371

Notes: This table examines the effects of ENEM adoption on test score gaps between students who did/did not engage in test prep activity. We define two different measures of students' test prep activity.

In columns (B)–(C), we define a set of "prep schools" whose curriculum is specifically focused on preparation for college admission exams. For this, we obtained lists of schools that use test-oriented curricula from the websites of four prominent test prep companies: *Sistema Anglo*, *Sistema pH*, *Elite Rede de Ensino*, and *Curso Objetivo*. These lists provided the names and street addresses of each test-oriented school. We then obtained names and addresses for each school in our sample from INEP's administrative school catalog data. We geocoded the addresses in both datasets using the `ggmap` package for *R*. Lastly, we matched schools whose addresses were within 100 meters of each other and manually checked all matches using school names. We define schools that matched using this procedure as "prep schools" (all of the matched schools are private high schools). Column (B) compares prep schools to public high schools. Column (C) compares prep schools to other private high schools in our sample.

In column (D), we use a variable from the ENEM questionnaire that indicates whether individuals took an entrance exam preparation course. The question does not distinguish between courses that focused on the ENEM exam and courses that focused on other *vestibular* exams. This question is only available in the 2009–2011 and 2013–2014 cohorts, and there are missing values for many students in these cohorts. Column (D) compares students who did/did not take a prep course in a sample of individuals with non-missing values of this variable.

The dependent variables are ENEM subject scores in SD units. "Average score (core subjects)" is the average score across math, language arts, natural science, and social science. The sample includes ENEM test takers in our high school senior sample (column C of Table 1). Some columns have slightly smaller sample sizes due to missing values of demographic variables.

The table displays β^{gap} coefficients from versions of equation (2) estimated at the individual level. Column (A) replicates our benchmark β^{gap} coefficients from column (E) of Table 3. Columns (B)–(D) display β^{gap} coefficients from a version of equation (2) in which we replace the Private_h dummy with an indicator for prep schools (columns B–C) and with an indicator for individuals who took a prep course (column D). The bottom row shows the gap in the average score (core subjects) in exam cohorts prior to each state's ENEM adoption year, i.e., cohorts with $\text{HighStakes}_{st} = 0$.

Parentheses contain standard errors clustered at the state level.

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

Table A8. Robustness to controls for affirmative action adoption

	(A)	(B)	(C)	(D)
		Controls for affirmative action adoption		
Covariates	Benchmark model	Federal universities	Federal & state univ.	All universities
Panel A. Math score				
ProportionENEM$_{s(h)t}$ × Private$_h$	0.158*	0.151*	0.157*	0.158*
	(0.079)	(0.078)	(0.079)	(0.080)
ProportionAA$_{s(h)t}$ × Private$_h$		0.005	0.039	0.029
		(0.088)	(0.122)	(0.116)
Panel B. Language arts score				
ProportionENEM$_{s(h)t}$ × Private$_h$	0.076***	0.080***	0.074***	0.084***
	(0.026)	(0.023)	(0.024)	(0.026)
ProportionAA$_{s(h)t}$ × Private$_h$		−0.109**	−0.105	−0.126
		(0.051)	(0.071)	(0.076)
Panel C. Natural science score				
ProportionENEM$_{s(h)t}$ × Private$_h$	0.065*	0.062*	0.063*	0.066*
	(0.034)	(0.033)	(0.034)	(0.035)
ProportionAA$_{s(h)t}$ × Private$_h$		−0.024	−0.043	−0.016
		(0.061)	(0.085)	(0.072)
Panel D. Social science arts score				
ProportionENEM$_{s(h)t}$ × Private$_h$	0.081***	0.075***	0.080***	0.076***
	(0.023)	(0.024)	(0.022)	(0.025)
ProportionAA$_{s(h)t}$ × Private$_h$		0.029	0.036	0.091
		(0.063)	(0.089)	(0.080)
Panel E. Average score (core subjects)				
ProportionENEM$_{s(h)t}$ × Private$_h$	0.110**	0.106**	0.108**	0.111**
	(0.040)	(0.040)	(0.040)	(0.042)
ProportionAA$_{s(h)t}$ × Private$_h$		−0.029	−0.021	−0.006
		(0.067)	(0.096)	(0.085)
Panel F. Writing score				
ProportionENEM$_{s(h)t}$ × Private$_h$	0.102*	0.083	0.103*	0.086
	(0.058)	(0.052)	(0.054)	(0.054)
ProportionAA$_{s(h)t}$ × Private$_h$		0.199**	0.206*	0.285**
		(0.078)	(0.106)	(0.133)
N (# exam takers)	2,512,214	2,512,214	2,512,214	2,512,214

Notes: This table examines the robustness of our results on private/public school test score gaps to controls for affirmative action adoption. The sample includes ENEM test takers in our high school senior sample (column C of Table 1). The dependent variables are the ENEM subject scores listed in the panel titles (in SD units). Column (A) replicates our benchmark results from column (E) of Table 3, which are the β^{gap} coefficients on the interaction between ProportionENEM$_{s(h)t}$ and a dummy for private high schools, Private$_h$ from equation (2). In columns (B)–(D) we add in a measure of the adoption of affirmative action at the state × year level, ProportionAA$_{s(h)t}$, and its interaction with Private$_h$. We compute ProportionAA$_{s(h)t}$ as the proportion of all new enrollees in state $s(h)$ and year t who were admitting through reserved quotas using higher education census data. Columns (B)–(D) define ProportionAA$_{s(h)t}$ using only federal universities, federal and state universities, and all universities, respectively.

Parentheses contain standard errors clustered at the state level.

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

Table A9. Heterogeneity by topic area — Language arts, natural science, and social science
Dependent variable: Proportion correct answers

(A)	(B)	(C)	(D)	(E)	(F)
		Public students		Private/public gap	
Question group	N_q	Mean	β (SE)	Mean	β^{gap} (SE)
Panel A. Language arts					
Communication (1–4)	54	0.468	−0.003 (0.009)	0.193	0.019 (0.009)**
Foreign language (5–8)	80	0.424	−0.009 (0.013)	0.234	−0.005 (0.008)
Body language (9–11)	29	0.521	−0.018 (0.010)*	0.139	0.016 (0.013)
Art (12–14)	42	0.425	0.005 (0.012)	0.188	0.019 (0.007)**
Literary text (15–17)	63	0.342	−0.004 (0.006)	0.157	0.022 (0.006)***
Linguistics (18–20)	41	0.449	−0.003 (0.015)	0.193	0.020 (0.012)
Argumentation (21–24)	63	0.450	0.000 (0.005)	0.171	0.013 (0.006)**
Portuguese (25–27)	39	0.410	−0.011 (0.008)	0.205	0.017 (0.010)
Social communication (28–30)	33	0.455	0.001 (0.011)	0.186	0.024 (0.007)***
All coefficients equal (p value)			0.034		0.021
Panel B. Natural science					
Human constructions (1–4)	57	0.274	0.002 (0.009)	0.166	0.013 (0.006)**
Technology (5–7)	34	0.255	0.005 (0.005)	0.102	0.010 (0.004)**
Environmental conservation (8–12)	58	0.360	0.005 (0.006)	0.178	0.004 (0.005)
Ecosystems (13–16)	54	0.316	−0.004 (0.008)	0.204	0.018 (0.011)
Scientific methods (17–19)	49	0.297	−0.006 (0.010)	0.189	0.006 (0.008)
Physics (20–23)	57	0.269	−0.010 (0.008)	0.158	0.009 (0.004)*
Chemistry (24–27)	65	0.246	−0.001 (0.002)	0.127	0.006 (0.007)
Biology (28–30)	31	0.416	−0.012 (0.014)	0.192	0.018 (0.014)
All coefficients equal (p value)			0.023		0.045
Panel C. Social science					
Culture (1–5)	73	0.385	−0.011 (0.006)*	0.184	0.022 (0.007)***
Geography (6–10)	66	0.353	0.004 (0.006)	0.214	0.006 (0.006)
Social institutions (11–15)	74	0.376	−0.006 (0.008)	0.173	0.021 (0.005)***
Technology (16–20)	61	0.376	−0.010 (0.005)*	0.186	0.009 (0.007)
Citizenship (21–25)	62	0.432	0.002 (0.007)	0.192	0.015 (0.009)
Society and nature (26–30)	69	0.405	−0.001 (0.008)	0.168	0.013 (0.007)*
All coefficients equal (p value)			0.000		0.000

Notes: This table shows how the increase in ENEM stakes impacted students' performance on different topic areas of the language arts (Panel A), natural science (Panel B), and social science (Panel C) exams. This table is analogous to the math exam results in Panel B of Table 5. The sample includes 2009–2017 ENEM test takers in our high school senior sample (column C of Table 1). Regressions are at the high school (h) × year (t) × exam question (q) level. The dependent variable is the proportion of correct answers in each htq cell. We estimate regressions separately for questions in the topic areas of each subject as defined by ENEM test designers.

Column (A) defines the group of questions for each regression. Column (B) shows the number of questions in each group. Column (C) shows the mean proportion of correct answers for public school students in cohorts prior to each state's ENEM adoption year (i.e., cohorts with $\text{HighStakes}_{st} = 0$). Column (E) shows the mean private/public gap in the proportion of correct answers in those cohorts. Columns (D) and (F) display the β and β^{gap} coefficients from equation (2) estimated for each group of questions. The last row of each panel reports p values from F tests that the topic area coefficients in columns (D) or (F) are jointly equal.

Parentheses contain standard errors clustered at the state level.

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

Table A10. Heterogeneity by topics in a *Me Salva!* study guide — Math exam
Dependent variable: Proportion correct answers

(A)	(B)	(C)	(D)	(E)	(F)
		Public students		Private/public gap	
Question group	N_q	Mean	β (SE)	Mean	β^{gap} (SE)
Panel A. Study guide words matched to question text					
Not in study guide	262	0.306	−0.007 (0.011)	0.186	0.026 (0.010)**
In study guide	143	0.267	−0.003 (0.008)	0.161	0.021 (0.008)**
cone (*cone*)	8	0.471	−0.010 (0.026)	0.210	0.024 (0.030)
cube (*cubo*)	8	0.179	−0.000 (0.008)	0.182	0.029 (0.010)***
cylinder (*cilindro*)	10	0.374	−0.011 (0.013)	0.157	0.021 (0.022)
directly (*diretamente*)	8	0.284	0.009 (0.011)	0.218	0.030 (0.010)***
median (*mediana*)	12	0.226	0.009 (0.010)	0.130	0.055 (0.021)**
parallelepiped (*paralelepípedo*)	9	0.314	0.001 (0.034)	0.342	0.043 (0.028)
possibilities (*possibilidade*)	8	0.275	−0.015 (0.006)**	0.126	0.021 (0.017)
possible (*possíveis*)	10	0.310	−0.015 (0.023)	0.168	0.011 (0.022)
prism (*prisma*)	8	0.368	−0.002 (0.015)	0.136	0.014 (0.025)
probability (*probabilidade*)	25	0.234	−0.010 (0.008)	0.131	0.028 (0.010)**
pyramid (*pirâmide*)	10	0.337	−0.035 (0.018)*	0.138	0.042 (0.017)**
rectangle (*retângulo*)	10	0.277	0.006 (0.011)	0.192	0.024 (0.014)*
square (*quadrado*)	35	0.256	−0.001 (0.005)	0.197	0.009 (0.006)
triangle (*triângulo*)	8	0.231	−0.002 (0.006)	0.120	0.038 (0.010)***
Fewer than 8 occurrences	40	0.268	−0.005 (0.007)	0.146	0.016 (0.008)*
In vs. not in study guide (p value)			0.310		0.157
Panel B. Study guide concepts matched to solutions					
Not in study guide	231	0.317	−0.005 (0.010)	0.193	0.023 (0.011)**
In study guide	173	0.259	−0.004 (0.009)	0.155	0.025 (0.008)***
Geometric formulas	64	0.261	−0.004 (0.009)	0.174	0.026 (0.007)***
Proportions (“Rule of 3”)	12	0.330	−0.011 (0.018)	0.205	0.058 (0.010)***
Manipulating fractions	58	0.248	−0.004 (0.008)	0.161	0.029 (0.009)***
Radicals	13	0.195	0.009 (0.006)*	0.139	0.024 (0.014)*
Combinatory and statistical analysis	25	0.224	−0.003 (0.004)	0.103	0.024 (0.015)
Probability	36	0.255	−0.010 (0.010)	0.138	0.025 (0.007)***
Trigonometric formulas	8	0.231	0.006 (0.006)	0.104	0.017 (0.007)**
In vs. not in study guide (p value)			0.791		0.591

Notes: This table shows how the increase in ENEM stakes impacted students' performance on math questions that are covered in a study guide by the test prep company *Me Salva!*. The sample includes 2009–2017 ENEM exam takers in our high school senior sample (column C of Table 1). Regressions are at the high school (h) × year (t) × exam question (q) level. The dependent variable is the proportion of correct answers in each htq cell. Panel A defines question groups based on whether the question text contains key words from the *Me Salva!* study guide, restricting to words that appear in 8+ questions. Panel B defines question groups based on the whether the text of the solution (prepared by another test prep company, *Descomplica*) requires concepts from the *Me Salva!* study guide. See Appendix C.4 for details on these data sources and our match between *Me Salva!* topics and ENEM questions.

Column (A) defines the group of questions for each regression. Column (B) shows the number of questions in each group. Column (C) shows the mean proportion of correct answers for public school students in cohorts prior to each state's ENEM adoption year (i.e., cohorts with $\text{HighStakes}_{st} = 0$). Column (E) shows the mean private/public gap in the proportion of correct answers in those cohorts. Columns (D) and (F) display the β and β^{gap} coefficients from equation (2) estimated for each group of questions. In both panels, the last row reports p values from an F test that the coefficients in the first and second rows are equal.

Parentheses contain standard errors clustered at the state level.

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

Table A11. OLS relationships between the informativeness and preppability of ENEM competencies with controls for question difficulty

	(A)	(B)	(C)	(D)
		Change in return to a correct answer from 1pp increase in β^{gap}		
Dependent variable	Mean return to correct answer	Bivariate regression	Controls for % correct	IRT controls
Panel A. Outcomes for all exam takers				
Enrolled in any college by 2019	0.1011	0.0043^{**}	0.0008	0.0019
		(0.0022)	(0.0007)	(0.0013)
Finished college within 5 years of ENEM	0.0294	0.0014	0.0006^{**}	0.0007^{*}
		(0.0009)	(0.0003)	(0.0004)
Earned a college degree by 2019	0.0707	0.0031^{*}	0.0018^{**}	0.0020^{*}
		(0.0016)	(0.0008)	(0.0010)
Appears in RAIS in 2016–2018	−0.0162	0.0012^{*}	0.0012	0.0001
		(0.0007)	(0.0007)	(0.0007)
N (# competencies)	120	120	120	120
Panel B. Outcomes for college enrollees				
Persisted in college for 1 year	0.0123	0.0005^{*}	0.0000	0.0002
		(0.0002)	(0.0001)	(0.0002)
Persisted in college for 3 years	0.0472	0.0016^{**}	0.0004	0.0010^{*}
		(0.0008)	(0.0003)	(0.0006)
Completed program within 5 years	0.0225	0.0012	0.0009^{**}	0.0009^{*}
		(0.0008)	(0.0004)	(0.0005)
Fraction of college credits completed	0.0481	0.0019^{**}	0.0011^{***}	0.0013^{**}
		(0.0009)	(0.0004)	(0.0006)
N (# competencies)	120	120	120	120
Panel C. Outcome for individuals in RAIS				
Hourly wage (BRL)	13.2852	0.5729^{**}	0.6169^{**}	0.5039^{*}
		(0.2846)	(0.2853)	(0.2884)
Log hourly wage	0.1534	0.0076^{**}	0.0067^{**}	0.0058^{**}
		(0.0030)	(0.0028)	(0.0028)
N (# competencies)	120	120	120	120

Notes: This table presents OLS relationships between the informativeness and preppability of ENEM competencies with controls for question difficulty. Our measure of informativeness is the average return to a correct answer at the competency level defined using the outcomes in the column header. Our measure of preppability is the β^{gap} coefficient from equation (2) estimated separately for groups of questions in each competency. See Table 7 for details.

Column (A) shows the mean return to a correct answer averaged across all subjects and competencies. Column (B) reports OLS coefficients from bivariate regressions of the mean returns to a correct answer on the β^{gap} coefficients using all subjects and competencies, which replicates the results in column (F) of Table 7. Column (C) is similar to column (B), but we include a quadratic in the mean proportion of correct answers in each competency interacted with subject dummies. Column (D) is similar to column (B), but we include a quadratic in the mean IRT parameters in each competency interacted with subject dummies. We normalize the OLS coefficients to represent a 1pp increase in β^{gap}. Parentheses contain robust standard errors.

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

B. Theoretical appendix

This theoretical appendix presents a simple framework that helps to interpret our main empirical results and shed light on the potential mechanisms.

B.1. **Framework setup.** We consider a population of exam takers that are characterized by their socioeconomic status (SES) and their abilities. Let X_i denote an observable measure of individual i's SES, e.g., an indicator for attending a private high school. We let a_i denote individual i's ability for performing well on a college admission exam, which is not directly observable in data. We refer to a_i as *test ability* to emphasize that it may be distinct from other abilities that help the individual perform well in college and in the labor market.

We assume that the stakes of the college admission exam vary across cohorts, and that test takers in high-stakes cohorts engage in more test prep. For simplicity, we suppose that individuals are randomly assigned to either a low-stakes or a high-stakes exam cohort, and we let H_i be a binary indicator for the high-stakes cohort.[43] Individuals have a stronger incentive to perform well in the high-stakes exam cohort, and so we assume that these individuals engage in additional test prep. This additional prep increases their test ability by an amount that we denote by e_i. We interpret e_i broadly; for example, it may include studying test prep books, taking preparatory courses, focusing more intently during the school year, or exerting more effort on the exam.

Thus a test taker's preparedness for the exam, as a function of their cohort, is given by:

$$\theta_i = a_i + H_i e_i. \tag{B1}$$

where we refer to θ_i as *test skill*. In other words, individuals in low-stakes cohorts have test skill $\theta_i = a_i$, while individuals in high-stakes cohorts have test skill $\theta_i = a_i + e_i$.

An individual's score on the college admission exam is a noisy measure of their test skill. We let T_i denote individual i's test score, which is observable in the data. We assume test scores are given by:

$$T_i = \theta_i + \epsilon_i^T, \tag{B2}$$

where ϵ_i^T is random noise that reflects variation in test performance due to factors like guessing and health on exam day.

Our interest is in the predictive power of the test score for measures of college success that matter to both individuals and colleges. We let Y_i denote an observable measure of college success, e.g., persisting in college after enrolling or completing a college degree. We assume Y_i is given by:

$$Y_i = \alpha a_i + \beta H_i e_i + \gamma X_i + v_i + \epsilon_i^Y. \tag{B3}$$

[43] Our empirical analysis relies on a parallel trends version of this assumption.

We allow college success to potentially depend on test ability, a_i, test prep, e_i, and SES, X_i. The parameter α represents the effect of test ability on college success; it is natural to assume $\alpha > 0$ since individuals with high test scores tend to perform better in college. Similarly, an individual's SES may help them succeed in college above and beyond their test ability, and so we assume $\gamma > 0$. The parameter β allows for the possibility that test prep directly affects college performance, although, as we discuss below, the sign of this parameter is less clear. In addition to these three factors, we allow college success to depend on other abilities that are unrelated to test skill, which we denote by v_i, and a random noise term, ϵ_i^Y.

B.2. **Effects of exam stakes on test score gaps.** In Section 4, we find that increasing the stakes of a college admission exam increases test score gaps between high- and low-SES students. In our framework, this result can be written as:

$$\text{cov}(T_i, X_i | H_i = 1) > \text{cov}(T_i, X_i | H_i = 0). \tag{B4}$$

In other words, the covariance between test scores, T_i, and SES, X_i, is larger in the high-stakes cohort ($H_i = 1$) than in the low-stakes cohort ($H_i = 0$).

Plugging in equations (B1)–(B2) and simplifying, expression (B4) can also be written as:

$$\text{cov}(X_i, e_i) > 0. \tag{B5}$$

Expression (B5) states that high SES students engage in more test prep than low SES students when the stakes of the exam increase.

B.3. **Effects of exam stakes on predictive validity.** In Section 5, we examine how an increase in stakes impacts the predictive power of exam scores for college success, as well as the potential mechanisms for this effect.

In Table 6, we find that scores from higher stakes exams have more predictive power for various measures of college success. In our framework, this result can be written as:

$$\text{cov}(T_i, Y_i | H_i = 1) > \text{cov}(T_i, Y_i | H_i = 0). \tag{B6}$$

Using equations (B1)–(B3), expression (B6) simplifies to:

$$\gamma\text{cov}(X_i, e_i) + (\alpha + \beta)\text{cov}(a_i, e_i) + \text{cov}(v_i, e_i) + \beta\text{var}(e_i) > 0. \tag{B7}$$

Expression (B7) shows that there are three broad channels through which additional incentives for test prep could increase the predictive validity of exam scores:

(1) **SES correlation:** $\gamma\text{cov}(X_i, e_i) > 0$. First, the high-stakes exam scores may be more predictive of college outcomes simply because they are more correlated with SES, which also tends to benefit students in college. For example, wealthy students may have greater access to test prep services, and family wealth also may help students

succeed in college. Since we find that $\text{cov}(X_i, e_i) > 0$, then all else equal, one would expect that higher-stakes exam scores would have more predictive validity.

(2) **Ability correlation:** $(\alpha + \beta)\text{cov}(a_i, e_i) + \text{cov}(v_i, e_i) > 0$. Second, the predictive validity of scores may increase if the induced test prep is correlated with ability. Students who have higher test ability on the low-stakes exam may engage in more test prep when the stakes increase, i.e., $\text{cov}(a_i, e_i) > 0$. Additionally, students with higher ability to succeed in college ability may engage in more test prep for the higher-stakes exam, i.e., $\text{cov}(v_i, e_i) > 0$. The sign of these covariance terms is *a priori* less clear. On the one hand, students who know that they are likely to succeed in college may have the strongest incentives to prep for the higher-stakes exam. On the other hand, test prep may disproportionately benefit high-income but lower-ability students who are unlikely to be admitted to top colleges without prepping. Thus we do not have a strong prior on whether an ability correlation plays a role in the observed increase in exam score validity.

(3) **Skill accumulation:** $\beta\text{var}(e_i) > 0$. Finally, test prep may be directly beneficial for college success. This channel is operative if the test skills that individuals accumulate from prepping, e_i, directly improve their college outcomes, Y_i. In our framework, this skill accumulation channel exists if $\beta > 0$. High-stakes exams are often criticized for creating incentives to engage in test-oriented learning that is not useful outside the exam. Thus it is possible that $\beta = 0$. It is also possible that $\beta < 0$ if test prep crowds out other useful learning.

There is a fourth potential channel that is outside the scope of our framework: high-stakes exam scores may have more predictive validity if they increase the "match quality" between individuals and college programs. This channel arises because an individual's test score, T_i, may have a causal effect on their outcome, Y_i, through its influence on which college and/or major they attend. Thus high-stakes exam scores may be more informative for college success if the distribution of these scores leads to better student/college matches, e.g., on the basis of academic preparation. To distinguish between this channel and the three mentioned above, we follow the standard practice that testing agencies use to measure predictive validity. Specifically, in addition to estimating raw correlations between test scores, T_i, and outcomes, Y_i, we also estimate correlations after de-meaning each variable *within* college programs.

Distinguishing between the correlational and skill accumulation channels is challenging because of unobserved abilities that impact both test scores and college success (i.e., a_i and v_i). But regardless of which channel is at play, our finding that test prep increases the informativeness of scores can explain why many colleges around the world use high-stakes tests for admissions.

C. Empirical appendix

C.1. **Variable definitions.** This section describes the main variables in our paper.

C.1.1. *Test scores.*

- **Subject scores.** The post-2009 ENEM scores, as reported to the public, are scaled to have a mean of 500 and an SD of 100 in the population of 2009 high school seniors who took the exam. Throughout the paper, we report ENEM scores in SD units relative to this population. For ENEM scores in math, language arts, natural science, and social science, our transformation is:

$$\text{Transformed subject score} = \frac{\text{Raw subject score} - 500}{100},$$

  After transformation, a score of zero in our paper is equivalent to the performance of the average high school senior who took the ENEM in 2009, and a score of one is 1 SD higher within this population. These transformations preserve the comparability of test scores across cohorts.

  The 2007–2008 ENEM reported only a single core-component score plus a writing score. To define scores for each subject, we first categorize the multiple choice questions into math, language arts, natural science, and social science, and then compute a separate score for each subject using the IRT parameters estimated from the response data.

  Since the reference populations differ for the 2007–2008 and 2009–2017 exams, in regressions where we use scores from both periods, we standardize the scores to have mean 0 and SD 1 within each year of our sample.
- **Average scores (core subjects).** The post-2009 average scores are calculated by taking the average of four subject scores, and then standardize relative to the reference population. In practice, our transformation is:

$$\text{Transformed average score} = \frac{\text{Raw average score} - 500}{86.7},$$

  where 86.7 is the SD of the average score in the reference population. After transformation, a score of zero in our paper is equivalent to the performance of the average high school senior who took the ENEM in 2009, and a score of one is 1 SD higher within this population.

  For the 2007–2008 ENEM, the average score is defined as the single core-component score. In regressions where we use scores from both the 2007–2008 and 2009–2017 exams, we standardize the scores to have mean 0 and SD 1 within each year of our sample.
- **Writing scores.** The post-2009 writing score is also standardize relative to the reference population. In practice, our transformation is:

$$\text{Transformed writing score} = \frac{\text{Raw writing score} - 597}{137},$$

In regressions where we use scores from both the 2007–2008 and 2009–2017 exams, we standardize the writing scores to have mean 0 and SD 1 within each year of our sample.

C.1.2. *Exam-taker characteristics.* These variables were collected from a survey that applicants completed as part of the ENEM exam process.

- **Race.** In Brazil, race is commonly classified in five groups: *branco* (white), *pardo* (brown), *preto* (Black), *amarelo* (yellow), and indigenous. Since Asian and indigenous people represent a small proportion of the population in Brazil (less than 3 percent in our sample), we use indicators for three major racial groups: *branco* (white), *pardo* (brown), and *preto* (Black). We set the indicator variables to missing if the students declined to declare their racial identities (2 percent of the students in our data).
- **Parental education.** The measures for mother's and father's education consist of 8 categories from "none" to "post-graduate". From these categorical variables, we derive two indicator variables "Mother attended college" and "Father attended college", which equal 1 if the respective parent achieves an educational level of "college" or "post-graduate".
- **Family income.** Family income is measured as multiples of the minimum wage in the year of the exam. We define an indicator variable "Family income > 2x min. wage" (or "High-income"), which equals 1 if the reported income is more than twice the minimum wage.
- **Private high school.** Throughout our analysis, we define "private high schools" to include both private and federal high schools (0.5 percent of students) since their students are comparable in terms of socioeconomic status and achievement. In contrast, "public high schools" include both state and municipality high schools.

C.1.3. *College and labor market outcomes.* These variables were collected from INEP's higher education census (*Censo da Educação Superior*) for the years 2010–2019 (INEP, 2022) and Brazil's employee-employer dataset, the RAIS (*Relação Anual de Informações Sociais*), for the years 2016–2018 (RAIS, 2022).

- **College enrollment.** We define an indicator variable "Ever enrolled in college" which equals 1 if the student has a record in the INEP's higher education census for the years 2010–2019.
- **College persistence.** The higher education census contains information on a student's enrollment year and last year on record. We define an indicator variable "Persisted in college for 3 years" which equals 1 if the student's last year on record is greater or equal to 3 years after the enrollment year.
- **College graduation.** The higher education census contains information on a student's enrollment and graduation year. We define an indicator variable "Ever graduated college" which equals 1 if the student has a non-missing graduation year. We define an indicator

variable “Graduated college within 5 years” which equals 1 if the student’s graudation year is within 5 years of the enrollment year.

- **Fraction of college credits completed.** The higher education census contains information on a student’s credits completed in his/her program and the total credits required in the program curriculum. We define the variable “Fraction of college credits completed” as the credits completed on the student’s last record in the census divided by the total credits in the program.
- **Appears in RAIS.** “Appears in RAIS” is an indicator variable that equals 1 if a given student has a matching record in RAIS, which indicates that the student has engaged in formal employment.
- **Hourly wage (BRL).** We compute an individual’s hourly wage as average monthly earnings divided by average monthly contracted hours, both available in RAIS. The wages are expressed in terms of Brazilian Reals.

C.2. **Data and merging.** Our base dataset contains all individuals who took the ENEM exam in 2007–2017. This dataset includes 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 student responses to each exam question, the question subject, and skill tested. From this dataset, we only keep high-school students with a valid score (i.e., non-zero and non-missing) on each subject test. These restrictions exclude, for example, individuals who took the exam after graduating from high school or who missed one of the testing days.

To measure long-run outcomes, we combine the 2009–2014 ENEM records with two individual-level administrative datasets using individuals’ national ID numbers (*Cadastro de Pessoas Físicas*). The linkage was conducted in the secured data room at the INEP facilities in Brasilia, Brazil. We exclude students with missing national ID (0.04 percent) and those who took the exam more than three times in our data (0.10 percent).

We measure college outcomes using Brazil’s higher-education census from 2010–2019. This dataset offers comprehensive information about all college enrollees, including their university of enrollment, major, the academic year when they enrolled, and their year of graduation. 65.6 percent of high-school seniors taking the ENEM during 2009–2014 appear in the census data.

We measure labor-market outcomes using an administrative employee-employer matched dataset called RAIS (*Relação Anual de Informações Sociais*) from 2016–2018. The RAIS contains data on workers employed in the formal sector. It does not include data on individuals working within the informal sector, those who are self-employed, or individuals who are currently unemployed. This dataset includes worker-level and firm-level information. The worker-level data includes educational attaintment, occupation, and earnings. The firm-level data includes total employee count, the industry they operate within, and their geographical location. 32.9 percent of high-school seniors taking the ENEM during 2009–2014 appear in the RAIS. The relatively low match rate might be attributed to the fact that some individuals could still be enrolled in college. However, even when considering individuals who took the ENEM in 2009, the match rate remains comparably low with only 31.7 percent of them being matched to the RAIS.

C.3. **Sample definition.** This section describes criteria needed for high schools to be included in the INEP annual reports we leverage to create our sample.

As noted in the main text, our *high school graduate sample* consists of the set of high schools that were in all the yearly ENEM-performance reports created by the INEP during 2005-2015. Only high schools that met two conditions were included in the annual performance reports. First, the high school needed to have at least 10 test-takers who declared that they would graduate that year. Before 2011, all individuals taking the ENEM were recorded as test-takers for the calculation of the participation rate, regardless of whether they completed the test or not. Since 2011, to be considered a test-taker, an individual has to complete the four subject tests plus the writing essay, and obtain a non-zero score in all subjects.[44] Second, starting in 2009, the INEP required a minimum *participation rate* to be included in the report. The participation rate is the total number of ENEM test-takers in the high school divided by the number of enrolled students in the final years of high school, based on the records of the High School Census. Between 2009 and 2010, the minimum participation to be included in the report rate was 2 percent. In 2011, the minimum participation threshold increased to 50 percent.

At the schools in our sample, the average ENEM participation rate over the 2005–2015 period was 70 percent. Table 2 presents balance tests that show that the number of exam takers in our high school graduate sample and the characteristics of these exam takers did not change significant when the stakes of the ENEM increased.

C.4. **Categorization of math questions.** This subsection provides details on the categorization of math questions that we use for the heterogeneity analyses in Table 5 and Appendix Table A10.

[44] Only the following grades are considered in the report: 3rd and 4th grade of regular high school (*ensino médio regular 3a e 4a série*), 3rd and 4th grade of teaching track high school *(ensino médio magistério 3a e 4a série*), non-serialized regular and teaching track high school (*ensino médio não-seriado, regular e magistério*), vocational education and high school for youth and adult education (*educação profissionalizante e ensino médio para educação de jovens e adultos*).

Table C1. Topic areas and competencies for ENEM math exam

(A) Topic area	(B) Competency (and reference number)	(C) Topic area	(D) Competency (and reference number)
Numbers	Recognize numbers (1)	Proportions	Solve problems using proportions (16)
Numbers	Identify numerical patterns (2)	Proportions	Use proportions to construct arguments (17)
Numbers	Solve problems using numbers (3)	Proportions	Evaluate interventions using proportions (18)
Numbers	Use numbers to construct arguments (4)	Algebra	Identify algebraic relationships (19)
Numbers	Evaluate interventions using numbers (5)	Algebra	Interpret Cartesian graphs (20)
Geometry	Project 3D objects into 2D space (6)	Algebra	Solve problems using algebra (21)
Geometry	Identify geometric shapes (7)	Algebra	Use algebra to construct arguments (22)
Geometry	Solve problems using geometry (8)	Algebra	Evaluate interventions using algebra (23)
Geometry	Use geometry to construct arguments (9)	Interpreting data	Make inferences using data in tables/graphs (24)
Measurements	Identify units of measurement (10)	Interpreting data	Solve problems using data in tables/graphs (25)
Measurements	Use scales in everyday situations (11)	Interpreting data	Use tables/graphs to construct arguments (26)
Measurements	Solve problems using magnitudes (12)	Statistics	Calculate statistical quantities from data (27)
Measurements	Use measurements to construct arguments (13)	Statistics	Solve problems using statistics (28)
Measurements	Evaluate interventions using measurements (14)	Statistics	Use statistics to construct arguments (29)
Proportions	Identify proportional relationships (15)	Statistics	Evaluate interventions using statistics (30)

Notes: This table shows the 7 topic areas (columns A and C) and 30 competencies (columns B and D) for the ENEM math exam. Labels are translated and shortened by the authors from the descriptions in `Matriz_Referencia_Enem.pdf`, which is included with the microdata.

In Table 5, we categorize questions into 7 *topic areas* (Panel B) and 30 *competencies* (Panel C) defined by ENEM test designers. These topics areas and competencies are in the `ITENS_PROVA_****.TXT` files of the microdata. The labels for each topic area and competency are defined in `Matriz_Referencia_Enem.pdf`, which is included with the microdata and is also available online at the link in this footnote.[45] Table C1 shows the translated and shortened labels that we use for Table 5. We also present results by topic area for language arts, natural science, and social science in Appendix Table A9.

In Appendix Table A10, we define groups of questions based on whether the questions are related to topics covered in a study guide created by *Me Salva!*, which is a well-known Brazilian test prep company. The study guide is called *The Approved Book: One topic per day to pass the ENEM*.[46] In Panel A of Appendix Table A10, we match key words from the *Me Salva!* study guide to the text of each question. In Panel B of Appendix Table A10, we match concepts from the *Me Salva!* study guide to solutions to each question that were created by another well-known test prep company called *Descomplica*. *Descomplica* creates solutions to each question from previous ENEM exams and makes them publicly-available online to help students prepare.[47]

[45] See: https://download.inep.gov.br/download/enem/matriz_referencia.pdf (accessed in June 2023).
[46] The Portuguese title is *O Livro do Aprovado: Um conteúdo por dia para passar no ENEM*. See: https://cdn.mesalva.com/uploads/medium/attachment/MS2018-livro-do-aprovado.pdf (accessed in June 2023).
[47] See: https://descomplica.com.br/gabarito-enem/questoes/?cor=azul (accessed in June 2023). One question is behind a paywall—Question 145 in the *Azul* book of the 2015 math exam)—and thus we exclude it from our analysis in Panel B of Table A10).

We match key words and concepts in the *Me Salva!* study guide to ENEM questions using text analysis. The *Me Salva!* study guide contains tips for solving questions in seven different content areas, and in each content area there are key words that appear in bold in the text. For Panel A of Appendix Table A10, we search the question text for the key words from the study guide. For Panel B of Appendix Table A10, we search the *Descomplica* solutions for both the key words and for regular expressions that indicate questions in which the solution is likely to depend on concepts from the study guide. The content areas and search terms that we use are as follows; the key words that we use in both Panels A and B appear in italics, and the regular expression searches that we use in Panel B appear in plain text:

- **Geometric formulas.** *cilindro, cone, cubo, equilatero, esfera, hexagonal, hexagono, losango, paralelepipedo, piramide, prisma, quadrado, quadrangular, retangulo, trapezio, triangular, triangulo.*
- **Proportions ("Rule of 3").** *diretamente, grandezas, inversamente, proporcionais, proporção, regra de três.*
- **Manipulating fractions.** Solutions that contain at least two fractions ([0-9]+/[0-9]+) and also an equals sign (=).
- **Radicals.** Solutions that contain a square or cube root sign ($\sqrt{}$ or $\sqrt[3]{}$).
- **Combinatory and statistical analysis.** *arranjo, combinação, mediana, moda, permutação, possibilidades.* Solutions that contain arrangement or combination notation (A[0-9],[0-9] or C[0-9],[0-9]).
- **Probability.** *combinação, possiveis, probabilidade.* Solutions that contain combination notation (C[0-9],[0-9]).
- **Trigonometric formulas.** *cos, cossec, cosseno, cotg, sec, sen, seno, tangente, tg.*

In all cases we trim plural, adjective, and masculine/feminine endings to words before matching. In Panel A, we display results for key words that appear in eight or more questions, and we group all other words into the "Fewer than 8 occurrences" category. In Panel B, we display results separately for each of the seven *Me Salva!* content areas. In both panels, we also show results pooling across all questions that do/don't match any search term in the study guide.