NBER WORKING PAPER SERIES

STAKES AND SIGNALS: AN EMPIRICAL INVESTIGATION OF MUDDLED INFORMATION IN STANDARDIZED TESTING

Germán J. Reyes
Evan Riehl
Ruqing Xu

Working Paper 32608
http://www.nber.org/papers/w32608

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2024

For helpful comments we thank Serena Canaan, William Dougan, Deborah Doukas, Jorge Luis García, Daniel Hamermesh, Navin Kartik, Pierre Mouganie, Eric Nielsen, Roberta Olivieri, Christiane Szerman, Christopher Walters, Russell Weinstein, and participants at various seminars and conferences. Adya Bhargava provided excellent research assistance. All errors are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Stakes and Signals: An Empirical Investigation of Muddled Information in Standardized Testing
Germán J. Reyes, Evan Riehl, and Ruqing Xu
NBER Working Paper No. 32608
June 2024
JEL No. I23,I24,J24,M5

ABSTRACT

We examine a natural experiment in Brazil in which similar students took the same standardized test as either a low-stakes school accountability exam or a high-stakes admission exam for the country's top universities. Using administrative data and a difference-in-differences design, we find that test score gaps between high- and low-income students expanded on the high-stakes exam, consistent with wealthy students engaging in test prep. Yet the increase in stakes made scores more informative for students' college outcomes. Thus the "muddling" of information on natural ability and test prep improved the quality of the score signal, although it also exacerbated inequality.

Germán J. Reyes
Middlebury College
1229 S Street Extension
Middlebury, VT 05753
greyes@middlebury.edu

Ruqing Xu
Cornell University
rx24@cornell.edu

Evan Riehl
Department of Economics
Cornell University
266 Ives Hall
Ithaca, NY 14853
and NBER
eriehl@cornell.edu

A data appendix is available at http://www.nber.org/data-appendix/w32608

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. Social scientists have long hypothesized that strategic manipulation erodes the reliability of signals, an idea that is known as Goodhart's Law (Goodhart, 1975) or Campbell's Law (Campbell, 1979).[2] 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.[3] But it is theoretically ambiguous 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 test that measured high school quality 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 test scores for individuals' academic potential.

It is *ex ante* unclear how test score inequality and informativeness change as the stakes of an exam increase. 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 common belief among education researchers

[1] Influential books such as *The Tyranny of Metrics* (Muller, 2018) have popularized the notion that agents strategically manipulate signals of their quality in various high-stakes settings.

[2] Goodhart's Law is named after economist Charles Goodhart, who wrote: "When a measure becomes a target, it ceases to be a good measure" (Goodhart, 1975). Campbell's Law is named after psychologist Donald 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).

[3] In this literature, gaming ability refers to the marginal cost of improving the signal, while natural action is the signal individuals send without stakes.

that low-stakes test scores are a better measure of student learning than high-stakes scores because there is less of an incentive to manipulate performance (e.g., Amrein and Berliner, 2002). Further, critics often argue that the use of high-stakes college admission exams helps wealthy students game the system through expensive test prep and other score-boosting strategies (Buchmann et al., 2010; Soares, 2015).[4] Thus an increase in exam stakes could widen demographic test score gaps and also 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 stakes could reduce socioeconomic test score gaps while also making exam scores more informative about college potential.

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 to measure school quality, 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 some ENEM participants took a low-stakes (from their perspective) school quality exam, while others took a high-stakes test that governed admission to the most selective universities in their state. We define a sample of high school seniors who were likely to take the ENEM 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,

[4] 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.

We have two main findings. First, test score gaps between advantaged and disadvantaged students widened on the higher stakes exam. 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. Racial and other socioeconomic test score gaps expanded by a similar percentage. The magnitude of our estimates implies that there was a significant increase in the selectivity of the university programs that private school students could gain access 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. The predictive power of scores increased both overall and measured among students who attended the same college programs, which shows that our findings are driven by an increase in the informativeness of scores rather than by a causal impact of the scores on the programs students attended. Our results on the informativeness of ENEM scores for earnings are inconclusive because our labor market data is measured too early in students' careers.

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 exam preparation. But our question-level analysis shows that private students performed better across a wide range of exam skills that are aligned with 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 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, private students' score gains reflected a broad set of skills that are informative for academic potential.

Our paper provides empirical evidence that informs theoretical work on how gaming impacts signals of performance. Many signaling models assume that the principal's goal is to minimize the agent's ability to manipulate their signal (e.g., Holmstrom and Milgrom, 1991; Ederer et al., 2018; Perez-Richet and Skreta, 2022; Ball, 2024). Models of "muddled information" (Fischer and Verrecchia, 2000; Frankel and Kartik, 2019, 2022) show that signals become relatively more informative about an agent's *gaming ability* when the stakes of the

interaction increase. But as Frankel and Kartik (2019) acknowledge, it is unclear whether gaming makes signals more or less informative for outcomes of interest to observers. We show that the influence of gaming ability on standardized tests makes the resulting scores more informative for academic potential. This challenges the common assumption that the gaming of college admission exams necessarily distorts or adds noise to the signal (Krishna et al., 2022; Lee and Suen, 2023). Our findings suggest that gaming ability can instead reflect beneficial characteristic like learning capacity or non-cognitive skills.

Relatedly, our paper is unique within research in the economics of education in showing that higher stakes can increase the informativeness of exams. A large empirical literature finds that educators strategically respond to high-stakes testing by teaching narrowly-defined skills (Jones et al., 1999; Jacob, 2005; Cohodes, 2016), manipulating the test-taking population (Figlio and Getzler, 2006; Cullen and Reback, 2006), prioritizing specific subjects or students (Neal and Schanzenbach, 2010; Reback et al., 2014), or even resorting to outright cheating (Jacob and Levitt, 2003). Related work finds that gains in high-stakes test scores from accountability policies do not always translate into 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 ours is the first paper in this literature to use data on longer-run outcomes to directly test how the informativeness of scores varies between high- and low-stakes exams. We show that, in the context of college admissions, higher stakes tests can provide a better signal of students' academic potential.

Lastly, our paper shows that there is a tradeoff between equality and informativeness in the use of high-stakes college admission exams. There is an ongoing and high-profile debate on whether selective U.S. colleges should use standardized exams in college admissions (Belasco et al., 2015; Bennett, 2022; Dessein et al., 2023; Dynarski et al., 2023). A central question in this debate is whether admission scores are "biased" in favor of high-income and white students because they have greater access to test prep.[5] A vast literature shows that there are large gaps in college admission exam scores by race and socioeconomic status (e.g., Bowen and Bok, 1998; Kane, 1998; Card and Rothstein, 2007; Goodman et al., 2020), while other work shows that admission scores are predictive of longer-run outcomes (Rothstein, 2004; Bettinger et al., 2013; Chetty et al., 2023; Riehl, 2023; Friedman et al., 2024).[6] But a limitation with

[5] For example, in 2019 a group of students and advocacy groups sued the University of California over its use the SAT and ACT exams. The plaintiffs' complaint quotes UC Regents Chairman John Pérez: "The highest predictive value of an SAT isn't in how well a student will do in school, but how well they were able to avail themselves of prep material. And access to that prep material is still disproportionately tied to family income" (Rosenbaum et al., 2020).

[6] Related research shows how the design and implementation of admission exams can either reduce or decrease inequality in college access (Hoxby and Turner, 2013; Bulman, 2015; Pallais, 2015; Goodman, 2016; Bhattacharya et al., 2017; Reyes, 2023).

papers in both of these literatures is that they focus on a static admissions environment, so it is unclear how test score gaps or informativeness would change if universities used alternative admission criteria. Consistent with the common criticism of admission tests, we find that racial and socioeconomic gaps in performance expand on higher stakes assessments, and we show that test prep is a likely mechanism for this result. Yet we do not find that test prep creates bias in scores as a measure of college performance. Thus our findings suggest that universities face a fundamental tradeoff between equality and informativeness in choosing whether or not to use admission tests.

The paper proceeds as follows. Section 2 provides background on Brazilian higher education and the ENEM exam. Section 3 describes our data and empirical strategy. Sections 4 and 5 present our results on test score inequality and informativeness. Section 6 concludes.

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 tertiary 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 at the top of national college rankings. The Brazilian higher education system additionally includes over 2,000 private universities and technical colleges that enroll roughly 80 percent of all college students. While a handful of these private institutions are elite and selective, the majority are moderately selective or have essentially open enrollment.

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.[7] But students from private high schools are vastly overrepresented in higher education; in 2009, they accounted for 40 percent of all incoming college students and 47 percent of federal university enrollees (Appendix Table A1).

2.2. **Federal university admissions and the ENEM exam.** Admission to federal universities is highly competitive and is based exclusively on test scores from entrance exams.[8] Before 2009, each federal university designed and administered its own admission tests, which

[7] Roughly 0.5 percent of Brazilian students attend a high school managed by the federal government (Appendix Table A1). Our empirical analysis defines "private high schools" to include both privately- and federally-managed high schools since their students are similar in socioeconomic status and achievement.

[8] During our sample period, many federal universities implemented reserved quotas for minority and/or public school students. Within these quota groups, admissions are based solely on entrance exam scores. Below we discuss the implications of the affirmative action quotas for our empirical analysis.

are called *vestibular* exams. This made it 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 providing public information on high school performance. Between 1998 and 2008, the ENEM exam consisted of 63 multidisciplinary 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 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*) that matches students to colleges based on 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. 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). Thus, some federal universities began using the ENEM immediately in 2009, while others adopted it five or more years later.[9] As we describe Section 3, our empirical strategy exploits this variation in the timing of ENEM adoption by federal universities.

2.3. **The market for test prep.** Preparation for college entrance exams is a central part of the lives of many Brazilian students as they approach the end of high school. 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 exam performance of prior cohorts to attract new students, and newspapers publish annual rankings of school-mean scores. 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).

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

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, test prep remains heavily concentrated in the private sector. Lower-income students often cannot afford to enroll in private high schools or *cursinhos*, and there is typically less emphasis on test prep in public school curricula (Silva, 2014). Our first empirical analysis asks whether these disparities in access to test prep contribute to inequality in access to selective colleges.

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. We also observe individuals' responses to each exam question, which allows us to measure the questions that individuals got right and wrong. The data for each question includes the learning objectives, the Item Response Theory (IRT) parameters, and the question text.

To measure longer-run outcomes, we link the ENEM data to two other administrative datasets at the individual level (see Appendix C.2 for details on the merge.). First, we measure college outcomes by linking to INEP's higher education census (*Censo da Educação Superior*) for the years 2010–2019 (INEP, 2022a). This dataset contains information on the universe of Brazilian college students, 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 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 role as a measure of high school performance.

Our analysis sample includes high school seniors at schools that met the criteria to be included in the government's performance 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).[10] Schools were included in the performance reports if a large fraction of their students participated in the exam.[11] 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. 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. Section 3.5 presents 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.[12]

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

[10] INEP published high school ENEM performance measures with the goal of "assist[ing] teachers, principals and other educational managers in identifying deficiencies and good practices" (INEP, 2019b). There were no financial incentives tied to school performance.

[11] 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.

[12] 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 details.

scoring 1.4 SDs higher than public school students on average.[13] 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 December 2009, only 28 percent of federal university students nationwide were admitted using the ENEM in the following year.[14] 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's 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 the state where the university is located. 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 close to home.

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 exploits all variation in ENEM adoption timing. In particular, $\text{ProportionENEM}_{st}$ reflects variation in ENEM 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

[13] For reference, the white/Black gap in the 2017 U.S. SAT math exam was 0.85 SDs (College Board, 2017).
[14] The ENEM is administered near the end of each year, and scores are used for admission to university cohorts that begin in the following calendar year.

value τ_s^* that yields the highest R^2 across these regressions.[15] 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^*.[16] 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 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 admission slots allocated using the ENEM increased by over 60 percent in the adoption year, and this share remains high in subsequent years. Appendix Table A2 shows the values of 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:

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

Y_{ht} is an average outcome for students who attended high school h and took the ENEM 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, ProportionENEM$_{s(h)t}$, which measures the stakes of the ENEM exam in state $s(h)$ and cohort t. In alternate specifications, we replace ProportionENEM$_{s(h)t}$ with our binary treatment variable, 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

[15] 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 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 HighStakes$_{st} = \mathbb{1}\{t \geq \tau_s^*\}$. We define one state (Sergipe) as a "never adopter" since the value of ProportionENEM$_{st}$ never exceeds 0.06.

[16] Throughout the paper, we refer to cohorts prior to each state's ENEM adoption year as "low stakes" cohorts for brevity. 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 (*O Programa Universidade Para Todos*) scholarships and eligibility for the federal FIES (*O Fundo de Financiamento Estudantil*) 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. Despite these other incentives, we argue that the adoption of the ENEM by federal universities significantly increased the exam's stakes for students who wished to attend a selective university close to home.

who took the college admissions version of the ENEM exam in 2009–2017, which holds the structure of the ENEM exam fixed over time. 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:

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

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

To address potential concerns about treatment effect heterogeneity in two-way fixed effects models (De Chaisemartin and d'Haultfoeuille, 2020), we present robustness results that restrict identification to clean comparisons based on states' ENEM adoption years. Our approach follows Callaway and Sant'Anna (2021) in estimating 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, and the 2011 adopters are our control ("not yet treated") group since these states had not yet adopted the exam. We define groups for all pairwise combinations of ENEM adoption years and, within each pair, we restrict the sample to exam cohorts prior to 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 groups. The resulting β^{gap} coefficients are regression-weighted averages of the pairwise treatment effects.[17]

3.5. **Identification assumptions and balance tests.** Our identification relies on a parallel trends assumption, which requires that the timing of federal universities' switch to the ENEM exam is unrelated to state-level trends in potential test score outcomes. This assumption could be violated if the adoption of the ENEM exam induced students to enroll in different high schools or impacted the types of students who took the exam.

[17] Our stacked specification uses our binary treatment variable, $\text{HighStakes}_{s(h)t}$, in the following regression:

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

In this specification, the dataset is at the high school (h) × year (t) × pairwise group (g) level, and we include state × group dummies, $\gamma_{s(h)g}$ and $\tilde{\gamma}_{s(h)g}$, and year × group dummies, γ_{tg} and $\tilde{\gamma}_{tg}$. Appendix Table A4 shows the pairwise groups and the structure of our stacked dataset.

Table 2 tests this assumption by examining whether the adoption of the ENEM was related to trends in exam taking and school enrollment (Panel A) or the demographic characteristics of students in our analysis sample (Panels B–C). Column (A) shows the mean of each dependent variable in cohorts prior to the state's ENEM adoption year. Columns (B)–(D) present 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 differences between the β coefficients in columns (C) and (D).

We find no evidence that the adoption of the ENEM exam caused students to attend different high schools or impacted the number of exam participants in our sample. Panel A of Table 2 shows regression results using three outcome variables: 1) the log number of ENEM participants per school/cohort at the high schools in our sample; 2) the log number of seniors per school/cohort at these high schools; and 3) the number of distinct schools that these seniors had attended in the past three years. We find that the increase in ENEM stakes did not significantly affect the number of exam takers or the number of seniors at either public or private schools in our sample. It also did not induce students to switch schools at a higher rate. This evidence is consistent with our prior that the incentives to attend a prep-oriented high school did not change significantly when federal universities switched from their own admission exams to the ENEM; rather, it mainly affected prep schools' incentives on *which* exam to focus on in their curriculum.

Panels B–C of Table 2 show that the increase in ENEM stakes also did not significantly impact the composition of students in our analysis sample. 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. We do find that a 100 percentage point increase in federal universities' ENEM adoption is associated with a 1.4 percentage point decrease in the fraction of ENEM participants who were female, but this effect is small and it is similar in both public and private high schools. 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). In Panel C, we also find no impacts on an index of *predicted* ENEM scores based on all of these demographic characteristics.

Finally, we find no systematic differences in the size, selectivity, or student body characteristics of federal universities that were early- vs. late-adopters of the ENEM exam (Appendix Table A3). 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. In net, the evidence supports the assumption of parallel trends in potential exam score outcomes within our analysis sample.

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.

The increase in the stakes of the ENEM led to a widening of private/public test score gaps. Private 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 students' scores did not change significantly on the core subjects (column D), and test score gaps between private and public school students increased (column E). Our point estimate for the average core subject scores implies that a 100 percentage point increase in the adoption of the ENEM by federal universities is associated with a 0.11 SD increase in the private/public test score gap. This effect is nine percent of the mean test score gap in lower-stakes cohorts (column A). We also find that the private/public gap in writing scores expanded 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, τ_s^*.[18] 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 that the private/public score gap increased by roughly 0.05 to 0.10 SDs in the first cohort after ENEM adoption. These wider gaps decline only slightly in subsequent cohorts. For example, the private/public gap in average core subject scores 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).

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

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

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. We also see clear evidence that our results are driven by a causal effect of ENEM adoption using the default event study figures from Callaway and Sant'Anna (2021)'s `csdid` Stata package (Appendix Figure A2).

The magnitudes of our estimates represent meaningful increases in private students' chances of gaining admission to federal university programs. The effect of higher stakes on the private/public gap in average core subject 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.[19] To put this magnitude in perspective, consider a private school student whose low-stakes ENEM score would have put them exactly at the cutoff for 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 racial and socioeconomic test score gaps also expanded on the higher stakes ENEM exam (see also Appendix Table A6). The white bars represent mean gaps in average core subject 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. 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.[20]

4.2. **Robustness to model specification.** Table 4 examines whether our results on private/public test score gaps are sensitive to model specification. 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 to 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

[19] 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.
[20] Our null result for gender differs from work that finds that male/female test score gaps are larger on high stakes exams (Ors et al., 2013; Azmat et al., 2016; Cai et al., 2019). This work interprets their results as evidence of gender differences in responses to competitive pressure. We think our results are more attributable to test preparation (see below), which occurs over a longer time span and is likely to be gender neutral.

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 core subject score remains statistically significant at $p < 0.05$.

Next, we examine the robustness of our results to potential concerns about two-way fixed effects models with treatment effect heterogeneity. For this, we use three different samples from the stacked dataset described in Section 3.3 (see also Appendix Table A4). 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). This sample includes 2007–2012 test takers because we require a pre-period to estimate a treatment effect for 2009 adopters. In column (F), we include all 2007–2017 test takers and all pairwise combinations in our stacked dataset.[21]

Our benchmark estimates are robust to each of these alternative specifications. 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 model that compares only 2009 and 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.

4.3. **Other robustness tests.** Our results are robust to alternate definitions of our analysis sample. Our benchmark sample includes seniors at high schools that met the criteria to be included in the government's ENEM performance reports in each year from 2005–2015 (see Section 3.2). Appendix Table A5 includes samples defined by both laxer criteria (e.g., schools that appear in *any* year) and stricter criteria (e.g., at least a 50% ENEM participation rate in all years). In all samples, we find positive and statistically significant estimates of β^{gap} for average core subject scores. With the exception of writing scores, the magnitudes of our estimates for each subject are relatively stable across samples.

Our estimates in Table 3 are mostly robust to an alternative method of statistical inference for settings with a relatively small number of clusters. Our benchmark estimates allow for

[21] Note that in columns (E)–(F) of Table 4, 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. 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.

correlated errors within each of Brazil's 27 states, which is below the rule of thumb for potential few-cluster issues in Angrist and Pischke (2010). Appendix Table A7 shows that our estimate of β^{gap} for average core subject scores remains significant at $p < 0.05$ using p-values from the wild t bootstrap procedure recommended by Cameron et al. (2008). The score gap estimates for other subjects remain significant at $p < 0.10$ except for natural science and writing.

Lastly, our results are not driven by federal universities' adoption of affirmative action or changes in the size of their admission quotas. Many federal 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 Brazil's 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 of β^{gap} remain positive and significant with these controls. Appendix Table A9 shows that our treatment variable is unrelated to the total number of federal university enrollees, which suggests that federal universities did not alter the number of admission slots when they switched to the ENEM exam.

4.4. **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 the increase in staked induced students to exert 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, so 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).[22] Thus, moderate increases in ENEM scores were unlikely to significantly affect the admission chances of many public students. Although we cannot observe student effort in our data, this may partly explain why we only find increases in private students' scores.

The increase in private students' scores could also reflect an increase in ENEM-specific test preparation. Students who wished to attend selective college programs may have switched their preparation efforts from the *vestibular* exams to the ENEM exam when federal universities adopted the ENEM. Anecdotally, many private schools and test-prep companies

[22] Some individuals in our sample would have been eligible for admission through reserved quotas for disadvantaged students, but these quotas were not fully implemented at many federal universities until 2016. Appendix Figure A1 shows that most public students' ENEM scores were also well below the distribution of cutoff scores for these reserved quotas.

altered their curriculum to focus more on the skills and content covered by the ENEM. By contrast, test prep is not a major focus of the curriculum at most public high schools, in part because the ENEM is a very difficult test for the typical public student. Thus the teaching practices at public high schools were less likely to change with the ENEM rollout, which may explain the expansion of the private/public ENEM score gap.[23]

We empirically examine the role of exam preparation using two measures of students' test prep activity. 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*) and 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.[24] 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, and, likely for this reason, we do not find evidence that ENEM adoption impacted the proportion of students who took a test prep course. Yet, if these courses were more likely to focus on the ENEM exam after its adoption by federal universities, this could raise the average ENEM scores of students who took preparation courses. Appendix Table A10 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 led to larger test score gaps between students who did/did not engage in test prep as defined by these measures. ENEM adoption led to a 0.18 SD increase in the gap in mean ENEM scores between prep-focused private schools and public schools (second row in Panel B), which is roughly 60 percent larger than our point estimate for the overall 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 this gap closed by 0.08 SDs ($p < 0.05$) following ENEM adoption. Lastly, we find that the increase in ENEM stakes led to a large increase in the 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 the private/public test score gap.

In sum, this section showed that test score gaps between advantaged and disadvantaged students expanded when the stakes of the ENEM exam increased. Critics of high-stakes tests

[23] Another possibility is that public school curricula was slower to respond to the ENEM rollout than private school curricula, which may explain the slight fade-out of some of our test scores results in Figure 3.

[24] 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 some other private schools in our sample also have test-oriented curricula. Thus, the coefficients that we estimate may be attenuated due to under-classification of prep schools.

often argue that they give wealthy students a leg up in the college admission process. Our findings are consistent with this claim. Yet is it unclear whether higher stakes cause exams to be "biased" in favor of high-income students, as the question of bias depends on what the exam is intended to measure. To shed light on the relationship between exam stakes and informativeness, we now turn to our second empirical analysis.

5. Exam stakes and the informativeness of scores

5.1. **Potential channels.** Are high-stakes exam scores more or less informative for students' academic potential than low-stakes scores? The answer, according to both economic literature and public debates on standardized testing, remains unclear.[25]

On the one hand, an increase in exam stakes may reduce the informativeness of scores by distorting effort toward activities that improve test performance rather than 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 an increase in the stakes of an exam causes test scores to become relatively more informative about individuals' *gaming ability* and relatively less informative about their *natural action* that would arise in the absence of signaling concerns. Similarly, critics of high-stakes testing often argue that the ability to "game the system" through test prep 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 testing encourages students to reveal or develop skills that are beneficial for their academic careers. As Frankel and Kartik (2019) note, an individual's gaming ability for standardized tests could reflect work ethic or the capacity to learn new material. Moreover, if high-stakes scores are more correlated with family income, they may be more predictive of college outcomes because family resources help students succeed in school. High-stakes testing may also compel students to accumulate new skills that benefit them beyond the exam. Students may learn important academic material if the exams are well-aligned with school curricula (Lazear, 2006). They may also develop non-cognitive skills like cognitive endurance (Brown et al., 2022; Reyes, 2023) or a growth mindset (Dweck, 2006) while preparing for exams.

5.2. **Outcome variables and regression model.** To shed light on the relative strength of these mechanisms, we ask how the increase in the stakes of the ENEM impacted the informativeness of ENEM scores for students' college and labor market outcomes. Our

[25] Appendix B.1 presents a theoretical framework that illustrates the potential channels through which exam stakes can impact test score gaps and exam informativeness. This section briefly describes the intuition.

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 because of a data issue.[26] We linked this sample to Brazil's higher education census to measure enrollment, persistence, and graduation outcomes at all Brazilian colleges in the years 2010–2019. We measure earnings outcomes using Brazil's national employer-employee data for the years 2016–2018.

We define three types of outcome variables based on the sample for which we can measure each outcome (see Appendix C.1 for details). Our first set of variables include outcomes that we can define for all ENEM participants in our sample. These 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 define measures of college persistence and graduation within the subsample of individuals who enrolled in college. These outcomes include indicators for persisting in college several years after enrolling and an indicator for completing the program within five years. Lastly, for individuals who appear in the RAIS dataset, we measure labor market earnings using an individual's mean hourly wage from 2016–2018 (measured in both logs and levels). Many of the ENEM participants in our sample were still in college during 2016–2018, and even those who had left college were early in their careers. Thus it is important to stress that our earnings outcomes do not reflect the long-run returns to individuals' college investments.

We modify our regression model to examine how federal universities' adoption of the ENEM impacted the *correlation* between students' ENEM scores and longer-run outcomes. Our regression model for exam informativeness is a state × year version of equation (1):

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

The dependent variable, Y_{st}, is the correlation coefficient between ENEM scores and a longer-run outcome among students who attended high school in state s and took the ENEM in year t. This specification follows testing agencies' standard practice of measuring exam validity using correlations coefficients (e.g., Kobrin et al., 2008).[27] As above, the variable of interest, $\text{ProportionENEM}_{st}$, is the proportion of federal university enrollees in state s and year t

[26] 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 do not observe many of the longer-run outcomes in these cohorts given the timing of our data.

[27] Another way of measuring the information content of an exam is to "anchor" scores to an economic outcome of interest (e.g., Cawley et al., 1999; Jacob and Rothstein, 2016; Nielsen, 2023). We focus on correlation coefficients between scale scores and outcomes because scale scores are what colleges use to admit students. Our measure of informativeness is closely related to the theoretical concept of Blackwell informativeness used in Frankel and Kartik (2019) and subsequent studies on muddled information. Appendix B.2 proves that with binary states and signals, an increase in a signal's Blackwell informativeness necessarily leads to a higher correlation between states and signal realizations.

who were admitted using the ENEM exam. Thus β measures the impact of a 100 percentage point increase in ENEM adoption on the correlation between ENEM scores and outcomes.

In complementary regressions, we also follow exam validity research in restricting comparisons to students who attended the same college programs. In addition to the potential mechanisms discussed in Section 5.1, a student's ENEM scores may be correlated with their longer-run outcomes through their direct impact on which college and/or major they gained admission to. Further, many federal universities that adopted the ENEM simultaneously joined the SISU centralized college admission system, which likely 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 residualize both ENEM scores and outcomes on dummies for college $\times$ major pairs and then compute correlation coefficients using these residuals. These residual correlations reflect variation in the informativeness of scores among students who attended the same college programs.[28]

5.3. **Effects on informativeness.** Table 5 shows our main results on how federal university adoption of the ENEM exam impacted the informativeness of ENEM scores. Column (A) shows the mean correlation coefficient between average core subject scores and each outcome variable in cohorts prior to ENEM adoption. Columns (B)–(C) present β coefficients from equation (5) using our continuous treatment variable, $\text{ProportionENEM}_{st}$. Columns (D)–(E) present β coefficients using our binary treatment variable, HighStakes_{st}. In columns (B) and (D), the dependent variables, Y_{st}, are raw correlation coefficients. In columns (C) and (E), the dependent variables are correlation coefficients measured after residualizing ENEM scores and outcomes on college $\times$ major dummies.

We find that scores on the higher-stakes ENEM exam were *more* informative for students' college enrollment, persistence, and graduation outcomes. Panel A of Table 5 shows that the informativeness of average ENEM scores for 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 average ENEM scores and an indicator for college enrollment, and a 0.033 increase in the correlation between

[28] Rothstein (2004) points out that it is difficult to justify the sample selection assumptions that are implicit in many validity studies, and he offers a new estimator for settings in which the variables that determine admission to selective colleges are observable (e.g., the University of California system). We cannot implement his estimator because we do not observe the *vestibular* scores that federal universities used in admissions prior to ENEM adoption. Our analysis partially addresses Rothstein's concerns by including some outcomes that are defined for our entire sample (e.g., college enrollment and unconditional college graduation). Further, our regression model (5) estimates *changes* in exam informativeness across cohorts, which differences out the impact of unobservable determinants of sample selection that do not vary over time. Like Rothstein (2004), our benchmark results rely on the strong assumption that students' decisions about which college program to attend are ignorable, but we also present results that examine within-program validity.

average ENEM scores and an indicator for completing a college degree by 2019. We find similar (but less precise) estimates using the binary treatment variable (column D). The results are also similar we when compare degree attainment outcomes for students who attended the same college programs (columns C and E). This suggests that our findings are not driven by direct impacts of ENEM performance or the SISU system on the program that students attended.

The increase in the informativeness of ENEM scores is even more pronounced for college persistence outcomes measured within the population of college enrollees. In Panel B of Table 5, the 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 the estimated β coefficients in columns (B)–(E) are broadly similar 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 a 100 percentage point increase in ENEM adoption is associated with a 23 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. Appendix Table A11 shows that the results are mostly robust to the wild t bootstrap procedure.

We do not find conclusive evidence on the relationship between exam stakes and the informativeness of scores for hourly wages. Panel C of Table 5 shows 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. Further, the coefficients for both wage levels and log wages lose statistical significance when we use a wild t cluster bootstrap for inference (Appendix Table A11). These inconclusive results are likely due to the fact that we observe wages for only one-quarter of students in our sample because many individuals were still in college during our data period.

Figure 5 shows that informativeness increased in all four core subjects. The translucent areas depict the mean correlations in lower-stakes cohorts (analogous to column A of Table 5), and the darker areas depict the β coefficients from our benchmark regression model (analogous to column B of Table 5). The increase in exam stakes caused ENEM scores to become more correlated with degree completion, 3-year college persistence, and 5-year program completion in each of the four core subjects, and the magnitudes are relatively similar in each subject. We also find increases in the informativeness of writing scores, but these are smaller in magnitude and imprecisely estimated.

5.4. **Correlation with demographics.** Why were high-stakes ENEM scores more predictive of college success? As noted above, one possibility is that high-stakes scores may have been more correlated with other individual characteristics that help students succeed in college, such as family income or parental education. To assess this potential channel, we show how demographic controls impact the results on exam informativeness.

Figure 6 plots β coefficients from different specifications of equation (5). The white bars reproduce the results from our benchmark specification (Table 5, column B). The darker-colored bars show β coefficients including four sets of demographic controls: 1) a dummy for private high schools; 2) dummies for family income, parental education, and race; 3) gender and age; and 4) high school dummies. In each case we include the new control variables plus the controls from the previous specification. Thus, these β coefficients show how ENEM adoption impacted the exam's capacity to identify academic potential among students from similar demographic groups. Appendix Table A12 shows the corresponding regression results.

Demographic variables explain some, but not all, of the increase in the informativeness of ENEM scores. For example, the β coefficient for 3-year college persistence rates falls by 37 percent when we control for private high school and socioeconomic variables, and the coefficient for 5-year program completion rates falls by 60 percent. The impact of private high school and socioeconomic controls is similar or more modest for most other outcomes. Gender and age do not explain much of the increase in informativeness for most outcomes. The inclusion of high school dummies reduces the β coefficients by more than 80 percent for college persistence and degree completion, but these dummies do not have additional explanatory power for college enrollment or unconditional college graduation.

Importantly, these findings show that the higher-stakes ENEM scores provided new information on students' academic potential above and beyond easily observed demographic characteristics. By the end of our sample period, most federal universities had reserved quotas for public high school, low-income, and/or Black applicants, but otherwise admissions were based solely on ENEM performance. Thus our specification with private school, family income, and race controls shows that the higher-stakes ENEM exam helped federal universities identify applicants who were more likely to succeed within the set of demographic characteristics that they considered in admissions.

5.5. **Narrow vs. broad-based learning.** To further explore mechanisms, we use question-level ENEM data to examine whether the gains in private students' scores were driven by narrowly-targeted or broad-based learning. ENEM questions are based on a reference matrix of skills that educators regard as important for students to learn in high school. For example, math questions cover *topic areas* such as algebra, geometry, and statistics. Within each

topic area, questions are designed to test specific *competencies* such as identifying concepts, solving problems, and constructing arguments. Our data includes individuals' responses to each question, which allows us to estimate our regression model (2) separately for questions in each topic area or competency.[29] 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. This would allow us to measure the impact of ENEM adoption on students' question-level performance. 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 amenable to test prep (Riehl and Welch, 2023). Appendix Table A13 shows results for questions in language arts, natural science, and social science.

Panel A of Table 6 begins by showing results that pool across all 405 math questions in our sample (9 years $\times$ 45 questions per exam). Column (C) reports the mean proportion of correct answers for public students in cohorts prior to the state's ENEM adoption year, and column (E) reports the mean private/public gap in these cohorts. The average public school student answered 29.1 percent of the questions correctly, and the private/public gap in correct answers was 17.6pp. Columns (D) and (F) report the β and β^{gap} coefficients from equation (2). The estimate of β^{gap} in Panel A implies 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 averaged across all questions. This is consistent with our finding for scale scores in Table 3.

Panels B–C of Table 6 show that the private/public gap in correct responses increased across a wide range of math topic areas and competencies. These panels display results from estimating equation (2) separately for each topic area (Panel B) and competency (Panel C). 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 for questions on proportions. 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. Appendix Table A13 shows that the private/public gap in correct responses also increased across a wide range of topic areas in language arts, natural science, and social science.

Although the estimates of β^{gap} in Table 6 are uniformly positive, there is substantial variation in these coefficients across exam skills. In Figure 7, we ask whether the exam skills with the largest gains in private students' performance are more or less informative

[29] This analysis follows Jacob (2005) and Cohodes (2016) in using item-level data to shed light on the mechanisms for test scores gains. Unlike these papers, we are also able to link item-level data to longer-run outcomes to directly measure the informativeness of different exam skills.

for college persistence. The x-axes depict the β^{gap} coefficients from column (F) of Table 6 estimated at the topic area (Panel A) and competency (Panel B) levels. The y-axes display the informativeness of each topic area or competency for college persistence rates. We define informativeness as the average difference in 3-year persistence rates between individuals who answered each question correctly and those who did not, calculated across all questions within the same topic area or competency.[30] For example, our measure of informativeness for the algebra topic area is 0.046, which means that students who correctly answered a typical algebra question were 4.6pp more likely to persist in their college program for three years than students with an incorrect answer. Figure 7 includes exam skills from each of the four core subjects, as illustrated by the marker colors and symbols. The dashed line shows the linear relationship between exam skill informativeness and the β^{gap} coefficients.

Exam skills that experienced larger increases in the private/public gap in correct responses tend to be *more* informative for college persistence. Correct answers on the ENEM are highly informative for individuals' academic potential; the y-axes of Figure 7 show that students with a correct answer to an average question were roughly 5pp more likely to persist in college for three years than students with a wrong answer. Exam skills that contributed more to the expansion of the private/public test score gap (larger β^{gap} coefficients) also tend to be more informative for college persistence. For example, at the topic area level, math questions on proportions and interpreting data have the largest β^{gap} coefficients, and they are also among the most informative exam skills.

There is also a positive relationship between β^{gap} and skill informativeness for most other college and labor market outcomes. Appendix Table A14 shows competency-level OLS regressions of informativeness on β^{gap} for each of the outcomes in Table 5. We estimate regressions that pool across all four core subjects as well as separate regressions for each subject. In the pooled regressions, we find positive and mostly significant OLS coefficients for every outcome, with a particularly strong relationship between the β^{gap} coefficients and informativeness for wages. The positive relationship between informativeness and β^{gap} also arises within competencies on the math and language arts exams; we find no significant relationship in natural science or social science. This positive relationship is robust to including controls for question difficulty and other IRT parameters (Appendix Table A15). Although the magnitudes of the OLS coefficients are slightly reduced, we find no evidence that skills with larger β^{gap} coefficients are less informative for student outcomes.

[30] This measure of informativeness is closely related to the correlation coefficient between ENEM scores and outcomes, which is the measure of informativeness we use in Table 5. The correlation between an indicator for correctly answering question j, C_{ij}, and an outcome, Y_i, can be written as a function of the difference in the mean outcomes of students who got the question correct and incorrect, $\text{corr}(C_{ij}, Y_i) = (\mathbb{E}[Y_i|C_{ij} = 1] - \mathbb{E}[Y_i|C_{ij} = 0])\sigma_{C_j}/\sigma_Y$, where σ_{C_j} and σ_Y are the standard deviations of C_{ij} and Y_i.

5.6. **Discussion.** The results in this section run counter to the common criticism that high-stakes exams cause students to prepare for narrow exam skills that do not benefit them outside of the test. The adoption of the ENEM by federal universities increased the overall predictive power of scores for college enrollment, persistence, and graduation. Further, the increase in ENEM stakes caused private school students to perform better across a wide range of exam skills, and these performance gains were driven by skills that are more informative for longer-run outcomes.

Our mechanism analysis indicates that the increase in informativeness is partly due to an increase in the correlation between test scores and socioeconomic factors that influence both access to test prep and success in college. Yet our findings suggest that the higher stakes exam also helped students reveal or develop other harder-to-measure dimensions of academic potential. One of the government's objectives in redesigning the ENEM exam was to create a test that is 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 and/or rewarded students who had learned this material in high school.

6. Conclusion

Every year, admission committees at elite universities allocate scarce slots among many applicants on whom they have limited information. These committees seek to bolster their schools' reputations by admitting talented students (MacLeod et al., 2017), and so they consider signals of candidate quality such as standardized test scores, high school grades, and personal essays. Yet candidates have a strong incentive to manipulate these signals due to the perceived value of attending an elite university. This process gives an advantage to individuals who have the know-how and resources to improve their application credentials.

This paper examined how incentives to manipulate performance affect the distribution and informativeness of university admission scores. On the one hand, we found that test score gaps between private and public high school students increased when elite Brazilian universities began using the exam in admissions. In the language of theoretical work on muddled information (Frankel and Kartik, 2019), private school students had relatively higher "gaming ability" for standardized admission tests. Gaming ability may partially reflect access to test preparation resources; indeed, many wealthy Brazilian students take expensive prep courses and attend private high schools that focus on exam preparation. In this sense, our findings corroborate the concern that high-stakes admission exams give a leg up to wealthy students.

On the other hand, our paper showed that incentives to manipulate exam performance can actually *improve* the quality of the test score signal from the standpoint of admission

committees. We found that the predictive power of exam scores for college enrollment, persistence, and graduation increased when top Brazilian universities adopted the exam in admissions. This suggests that gaming ability may also reflect characteristics that help students succeed in college such as a willingness to exert effort, the capacity to learn new material, or family resources. This finding runs counter to the common narrative that test prep causes exam scores to be biased as a measure of academic potential.

Our findings highlight the challenge that universities face in seeking to admit both academically-prepared and diverse student bodies. Our paper shows that this problem is *not* solely due to demographic gaps in pre-college achievement; rather, the high-stakes nature of university admissions exacerbates the tension between diversity and informativeness. Brazilian universities balance these objectives by reserving admission slots for disadvantaged applicants and then admitting students who have the highest test scores within each pool. U.S. colleges are more constrained in their ability to consider demographic characteristics, particularly in the wake of the Supreme Court's 2023 ban on race-based affirmative action. Improved access to low-cost test prep through organizations like *Descomplica* in Brazil and the Khan Academy in the United States may help, but there is limited evidence on their effectiveness.

More broadly, the tradeoff between diversity and informativeness is likely to arise in other high-stakes settings such as recruiting at prestigious firms. We hope future research will shed light on the consequences of muddled information in other education and labor markets.

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.

Angrist, J. D. and J.-S. Pischke (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives 24*(2), 3–30.

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.

Ball, I. (2024). Scoring strategic agents. *American Economic Journal: Microeconomics (Accepted)*.

Belasco, A. S., K. O. Rosinger, and J. C. Hearn (2015). The test-optional movement at America's selective liberal arts colleges: A boon for equity or something else? *Educational Evaluation and Policy Analysis 37*(2), 206–223.

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

Bennett, C. T. (2022). Untested admissions: Examining changes in application behaviors and student demographics under test-optional policies. *American Educational Research Journal 59*(1), 180–216.

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.

Bowen, W. G. and D. Bok (1998). *The Shape of the River: Long-term Consequences of Considering Race in College and University Admissions*. Princeton University Press.

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.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics 90*(3), 414–427.

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.

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

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.

Chetty, R., D. J. Deming, and J. N. Friedman (2023). Diversifying society's leaders? the causal effects of admission to highly selective private colleges. NBER Working Paper No. 31492.

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.

Dessein, W., A. Frankel, and N. Kartik (2023). Test-optional admissions. arXiv preprint arXiv:2304.07551.

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

Dynarski, S., A. Nurshatayeva, L. C. Page, and J. Scott-Clayton (2023). Addressing non-financial barriers to college access and success: Evidence and policy implications. In *Handbook of the Economics of Education*, Volume 6, pp. 319–403. Elsevier.

Ederer, F., R. Holden, and M. Meyer (2018). Gaming and strategic opacity in incentive provision. *The RAND Journal of Economics 49*(4), 819–854.

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.

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.

Friedman, J., B. Sacerdote, and M. Tine (2024). Standardized test scores and academic performanceat ivy-plus colleges. Technical report, Opportunity Insights.

Goodhart, C. (1975). Problems of monetary management: the uk experience in papers in monetary economics. *Monetary Economics 1*.

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 (2022a). 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).

INEP (2022b). Microdata for School Census (Censo Escolar). 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.

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.

Kane, T. J. (1998). Racial and ethnic preferences in college admissions. *Ohio St. LJ 59*, 971.

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.

Krishna, K., S. Lychagin, W. Olszewski, R. Siegel, and C. Tergiman (2022). Pareto improvements in the contest for college admissions. NBER Working Paper No. 30220.

Lazear, E. P. (2006). Speeding, terrorism, and teaching to the test. *The Quarterly Journal of Economics 121*(3), 1029–1061.

Lee, F. X. and W. Suen (2023). Gaming a selective admissions system. *International Economic Review 64*(1), 413–443.

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.

Muller, J. (2018). *The tyranny of metrics*. Princeton University Press.

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.

Perez-Richet, E. and V. Skreta (2022). Test design under falsification. *Econometrica 90*(3), 1109–1142.

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.

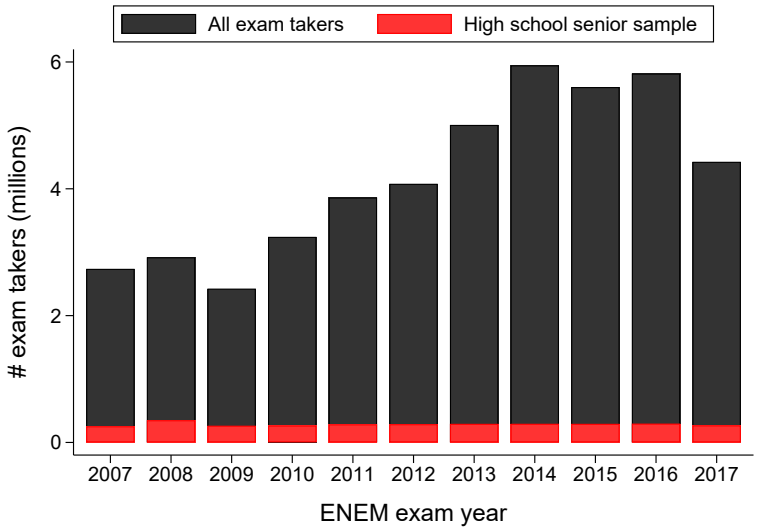
Riehl, E. and M. Welch (2023). Accountability, test prep incentives, and the design of math and English exams. *Journal of Policy Analysis and Management 42*(1), 60–96.

Rosenbaum, M., A. Savage, G. A. Ellis, K. B. Farkas, and M. L. Lavetter (2020). Kawika Smith et al. vs. Regents of the University of California, Superior Court of the State of California, County of Alameda, case no. RG19046222. https://publiccounsel.org/wp-content/uploads/2022/02/2020.06.15-Amended-Complaint.pdf.

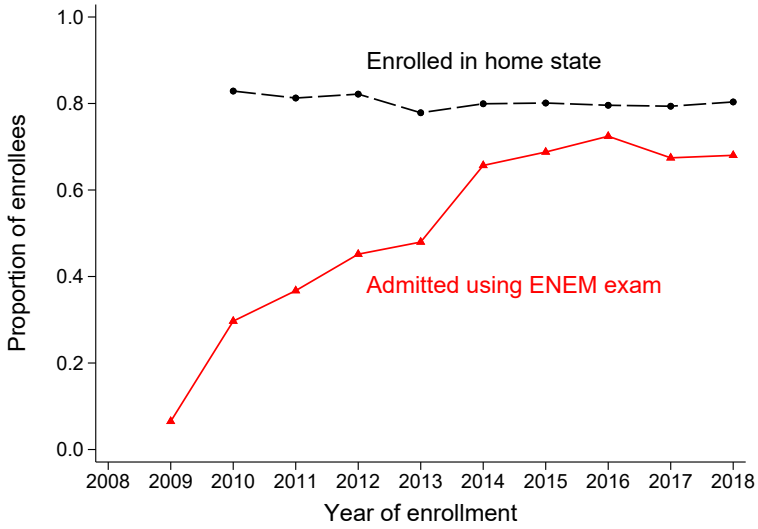
Rothstein, J. M. (2004). College performance predictions and the SAT. *Journal of Econometrics 121*(1), 297–317.

Silva, R. (2014, Dec). Escolas privadas e técnicas dominam o Enem. *Jornal Cruzeiro do Sul*. https://www2.jornalcruzeiro.com.br/materia/587165/escolas-privadas-e-tecnicas-dominam-o-enem (Accessed June 2023).

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



Panel A. Number of ENEM exam takers

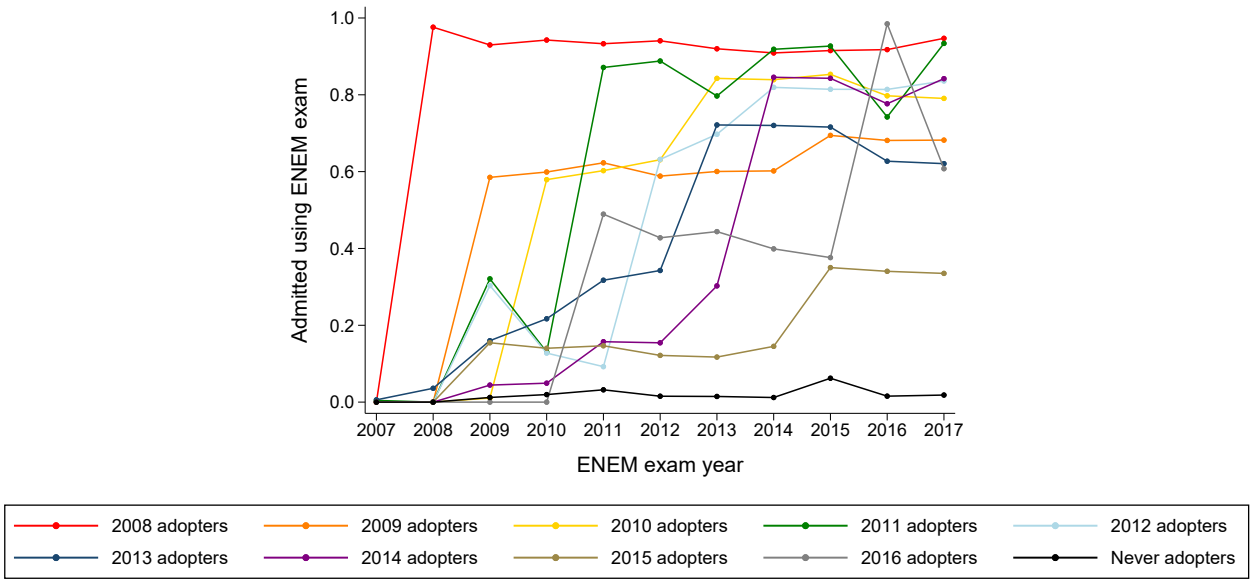


Panel B. Federal university enrollment

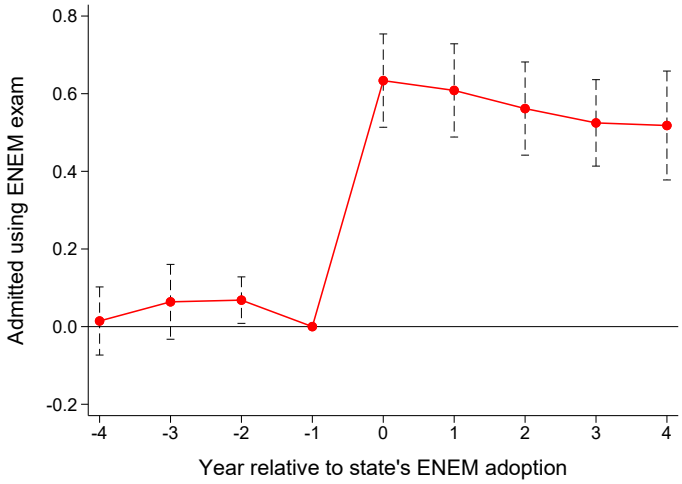
Figure 1. Adoption of ENEM exam by federal universities

Notes: This figure illustrates time variation in the number of ENEM exam takers and the proportion of federal university enrollees who were admitted using the ENEM.

Panel A shows the total number of individuals who took the ENEM each year from 2007 to 2017. Each bar displays the overall number of ENEM exam takers (black bars) and the number of exam takers in our analysis sample (red bars). Panel B shows the proportion of new enrollees in federal universities admitted through the ENEM exam in each year from 2009 to 2018 (red solid line) and the proportion of new enrollees in federal universities who attended a university in their birth state from 2010 to 2018 (black dashed line).



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 adoption by federal universities across states and years

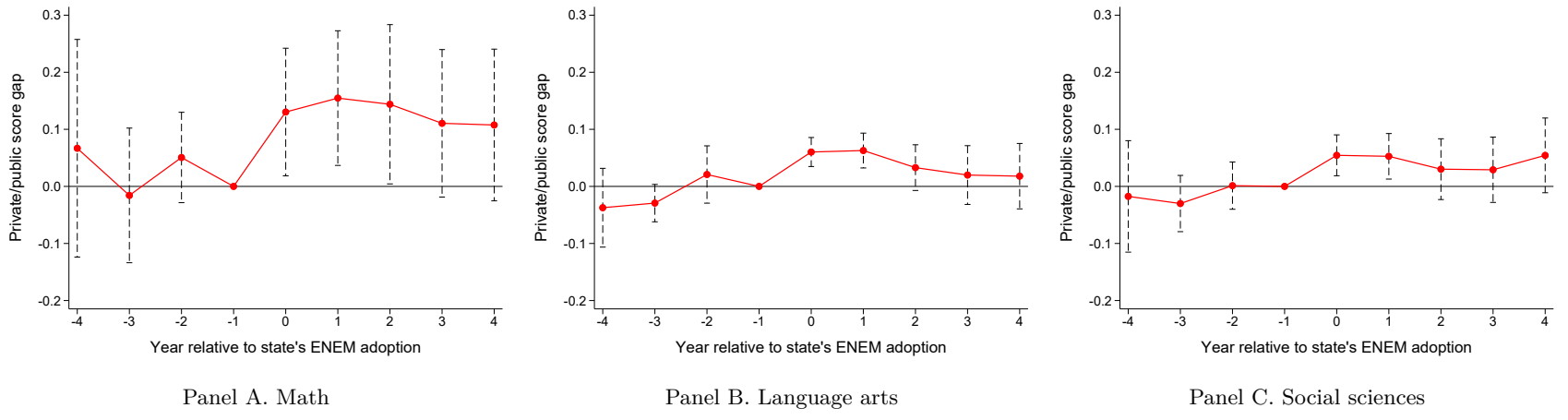
Notes: This figure illustrates the staggered adoption of the ENEM exam by federal universities. The outcome in both panels is the proportion of new enrollees in federal universities in state s who were admitted using the ENEM exam administered in year t (the calendar year prior to enrollment), denoted as $\text{ProportionENEM}_{st}$.

Panel A plots the mean of $\text{ProportionENEM}_{st}$ for groups of states based on their ENEM adoption year, τ_s^*, as indicated in the legend. See Section 3.3 for the definition of ENEM adoption years, τ_s^*.

Panel B plots event-study coefficients, β_l, from the state (s) $\times$ year (t) $\times$ pairwise group (g) level regression:

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

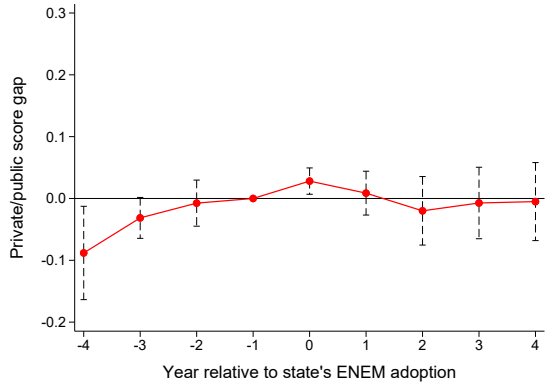
where l denotes years relative to τ_s^*. Pairwise groups (g) are defined described in Section 3.4. The regression includes state $\times$ group dummies (γ_{sg}), year $\times$ group dummies (γ_{tg}), and dummies for years l ($\mathbb{1}\{t - \tau_s^* = l\}$), with $l = -1$ omitted as the reference year. The graph plots the β_l coefficients from $l = -4$ to $l = 4$. Dashed lines depict 95% confidence intervals using standard errors clustered at the state level.



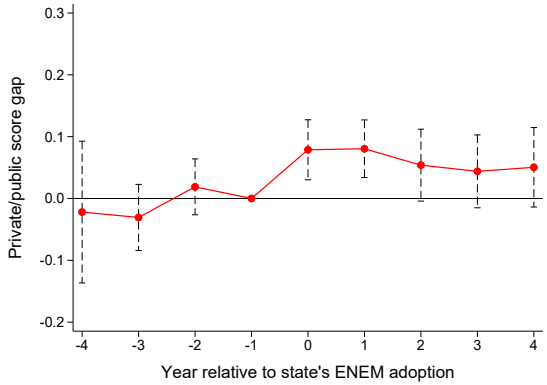
Panel A. Math

Panel B. Language arts

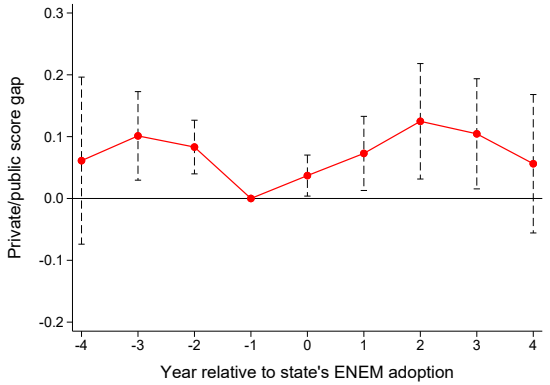
Panel C. Social sciences



Panel D. Natural sciences



Panel E. Average (core subjects)

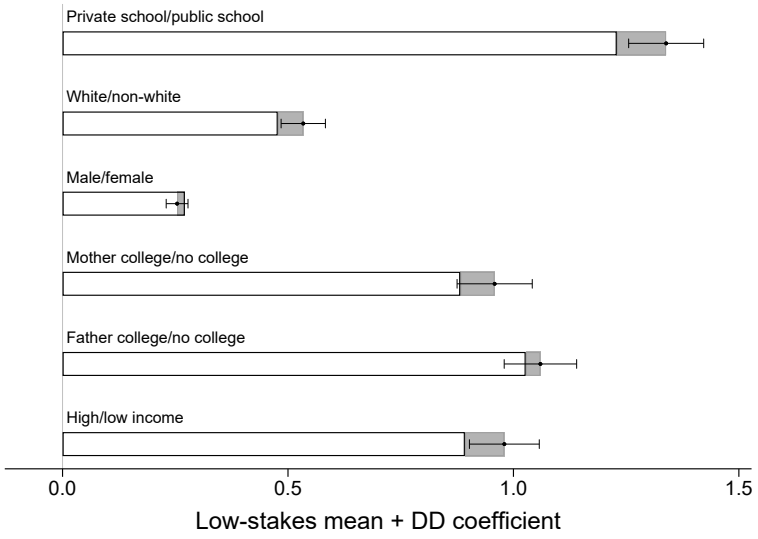


Panel F. Writing

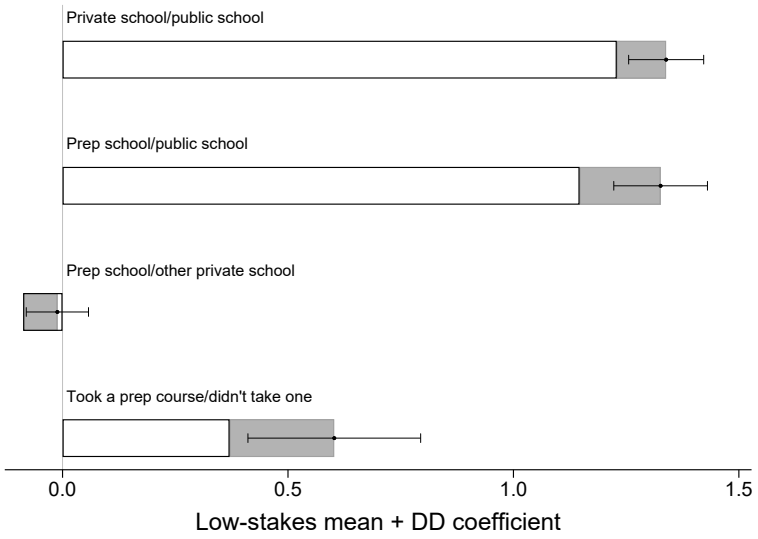
FIGURE 3. Event studies for effects of ENEM adoption on private/public test score gaps

Notes: This figure presents event study estimates of the impact of ENEM adoption on test score gaps between private and public school students. The sample includes all pairwise combinations of ENEM adoption years for which treatment effects can be estimated using 2009–2017 exam takers (the boxed cells in Appendix Table A4). Pairwise groups are defined described in Section 3.4.

Each panel plots the β_l^{gap} coefficients (y-axis) from equation (4) for years $l = -4$ to $l = 4$ relative to the state's ENEM adoption year, $\tau^*_{s(i)}$ (x-axis). The dependent variables are ENEM subject scores in SD units, as listed in the title of each panel. "Average (core subjects)" is the average score across math, language arts, natural science, and social science. 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 subjects) scores

Notes: This figure shows the impact of ENEM adoption on various gaps in average ENEM scores.

Panel A shows impacts on demographic test score gaps. "High-income" individuals are defined as those with a family income greater than or equal to twice the minimum wage. Panel B shows impacts on test score gaps between students who did and did not engage in test prep activities, as defined by two different measures. We define "prep schools" as private schools whose curriculum is specifically focused on preparation for college admission exams. To define this measure, 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 the student took an entrance exam preparation course. See Appendix Table A10 for details on these measures of test prep activity.

White bars represent mean gaps in average (core subjects) ENEM scores in low-stakes cohorts for each demographic/test prep group. Gray bars show estimates of β^{gap} from a specification of equation (2) that replaces Private_h with a dummy variable for the first group listed in the heading (e.g., high-income, prep school, etc.). For this figure, we estimate equation (2) at the individual level rather than at the high school × year level. Black bars represent 95% confidence intervals using standard errors clustered at the state level.

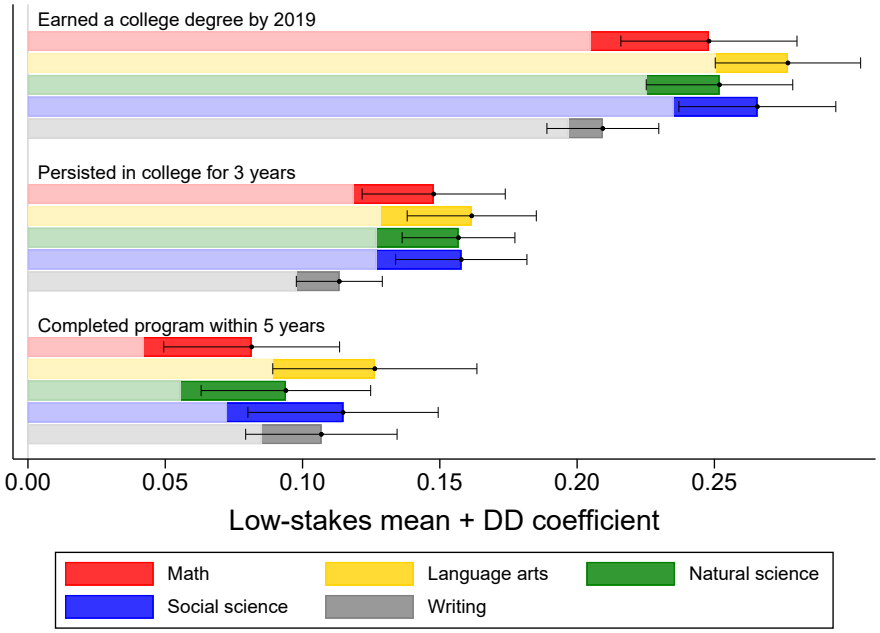


FIGURE 5. Effects of ENEM adoption on the informativeness of subject scores for longer-run outcomes

Notes: This figure shows the impacts of ENEM adoption on the informativeness of subject-specific ENEM scores for longer-run outcomes. Lighter-shaded areas depict the average correlation coefficients between subject scores and outcomes in low-stakes cohorts (i.e., cohorts where $\text{HighStakes}_{st} = 0$). Darker-shaded areas depict β estimates from equation (5), where the outcome variables are state $\times$ year correlation coefficients between subject scores and outcomes. The black bars represent 95% confidence intervals using standard errors clustered at the state level.

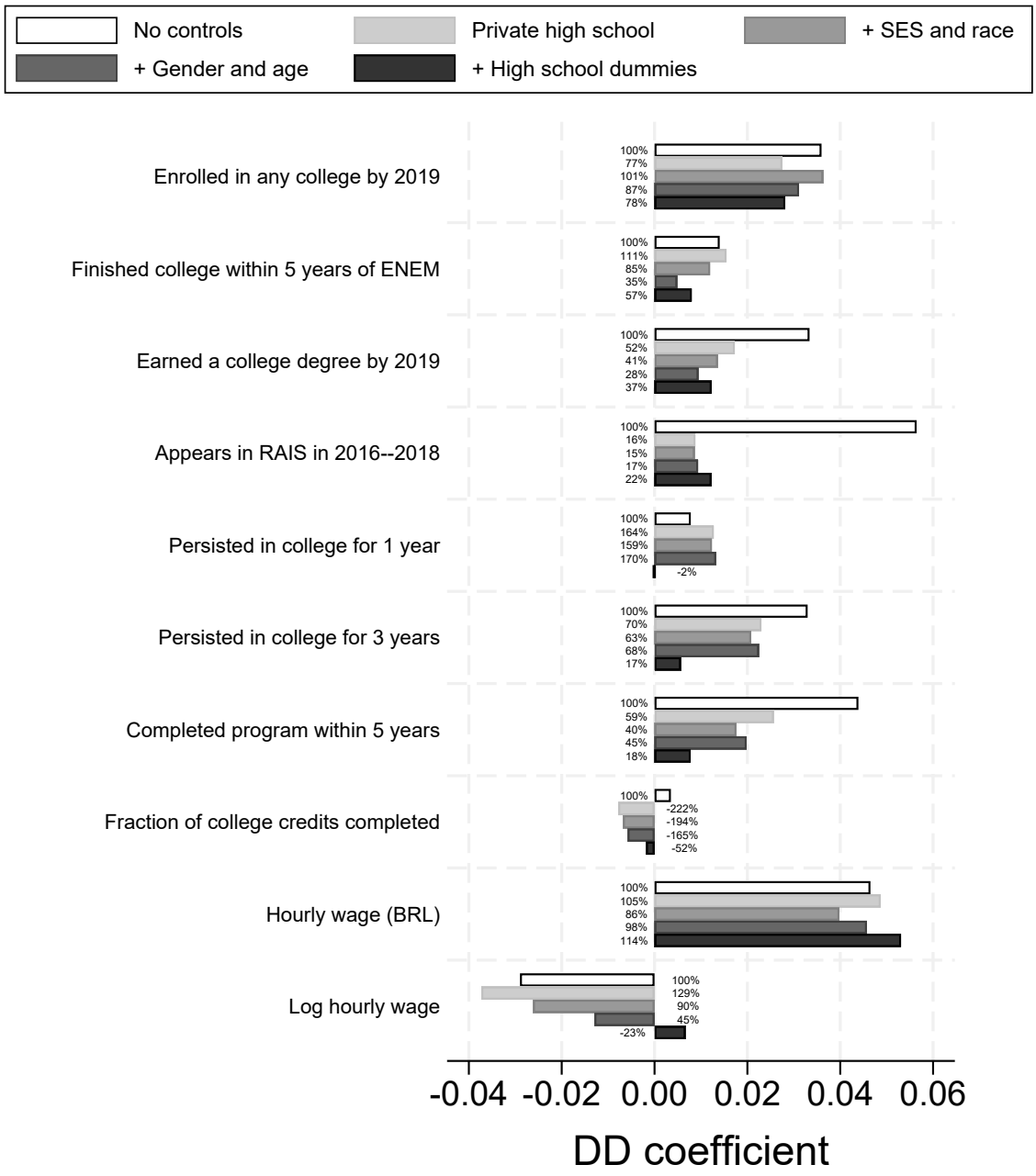
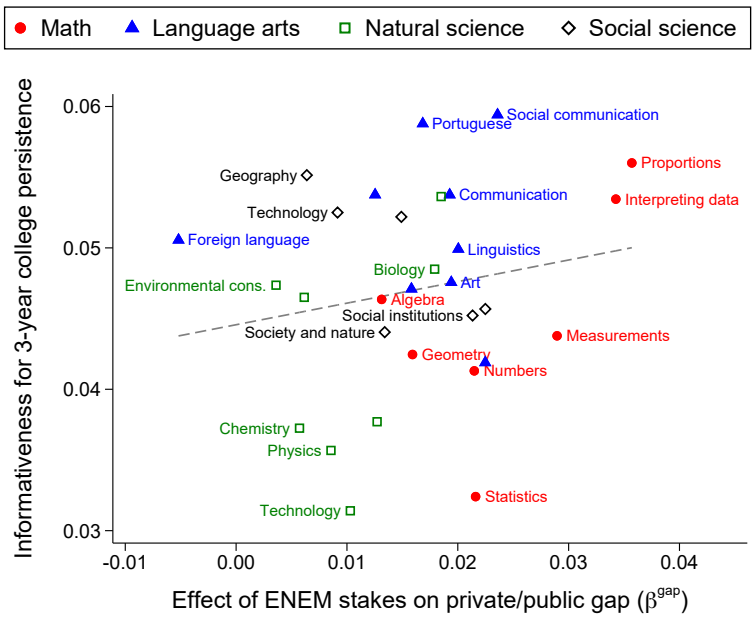


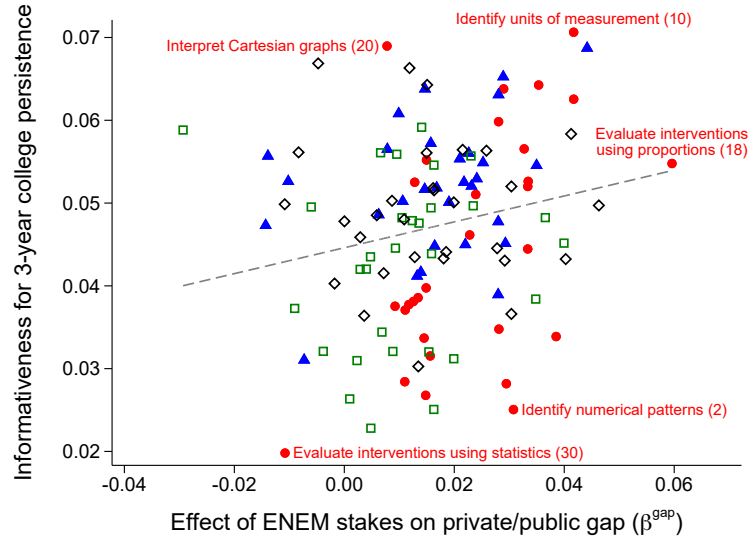
FIGURE 6. Effects of ENEM adoption on the correlation between outcomes and test scores controlling for demographics

Notes: This figure shows the impacts of ENEM adoption on the informativeness of average (core subjects) scores for longer-run outcomes after controlling for demographic characteristics.

The bars depict β coefficients from equation (5). The dependent variable is correlation coefficient between average (core subjects) ENEM scores and the longer-run outcome in state s and ENEM cohort t. The white bars reproduce the estimates from column (B) of Table 5. The darker-colored bars show the β coefficients in specifications that control for four sets of demographic controls: 1) a dummy for private high schools; 2) dummies for family income categories, mother's education, father's education, and race; 3) a gender dummy and age; and 4) high school dummies. Each darker-colored bar includes the new controls along with all controls from the previous bars. Percentages indicate the ratio between the β coefficient for the given bar and the β coefficient for the specification without demographic controls (white bars).



Panel A. Topic areas



Panel B. Competencies

FIGURE 7. Informativeness for college persistence vs. impact of exam stakes on private/public gap by exam skill

Notes: This figure shows the relationship between the informativeness of ENEM exam skills for college persistence (y-axis) and the effect of ENEM stakes on the private/public gap (x-axis). To define the informativeness of each exam skill for 3-year college persistence (y-axis), we compute the mean difference in 3-year persistence rates between individuals who got each question correct and incorrect using 2009–2014 (excluding 2011) ENEM participants in our analysis sample, and then average these differences across all questions in the same topic area (Panel A) or competency (Panel B). Our measure of the effect of ENEM stakes on the private/public gap (x-axis) 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 analysis sample. Marker colors and shapes depict exam subjects, as described in the legend. The dashed line shows the linear relationship between informativeness and the β^{gap} coefficients.

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 completed college	0.15	0.18	0.27	0.56	0.13	0.44
Father completed 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 in 2016–2018			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 ENEM exam takers. Column (A) includes all individuals who took the ENEM exam in 2009–2017 who have a non-zero, non-missing test score in all four core subjects. Column (B) includes all exam takers in these years who were high school seniors at the time of the exam. Column (C) includes exam takers in our analysis sample. Columns (D) and (E) include present results for private an public high school students without our analysis sample, respectively. Column (F) displays the difference between columns (D) and (E).

Panel (A) presents 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 average ENEM scores (in SD units). "Average score (core subjects)" is the average score across math, language arts, natural science, and social science. Panel (C) displays college and labor market outcomes for the exam takers in our analysis sample. The last two rows report the number of exam takers and high schools. See Appendix C.1 for details on variable definitions.

TABLE 2. Balance tests for analysis sample

	(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. Exam taking and school enrollment					
Log # exam takers in school	4.713	0.076	−0.004	0.089	−0.093
		(0.065)	(0.118)	(0.056)	(0.107)
Log # HS seniors in school	5.118	−0.033	−0.072	−0.033	−0.039
		(0.022)	(0.053)	(0.032)	(0.070)
# schools attended in past 3 years	1.220	−0.010	−0.023	−0.010	−0.013
		(0.012)	(0.015)	(0.016)	(0.017)
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 completed college	0.258	0.001	0.009	0.000	0.009
		(0.008)	(0.010)	(0.005)	(0.010)
Father completed 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
N (# HS seniors)	707,255	3,283,616	913,767	2,369,849	3,283,616

Notes: This table presents balance tests for 2009–2017 ENEM test takers in our analysis sample (column C of Table 1). In Panel A, the dependent variables are: 1) the log number of exam takers per school; 2) the log number of high school seniors per school; 3) the average number of schools each senior attended in the past three years. In Panel B, the dependent variables are exam taker demographic characteristics. In Panel C, the dependent variable is the predicted value from a regression of average ENEM score (core subjects) on age, gender, and dummies for race, mother's education, father's education, and family income bins. The dependent variables are high school × year totals (log counts) and averages (all other 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 the effect of ENEM adoption on the test scores of private and public high school students. The sample includes 2009–2017 ENEM exam takers in our analysis 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 school students, and public school students, respectively. Column (E) displays β^{gap} coefficients from equation (2) estimated using all students in the analysis sample. 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 school test score gaps

Dependent variable	(A) Benchmark model	(B) Demographic 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 of the effects of ENEM adoption on private/public test score gaps.

In columns (A)–(D), the sample includes 2009–2017 ENEM exam takers in our analysis sample (column C of Table 1). In columns (E)–(F), the sample also includes 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 a mean of 0 and SD of 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 four subjects and then estimating a scale scores using the IRT parameters. See Appendix C.1 for details.

Column (A) reproduces the estimates from column (E) of Table 3, which are the β^{gap} coefficients from equation (2). Column (B) shows estimates of 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) shows estimates of equation (2) replacing the continuous treatment variable, $\text{ProportionENEM}_{st}$, with the binary treatment variable, HighStakes_{st}. Columns (D)–(F) show estimates of β^{gap} from equation (3) using the 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. Effects of ENEM adoption on the informativeness of ENEM scores for longer-run outcomes

	(A)	(B)	(C)	(D)	(E)
	Low-stakes mean	Benchmark model DD coefficients		Binary treatment DD coefficients	
Dependent variable: Correlation between average ENEM scores and...	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 the impacts of ENEM adoption on the informativeness of ENEM scores for longer-run student outcomes. Our measure of informativeness is the correlation coefficient between the outcome in the column header and average (core subject) ENEM scores using data from our analysis sample for each state × year pair in 2009–2014 (excluding 2011). We estimate equation (5) using these correlation coefficients as dependent variables, weighting each state × year observation by the number of ENEM test takers for whom the outcome is defined.

Column (A) shows the mean correlation coefficients in exam cohorts prior to each state's ENEM adoption year. Columns (B)–(E) display 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 the continuous treatment variable, $\text{ProportionENEM}_{st}$. Columns (D)–(E) use the binary treatment variable, HighStakes_{st}.

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 math exam performance by topic area and competency

(A)	(B)	(C)	(D)	(E)	(F)
		Outcome: Proportion of correct answers			
		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 the impacts of ENEM adoption on performance in different topic areas and competencies on the math subject test.

The sample includes 2009–2017 ENEM exam takers in our analysis 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 for questions on the math subject test. Panel A presents results from a regression that includes all math questions. Panels B and C present results from separate regressions for each of 7 math topic area and 30 math competencies, as defined by ENEM test designers. Panel C displays only the top 5 and bottom 5 competencies based on the values in Column (F). See Appendix C.4 for details on math topic areas and compentencies.

Column (A) specifies the group of questions contained in each regression. Column (B) indicates 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 these 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.

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