Performance gaps in high-stakes testing: the role of textual context

Cristóbal Ruiz-Tagle

October 18, 2024

Click for the most recent version

Abstract

Standardized tests are critical for determining educational and professional opportunities, but often assume uniform performance across diverse groups. While much research has examined how testing environments contribute to performance gaps, the impact of textual content within test questions has received less attention. This paper investigates how the contextual features of questions in Brazil's ENEM—the second-largest college admission test globally—correlate with performance disparities related to socioeconomic status (SES), gender, and ethnicity. Using data from over 3.8 million senior high-school test-takers across 13 years (2010–2022), I analyze item-specific gaps and link them to the multidimensional space of words used in each question. Through bag-of-words and topic modeling combined with penalized regressions, I identify specific words and topics in the question text that strongly predict gaps. Hypotheses are generated independently by interpreting the common patterns in these words and topics, with the interpretation provided by ChatGPT. Six hypotheses are produced: two each for SES, gender, and ethnicity, focusing on widening and reducing channels of performance. These are tested using a rich set of fixed effects at the individual-question level. The results reveal that SES gaps increase by 1.4 percentage points (23% of the overall SES gap) when questions feature financial concepts, especially among higher-ability test-takers. Gender gaps widen by 1.1 percentage points (30% of the average gender gap) when questions are framed using abstract scientific contexts, but this effect only emerges among high-ability female test-takers and is not driven by domains where females generally underperform. The presence of female characters tends to offset the widening effect, whereas the presence of underprivileged characters amplifies the SES gap. Additionally, practical problem-solving scenarios narrow gender gaps by 0.6 percentage points across all ability levels. No significant textual features were found to affect ethnic gaps. These findings offer crucial insights for test design and suggest a data-driven approach to improving fairness in other testing contexts.

Keywords: standardized testing; question design; fairness; text analysis; performance gaps

JEL Codes: I23; I24; I31; D63.

I am deeply grateful to Pamela Giustinelli and Sarah Eichmeyer for their invaluable mentorship, patience, and guidance throughout this project and my entire PhD journey. I also acknowledge the financial support provided by ANID-Chile. This work has greatly benefited from the valuable feedback of Raphaelle Aulagnon. As always, the usual disclaimers apply.

Affiliation: Bocconi University. Department of Economics. Mail to cristobal.ruiztagle@unibocconi.it.

First version: August 26th, 2024.

1 Introduction

Standardized tests are widely employed to assess the abilities of college applicants, monitor the quality of schools, and identify the most suitable candidates for various job positions.[1] In most cases, standardized tests are high-stakes, often serving as gateways to better opportunities. Despite their widespread use, these systems are increasingly criticized by advocacy groups who argue that they fail to measure underlying ability and merely perpetuate the inequalities present in educational systems.[2]

While extensive literature in economics examines how various aspects of standardized test administration—such as time pressure, penalty for wrong answers, the ability being measured, among others—affect group disparities, relatively little attention has been given to the role of contextual content. This paper seeks to bridge that gap by exploring how the textual context used to frame a question (hereinafter, "item" and "question" will be used interchangeably) influences performance disparities across socioeconomic, gender, and ethnic groups. By "textual context", I refer to elements like language or terminology tied to specific backgrounds, framing, or character representation that triggers particular emotions or experiences. These contexts may evoke different interpretations or cognitive loads across groups, potentially contributing to performance disparities.

The setting for this research is the Brazilian college admission process, with a focus on the ENEM, the second-largest college admission test in the world. Each year, millions of Brazilians take this test, which consists of 180 multiple-choice questions evenly distributed across mathematics, language, social sciences, and natural sciences. ENEM scores are the sole criterion for admission to free public universities through the centralized SISU system, making it a highly competitive and high-stakes test. For instance, in 2016, over 9 million individuals registered to take the ENEM, around 3,5 million submitted preference by the SISU, but only 150,000 were offered a place (Otero et al., 2021).

In this paper, I proceed in three steps: first, I estimate question-level performance gaps; second, I systematically explore patterns in these gaps by linking the estimated gap to specific words or topics, generating hypotheses about their contextual drivers; and third, I test whether these patterns hold at the individual level to explore heterogeneous group responses. These steps are essential because, first, I need a metric to establish the gaps at a question-by-question level. Then, the second step is crucial for identifying potential sources of these gaps from the vast space of possible words and generating meaningful, testable hypotheses. However, given the complexity of the relationships and limitations such as sample size, the

[1] Currently, several countries adopt this practice, including the SAT and ACT (USA), A-levels (UK), Baccalauréate (France), PAU (Spain), Abitur (Germany), Gaokao (China), PAES (Chile), HSC (Australia), and USE (Russia). They are also used to screen international graduate students via the Graduate Record Examination (GRE) or to assess foreign language proficiency through exams like TOEFL or IELTS. Similarly, standardized assessments are used to license professionals in fields such as medicine and law, with examples including the bar exams in law, the United States Medical Licensing Examination (USMLE), and the Medico Interno Residente (MIR) in Spain. Additionally, standardized tests are utilized to compare educational systems across countries, with prominent examples being the National Assessment of Educational Progress (NAEP) in the United States, the Program for International Student Assessment (PISA), and the Trends in International Mathematics and Science Study (TIMSS). Finally, standardized tests are also frequently employed for hiring and screening candidates in competitive labor markets (Rudner, 1992; Schmidt and Hunter, 1998).

[2] A significant debate has recently emerged following the decision by several Ivy League-Plus universities to reinstate standardized test scores in their admissions processes after a pandemic-induced pause. Numerous advocacy organizations are actively engaged in this discussion, such as FairTest https://fairtest.org/higher-ed/.

method does not always have perfect predictive power, leaving room for the misidentification of drivers behind these gaps. Therefore, the third step becomes indispensable, as it rigorously tests whether the identified patterns persist at the individual level. This step ensures that the aggregate-level findings are not mechanically applied and helps account for the uncertainty, uncovering nuanced heterogeneous effects that cannot be captured through broader analyses.

In the first step, I estimate performance gaps for 2,201 questions using a sample of 3.8 million high-stakes test-takers from 2010–2022.[3] Gaps are estimated at the question level by regressing an indicator for having answered a question correctly on SES, gender, and conditional ethnicity dummies in separate OLS regressions. By construction, this approach estimates performance gaps that may conceal underlying drivers. These drivers can be uncovered by studying the relationship between the estimated gaps and various question characteristics, allowing for a deeper understanding of what contributes to these disparities.[4] Consistent with previous literature, low-SES students tend to underperform overall, though some questions show them outperforming, with similar patterns for ethnicity. Female test-takers lag behind male peers in all subjects except language, where gaps are slightly negative and centered around zero. On average, SES gaps are 6.8 p.p., gender gaps are 3.2 p.p., and conditional ethnicity gaps are 1.7 p.p., with a 32% correct-answer rate overall. When examining how much of the gaps can be explained by competence fixed effects—accounting for within-question variation, such as questions on probability versus algebra, and thus adjusting for domains that are more stereotyped or require higher-order abilities—and controlling for objective non-contextual measures like word count, sentiment score, presence of graphs or figures, or readability –which aim to account for differences in engagement or effort–, I find that only 41% of the SES gap, 35% of the gender gap, and 38% of the conditional ethnic gap are explained. This suggests there are additional factors, beyond these traditional metrics, that contribute to the observed disparities.

In the second step, I aim to generate hypotheses by investigating how specific topics and words appearing in the questions' text, alongside non-contextual factors and competencies fixed effects, correlate with the observed gaps by dimension. To achieve this, I use a two-step process: First, I estimate the impact of individual unigrams (single words) on the gaps using a Ridge regression model, identifying the unigrams that contribute most significantly to the observed gaps. Second, I apply Latent Dirichlet Allocation (LDA) to generate 114 topics that maximize the coherence score of the text, independent of performance gaps and based solely on the structure of the question text. Each question is represented by a distribution of topics derived from the words it contains. I then run Lasso regularization regressions for each dimension (gender, SES, and ethnicity), where the left-hand side consists of the estimated gaps observed for question i, and the right-hand side includes the distribution of topics characterizing that question, along with competence fixed effects and non-contextual features of i. By controlling for these factors, the method helps isolate the contextual channels driving the performance gaps. To generate the

[3]This refers to senior-high school test-takers who submitted all ENEM components so can apply to one slot in a public university. Figure A1 shows yearly shares across the full sample. The number of 2,201 questions corresponds to all questions that were asked in the time window in Portuguese that passed the inclusion criterion from the total of 2,340.

[4]Each question has three regressions. Ethnicity gaps are estimated conditional on being low-SES due to the high correlation (0.98) between SES and ethnicity gaps.

hypotheses, I feed ChatGPT both the top unigrams identified in the Ridge model and the composition of the top topics that emerged from the LDA. ChatGPT synthesizes these inputs to extract the common drivers of the widening and reducing channels for each dimension, generating one hypothesis per channel (widening or reducing) for each gap (gender, SES, ethnicity). This results in six testable hypotheses. I specifically ask ChatGPT to focus on non-competence-related elements to ensure the analysis captures contextual effects, rather than domain-specific factors. The results reveal that for gender gaps, widening topics involve abstract scientific concepts, while reducing topics focus on practical problem-solving. For SES gaps, financial issues exacerbate the gap, while everyday scenarios reduce it. For ethnicity gaps, abstract social phenomena widen the gap, while historical and cultural contexts mitigate it. Tagged questions are present across all subjects, and their classification cannot be predicted by non-contextual features or IRT factors, indicating that they are comparable to non-tagged questions in terms of these characteristics.

In the third step, I test the generated hypotheses using detailed individual-level data across all questions and subjects, incorporating extensive fixed effects—including individual fixed effects, competencies fixed effects, question difficulty, as well as region and year fixed effects. This step is crucial because, although the second step controls for various factors and identifies topics associated with performance gaps, the topics may be challenging to interpret accurately and prone to misclassification errors. By testing at the individual level, I can explore within-group heterogeneity (i.e., ability) and validate whether the relationships observed hold consistently across different contexts. This allows me to examine whether the cognitive processes underlying these performance gaps—such as how individuals respond to contextual cues—are consistent regardless of the specific subject matter, which ensures that the identified contextual factors are truly driving performance disparities. The results show strong support for both channels in the SES gap: questions in the widening channel increase the gap by 1.4 percentage points (23% of the SES gap), while items in the reducing channel decrease it by 0.8 percentage points, particularly for lower-ability test-takers. For the gender gap, only the reducing channel holds, narrowing the gap by 0.6 percentage points consistently across ability levels. The widening channel has a positive, unexpected sign, contradicting the hypothesis. When extending the methodology to assess possible spillover effects, I observe that all significant effects are localized, with no spillover to subsequent questions.

Heterogeneity analyses reveal important insights, particularly regarding differences by ability, subject, difficulty, the degree to which competencies are stereotyped, and the presence of characters that mirror the test-takers' characteristics—the "mirroring effect." First, the widening effects are more pronounced among higher-ability test-takers, a trend observed across both SES and gender gaps, though not in the reduction channel, which shows no clear pattern for SES and remains stable for gender. This intensification with ability occurs despite the overall rise in gaps, likely due to floor effects, where low-ability test-takers struggle with more difficult questions, limiting the potential for gaps to emerge.[5] Second, the widening effects for SES are present across all four subjects, both in the full sample and when focusing only on test-takers in the fourth quartile, while for gender, these effects only emerge among high-ability test-takers.

[5]If a question is too difficult relative to the test-taker's ability, no gaps can emerge since low-ability test-takers are unlikely to answer it correctly, regardless of the dimension considered.

Third, widening effects Fourth, tagging questions based on whether the competence involved is stereotyped—using sub-competencies that show the largest increases in the gender performance gap—reveals that most widening effects stem from stereotyped questions. Finally, for gender gaps, a mirroring effect occurs: when female test-takers encounter a female character, the widening effect is entirely offset, while the reduction channel remains unaffected. In contrast, for SES gaps, the mirroring effect amplifies the widening channel. These findings suggest that pairing gender-related items with female characters could mitigate widening effects, but no similar strategy exists for SES gaps.

Contributions. This paper contributes to three strands of the literature. First, it adds to the priming and performance literature (e.g.: (Lodder et al., 2019; Afridi et al., 2015; Hoff and Pandey, 2014; Fryer et al., 2008; Vohs et al., 2006)) by examining how different groups respond to textual context in a high-stakes, real-world setting. The closest comparison is Duquennois (2022), who finds that fictional money widens SES gaps by 1.2 percentage points in TIMSS,[6] a international cross-country standardized tests which is a low-stakes setting. While her study relies on manually tagging monetary items, this paper uses a data-driven approach, finding a slightly larger SES gap (1.4 percentage points) in ENEM, a high-stakes exam. Moreover, Duquennois (2022) identifies spillovers across four consecutive items, which she attributes to attention capture—where one item influences subsequent performance due to cognitive focus on a particular context. In contrast, the ENEM effects are localized, suggesting that high-stakes incentives limit these spillovers. This difference underscores how context matters and how the incentives that test-takers face in different settings affect their performance. More broadly, the method proposed in this paper demonstrates the potential of using test booklet text analysis to uncover disparities in performance across various assessments. It could be extended to hiring tests or certification exams, providing insights into priming effects in other high-stakes environments, with important implications for equity and fairness in the composition and representation of disadvantaged groups in key areas such as education, the workforce, and professional qualifications.[7]

Second, this paper contributes to the literature on stereotypes and their consequences in high-stakes contexts. Extensive research has demonstrated how stereotypes shape real-world outcomes, such as high school track choices (Buser et al., 2014), college major decisions with different returns (Zanella, 2021), job applications and wage expectations (Kiessling et al., 2024), hiring decisions by managers (Coffman et al., 2021), and promotions (Roussille, 2024), as well as group decision-making (Coffman, 2014). However, less is known about how stereotypes affect test performance. A relevant but inconclusive body of work has explored the role of stereotype threat in testing (Steele and Aronson, 1995; Schmader et al., 2015),[8] with concerns about underpowered samples and publication biases (Priest et al., 2024; Shewach et al., 2019;

[6]Trends in International Mathematics and Science Study (TIMSS). More information in https://timssandpirls.bc.edu/timss-landing.html

[7]Recently, the U.S. Department of Justice found that hiring tests used by the Maryland Department of State Police and Durham's Fire Department disproportionately disqualified female and non-white candidates, reducing workforce diversity and undermining public safety. See more at https://www.justice.gov/opa/pr/justice-department-secures-agreement-maryland-department-state-police-resolve-allegations and https://www.justice.gov/opa/pr/justice-department-secures-agreement-durham-north-carolina-end-discriminatory-hiring.

[8]This phenomenon is not limited to cognitive domains, as it has been documented in sports (Beilock and McConnell, 2004) and entrepreneurship (Zhang et al., 2023).

Flore et al., 2018), particularly in high-stakes settings (Shewach et al., 2019; Cullen et al., 2006). This paper adds to this literature by showing that stereotype threats can emerge even in real-world situations without social monitoring or public exposure of individual performance, through the mechanism of self-stereotyping. In a lab experiment, Coffman (2014) find that individuals are less willing to contribute ideas in domains stereotypically outside their gender, driven by self-assessments rather than fear of discrimination. The present study's findings can also be interpreted through the lens of self-assessments, extending the evidence of these effects to the educational testing setting. Moreover, this paper enhances understanding of where these disparities arise, showing how stereotype reminders particularly affect test-takers with higher latent ability (i.e., those more identified with the domain, as predicted by theory (Smith and White, 2001)). Additionally, it provides evidence that some negative effects can be mitigated through a mirroring effect: gender-related widening effects are reduced when items feature a female character, while they increase when items remind low-SES individuals of their identity.

Finally, this paper contributes to the growing literature on testing and fairness. Research in this area is divided between studies conducted in low-stakes settings—such as global academic tests like PISA and TIMSS—and those in high-stakes environments, where the stakes themselves can exacerbate disparities. High-stakes contexts often magnify inequalities due to differences in stress management, effort, and preparation. For example, Cai et al. (2019) show that gender gaps widen when comparing mock versus actual Gaokao exams, while Attali et al. (2018) find larger gender and ethnic gaps in the high-stakes GRE exam compared to voluntary practice tests, attributing this to differences in effort. In the Brazilian context, Reyes et al. (2023) explore how varying stakes in the ENEM exam affect performance gaps, while Ofek-Shanny (2024) emphasize the sensitivity of performance gaps to stakes more broadly. Among studies conducted in low-stakes settings (Duquennois, 2022; Griselda, 2022; Brown et al., 2022; Cai et al., 2019; Baldiga, 2014; Muskens et al., 2024; Ofek-Shanny, 2024; Anaya et al., 2022), this paper extends the discussion by showing the importance of textual context and the need for neutral test design, an often overlooked factor in generating disparities. In relation to studies in high-stakes settings (Cohen et al., 2023; Attali et al., 2018; Goodman et al., 2020; Cai et al., 2019; Ebenstein et al., 2016; Coffman and Klinowski, 2020; Franco and Povea, 2024), this paper broadens the scope by demonstrating that disparities also emerge outside mathematics, offering the first evidence of question-by-question variation rather than focusing on overall score effects. Additionally, it highlights how minor contextual variations, such as the saliency of textual context, can significantly affect performance, particularly in high-pressure environments, leading to crucial consequences for access to better opportunities. This has far-reaching implications, as shown by Landaud et al. (2024), who find that Norwegian students randomly assigned to subjects in which they excel attend better college programs and earn higher wages.

The structure of the paper is as follows: Section 2 provides an overview of the Brazilian context and the ENEM admission test. Section 3 outlines the dataset and sample selection. Section 4 details the empirical strategy for generating and testing the hypotheses. Section 5 presents the findings and explores heterogeneity. Finally, Section 6 offers concluding remarks.

2 Background

Brazil has several features that make it an ideal setting for this analysis. It is home to 203 million people, with a real GDP per capita of USD 8,802 in 2022, and is a large federal country with 26 states and over 5,500 municipalities. Economic, social, ethnic, and geographic disparities are significant, with important inter-generational consequences (Pinotti et al., 2022). As highlighted by the most recent Wold Bank country report, despite Brazil's diversity, systemic ethnic and gender discrimination limit opportunities for many, perpetuating intergenerational poverty. In addition, Afro-Brazilians and Indigenous Peoples face reduced access to quality education and healthcare compared to whites, while women encounter substantial job discrimination, limiting their earning potential. Rural areas suffer from pervasive inequalities in accessing public services, hampering investments in human capital. Although there have been improvements in youth literacy, healthcare, and essential services, the wealthiest 1% of Brazilians own 32.2% of the nation's wealth, and a Gini coefficient of 0.518 highlights Brazil's position as one of the most unequal countries globally (World Bank, 2024).

2.1 Brazilian ENEM: a high-stakes setting

Established in 1998, the National High School Examination (ENEM) initially functioned as an assessment tool for the certification of high school completion. However, starting in 2009, it was expanded to also serve as the primary examination for admission into higher education institutions.[9] Figure A2 shows the share of enrollees in public universities that were admitted by the ENEM exam yearly. This means that every year the test serves a dual purpose: on the one hand, test takers of various ages who want to apply and pursue degrees in top public universities that participate in the centralized admission system, or in other universities that consider ENEM scores, and on the other hand, high school graduates who want to certify their degree.

Public universities in Brazil are analogous to flagship state universities in the United States, often being the most prestigious, elite, and selective institutions within their respective states. These universities are tuition-free and generally offer higher quality education than most private institutions, making them highly attractive to top-performing students from both low- and high-socioeconomic status (SES).

The ENEM admission test is massive, being the second largest in the world after China's National Higher Education Entrance Examination. While the number of test-takers varies each year, a relatively stable share of senior high school students take the test. Figure A1 shows the share of senior high school students who participate each year in relation to the total number of test-takers. The total number of test-takers fluctuates significantly year by year, always remaining above 4 million, while the number of senior-high school test-takers remains stable at around 500,000 annually.

When test takers receive their scores, they apply to post-secondary programs, most of which are

[9] This change involved modifications to both the content and length of the exam. The Ministry of Education designated ENEM as the entry exam for public universities, and these universities adopted it gradually and voluntarily. Reyes et al. (2023) and Machado and Szerman (2021) describe the implementation phase. Although some universities chose not to participate in the centralized admission system, they still considered ENEM scores in their independent entry assessments.

college degrees at both private and public institutions. To apply to the 23 most prestigious tuition-free universities, which depend on federal and national government funding, they must participate in the centralized clearinghouse system called SISU, as described in Machado and Szerman (2021). The only input considered by this system is the results in the subjects of the ENEM test. Students' priority scores are calculated using a weighted average of their test scores and degree-specific weights for each of the ENEM components. The clearinghouse system operates based on a deferral algorithm, ensuring that all students above the cut-off compete with the same probability, leaving no room for discretion in slot allocation that might benefit certain applicants. Because of the tuition-free policy, along with the selectivity and quality of the programs, college degrees from these institutions offer high private returns (Binelli and Menezes-Filho, 2019; Duryea et al., 2023).

Students applying to non-tuition-free public universities still face a high-stakes test, as most other institutions consider the ENEM results in their assessments. Additionally, the ENEM results are used to rank applicants for financial aid programs to finance private tuition fees, such as the University for All Program (ProUni) and the Student Financing Fund (FIES).

Additionally, since 2000, affirmative action policies have been in place, with full enforcement since 2018, favoring marginalized groups (Vaz, 2020; Otero et al., 2021). This aspect is particularly appealing to candidates from disadvantaged contexts.

ENEM is a test where significant inequalities emerge, particularly in the upper end of the performance distribution. As shown in Figure A3, in the top percentiles, the share of women, low-SES individuals, and non-whites is far below their representation in the overall test-taker population. While females are 57% of the test-takers, they fall to 40% among top performers. Similarly, non-white test-takers, who comprise 56% of the sample, drop to 30% in the top performers, and low-SES individuals, representing 27% of the sample, fall to less than 5% in the top performance percentile. This distortion is highly consequential, as it is precisely in these top percentiles where surpassing the cutoff is necessary for admission to more competitive college programs. Such disparities are well-documented in the literature, with several studies highlighting how unequal representation at the higher end of test performance can reinforce broader social inequalities in contexts like the SAT or ACT (Bordalo et al., 2016; Hyde et al., 2008; Freedle, 2003). Moreover, understanding the roots of these gaps—where they emerge, what drives them, and why they persist—is essential for creating informed interventions. Table A1 shows summary statistics in terms of gaps, in standardized deviation points, across different settings, suggesting that they are transversely present for these three dimensions in developing countries such as Chile, Colombia, and Brazil as well. The SES gap is a consistent finding, while the gender gap varies by country, with nations such as China, Finland, Greece, and Spain where girls on average outperform boys. Addressing these disparities and understanding their origins is critical for ensuring equitable access to higher education opportunities.

2.2 ENEM Features

The ENEM test is performed in two days, each day covering two subjects. From 2010 to 2016 the schedule was Social Sciences and Natural Sciences during the first day, and Language and Math in the second.

This switches in 2017 to be Language and Social Sciences in the first day. Apart from this, the design of the booklets are similar since 2010.

The ENEM covers four subjects: language, mathematics, social sciences, and natural sciences, plus a handwritten short 30-lines essay on a proposed topic equal for all test-takers. The tests are graded from 0 to 1000 using Item Response Theory (IRT) to determine the final score, and the marking of the multiple-choice part is entirely automated, without human interference.[10] For each subject, individuals receive a randomly assigned booklet from four daily available options. These booklets contain the same set of questions, but the order varies. All test-takers face the same items; only the sequence changes. There are no penalties for incorrect answers, so the incentive is to avoid leaving any questions unanswered.[11] Importantly, this test does not include adaptive sections based on previous performance in specific sections, unlike, for instance, the GRE.

Each subject has 45 questions, totaling 180 questions. The test is administered over two days: the first day spans 4.5 hours, and the second day spans 5.5 hours (the extra time is due to the handwritten essay). On average, each question is expected to be answered in 3 minutes. The exam takes place at the end of the academic year, in December, and its scores are only valid for that specific application process. To take the test, a fee of around USD 20 is required, but students from public schools and poorer households receive a waiver. The booklets include space for calculations, and only the final marked alternative is graded.

Each item in the ENEM belongs to repository of items proposed by experts and piloted in representative samples. The goals of each items is clearly and publicly ordered by the Matrix of Reference. This is a document that defines the categories of competences. [12] Each subject encompasses 6 to 9 areas that distribute 30 specific competences, which are targeted by the test items. The public data includes the code for each of the specific competences attributed to each item by the designers.

Questions in the ENEM exam follow a common structure: first, a heading, then an assignment, and finally, five alternatives to choose from (labeled from a to e). The heading usually includes the context of the question and a description, but sometimes it can be just an image or a graph. Assignments are typically short and direct, precisely describing what is being asked. Individuals need to gather information from the heading to complete the task and then choose the correct alternative from the provided options. The alternatives are displayed in the same order regardless of the booklet. As expected, some options are included to distract the applicant. The average length of the heading and assignment is 103 words. There is no optimal strategy for tackling these problems, but one potential approach is to skip the heading initially, start with the assignment, and then refer back to the heading strategically to minimize time.[13] Regardless of the booklet, the order of the alternatives remains the same. Figure A5 shows examples of

[10]The grading model is based on a 3PL model, where the three parameters—guessing, discrimination, and difficulty—are calibrated out-of-sample through a pilot. All parameters are freely available for each question, allowing the θ individual parameter for latent ability to be estimated. Section A.5 describes the grading model.

[11]This incentive is effective: less than 0.2% of the answers are left blank.

[12]In Section A.4 there is a description of the content they seek to measure by type.

[13]Note, however, that this procedure might require special meta-cognitive abilities that may differ across demographic groups.

these questions, illustrating the variation in topics as well as differences in item length, the use of resources such as character mentions, and references to real-world units like money. While Figure A6 presents the word cloud for each subject, illustrating not only the breadth of content but also how subjects overlap despite their distinct domains. In this context, subjects are not isolated silos; for instance, topics like technology can appear across different subjects. This overlap provides a unique opportunity to test how content influences performance independently of the subject domain.

3 Data

My data spans 13 years, from 2010 to 2022.[14] It is composed of three datasets. The first one is the dataset that includes individual performance in all the items and individual level characteristics such as age, type of high-school, gender, income, ethical self identification, among other. Year by year, these files include the information of all test-takers enrolled to perform the test, with the specific string of responses they provided for each of the 180 items that they faced in each of the booklets assigned. [15] Importantly this data set does not include an ID for each item, but it includes the correlational sequence in which the booklet was responded.[16] It also includes the IRT score in each of the four subject and the information about the assessment in the essay section. The second data set is the one that includes the information about each of the items, where the solution key is provided, the ID for each item, the subject to which the item belongs, the competence to which the item belongs based on the reference matrix and the parameters a, b, c necessary to identify the characteristic curve of each item (Figure A21 shows a general example).[17] Finally, the third data set is based on scrapped text from the booklets, from where I distinguish three elements: the heading, the assignment and the alternatives. [18] I also labeled items that included figures or graphs and tables.

3.1 Sample Selection

I center my analysis on three key dimensions, chosen for their relevance and based on existing literature, which demonstrates that these dimensions are particularly susceptible to disparities in standardized tests.[19] These dimensions are defined dichotomously across gender, socioeconomic status (SES), and self-reported ethnicity. Gender is coded 1 if the test-taker is female, low SES is coded 1 if the individual

[14] While the ENEM is available as entrance exam since 2009, I preferred to not consider this year for two reasons: first, as depicted in Figure A2, the share of admitted test takers by the ENEM in public universities was small, and second, because after an attempt of fraud, the test had to be postponed for 2 months with a high abstention rate of 40.6%.

[15] All of the information is freely available at: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem. Unfortunately with the information available in the public access dataset, it is not possible to identify individual across different years, so I am not able to identify test repeaters.

[16] This information does not include a good classifier for rural status, so based on the municipality that is imputed to the high-school I create a variable of rurality based on the Census information by IBGE.

[17] These parameters are also helpful to identify which items do not meet the convergence to the characteristic curve, and therefore where considered anomalous. The inclusion of these items might generate biases in my estimations

[18] As all booklets show the same set of items, but in different order, to identify which item is matched with its specific text, it is only necessary to scrape one booklet. I proceed systematically always with the order of the blue one for each year.

[19] In Figure A3 There is a clear indication of how harmful inequalities in these dimensions of interest are mainly in the top percentiles, those who are most likely to be admitted to quality public and free universities.

reports to earn less than 1 minimum wage, and the ethnicity+ variable is coded 1 if the test-taker identifies either *pardo* (mestizo) or *preta* (black). In my dataset, 57% of respondents are women, 28% are classified as low-SES, and 55% self-identify as *pardo* (mestizo) or *preta* (black). However, since SES and ethnic gaps are highly correlated (as shown in Figure A4), suggesting that they may capture different aspects of the same underlying disadvantage, I concentrate on gaps observed across race, conditional on being low-SES. Within this subset, 26% still identify as white. This approach allows me to isolate textual triggers that specifically contribute to ethnicity gaps, adding an extra layer of insight into disparities within an already underprivileged group.

To focus the sample on individuals in high-stakes situations, I impose several restrictions. First, I retain only high school seniors, ensuring that I compare individuals from the same cohort, which mitigates self-selection issues related to older test-takers and controls for variations in curriculum structure. Second, I include only those who followed the traditional academic track, excluding those in alternative tracks designed for special needs or adult education. Third, I exclude individuals who either failed to attend one of the test days or did not submit their handwritten essay. Missing one part of the test renders the application process incomplete, and such absences may reflect private information about performance that should not be factored in.[20] Additionally, I exclude booklets assigned to test-takers with special accommodations (e.g., larger font sizes) to maintain comparability.

This is a highly selective sample, justified by the goal of measuring the impact of contextual cues in high-stakes situations. Students who aspire to attend a high-quality public university in Brazil and who completed all parts of the test may not be fully representative of their broader demographic group, likely reflecting a more resilient subset. It's also important to note that I do not have access to any pre-existing ability measures, such as performance in other standardized tests or high school grades, which limits my ability to control for prior academic ability outside of ENEM performance.

4 Empirical Strategy

This section details the empirical strategy employed in this study. First, I describe the dataset and the criteria for sample selection, which includes cohorts from 2010 to 2022 and individuals in high-stakes testing situations covering around 3.8 million individuals. The empirical approach unfolds in three steps. In the first step, I estimate item-specific performance gaps across gender, socioeconomic status (SES), and (conditional) ethnicity for all individuals in the selected sample. Next, I map these estimated gaps to textual characteristics of the test items, utilizing vectorized text analysis to generate hypotheses about the factors driving the observed disparities. Finally, in the third step, I test these hypotheses using a random sample of 10,000 individuals per year, analyzing the 175 items they encountered and exploring both the widening and reducing effects on performance gaps.

[20] This is also important because, as highlighted in Reyes et al. (2023), some high school seniors take the test without intending to apply to college.

4.1 Item-by-item gaps as an input for hypotheses generation

To generate testable hypotheses about the drivers of performance gaps from individual item-level data, it is essential to aggregate the information to capture differences across dimensions. My approach involves several steps, which I outline below. The key idea is to map the item-by-item estimated gaps—measured as the difference from a reference group—onto objective features and vectorized text. This allows for the identification of factors with the greatest predictive power in explaining the gaps, both widening and reducing them. The top predictors associated with textual content are then translated into testable hypotheses, which are subsequently tested at the individual item level by tagging those items that exhibit the characteristics described in each hypothesis. Finally, the use of individual-level data enables the exploration of various sources of heterogeneity, which may help disentangle the underlying mechanisms driving these performance gaps.

4.1.1 Measuring item-by-item gaps

To measure the performance gaps across different demographic dimensions, I estimate item-specific gaps using the following regression model:

$$Y_i^q = \alpha_0^{q,r} + \alpha_1^{q,r} \cdot \mathrm{D}_i^r + \tau_b + \epsilon_i \tag{1}$$

where Y_i^q represents the performance outcome for individual i, $i = (1, \ldots, N)$ on item q, $q = (1, \ldots, Q)$, being 1 if the test-taker had the answer correct, and zero otherwise. [21] The coefficient $\alpha_1^{q,r}$ captures the effect of being in the category of interest for the demographic characteristic $r \in \{\text{Gender, SES, Ethnicity}\}$ on the performance for question q only. The term τ_b controls for the randomly assigned test booklet $d \in \{\text{red, yellow, blue, purple}\}$, which display the same set of questions but in different order, ensuring that any systematic variation due to booklet allocation is accounted for. This approach allows me to estimate item-by-item gaps, where $\widehat{\alpha}_1^{q,r}$ represents the specific performance gap for each dimension r across individual test items.

Then I merge these gaps with item's characteristics that likely might explain groups disparities such as: readability[22], length of the text measured as number of words, the use of figures, tables, the sentiment score[23] each item generates and if the text is grounded or not.[24] In my setting, these features are not considered contextual text factors as they are not expected to differentially affect specific groups based

[21] This implies that missing and double marked responses are coded as incorrect. The presence of these two possibilities is quite low in my sample, not exceeding 0.5%. The disadvantage of this is that it does not allow studying differences in effort or guessing through variations in this outcome.

[22] I estimate the a readability index based on the composition of four indexes using Anderson (2008) method: Flesch-Kincaid, Gunning fog, Automated Readability Index and Coleman-Liau. The measures have been manipulated so a greater value means a greater ease of reading. The calibration of the indexes from English to Portuguese has been done using Carvalho de Lima Moreno et al. (2022).

[23] The sentiment score has been estimated using the multilingual model BERT. It consider the probability of having a very good sentiment.

[24] This is a dummy variable hand-coded by myself following the definition by Koedinger et al. (2008) where grounded context are linked with real-world situations closer to test-takers experiences.

on the cues they evoke.

Table A2 presents the distribution of these topics by subject while Table A3 presents a summary statistics table showing how these characteristics vary by subject.[25] Importantly, when regressing the gaps estimated using this method with yearly fixed effects, there are no significant yearly trends, as shown in Figure A7. This suggests that test designers are not systematically adjusting the pool of questions to reduce these gaps, a key factor that supports the validity of my identification strategy.

4.1.2 Text-based hypotheses generation

The next step in the analysis involves mapping the estimated performance gaps to specific item features (e.g., sentiment score, number of words, presence of figures) and vectorized text data to identify the most predictive words and topics using regularization regressions. The use of regularization regressions, such as Ridge and Lasso, is particularly useful in this context as it allows me to identify the most relevant words or topics from a high-dimensional space, which includes all the words appearing in all items. By imposing a penalty on the size of the coefficients, these methods effectively shrink less important variables to zero, allowing the model to focus on the collection of words and topics that have the greatest explanatory power for predicting performance gaps. This approach helps reveal the key drivers of gap variation, both in terms of widening and reducing effects, while avoiding overfitting to noise in the data.

Text Processing. I begin by processing the text using natural language processing (NLP) techniques with SpaCy, specifically its Portuguese language model. This step includes lowercasing, tokenization, lemmatization, stop-word removal, and TF-IDF vectorization[26], ensuring that the text data is appropriately formatted for further analysis. The total size of the text, after this processing is 130,215 words.

Regularization Regression. For the regression models, I apply two different approaches: the Bag-of-Words model combined with Ridge regression and LDA (Latent Dirichlet Allocation) topic modeling combined with Lasso regression. Ridge regression is particularly suitable for the Bag-of-Words approach due to the high dimensionality of the data—each word is treated as a separate feature, leading to a large number of predictors. Ridge regression's ability to handle multicollinearity by imposing a penalty on the size of the coefficients makes it ideal for this high-dimensional setting. In contrast, Lasso regression is more appropriate for LDA topic modeling because it performs variable selection by shrinking some coefficients to zero, which helps identify the most relevant topics in a more parsimonious model.[27]

[25] There are 2,275 items available over the 13-year period. However, I focus my analysis on a subset of 2,201 items. I exclude 74 items because they exhibit characteristics that could bias my estimates. Specifically, these items fall outside the 99th percentile in terms of the b parameter (difficulty), the proportion of individuals whose parents have a college education, or because they failed to converge to the characteristic curve according to the IRT models.

[26] TF-IDF (Term Frequency-Inverse Document Frequency) vectorization transforms text into numerical values by calculating the importance of each word within a document relative to a collection of documents. It assigns higher weights to terms that are frequent in a document but rare across the corpus, enabling effective text analysis and feature extraction for machine learning models.

[27] The advantage of using L2 (Ridge) over L1 (Lasso) models in this context is that Lasso models have a higher tendency to shrink coefficients for words to zero, especially when compared to other features. This is problematic because the goal is to identify the impact of specific words while controlling for other features, rather than excluding them entirely.

Topic Modeling. The LDA topic modeling follows a data-driven approach, where the number of topics is determined by optimizing the coherence score, a metric used to assess the interpretability of the topics. Each topic represents a distribution of words, and each test item is characterized as a mixture of topics. Topics are abstract constructs that group together words frequently appearing in similar contexts, providing insights into the latent structure of the text. By identifying these word patterns, LDA allows me to uncover latent topics that can explain part of the variation in performance gaps However, interpreting these topics is notoriously difficult, as the meaning of a topic is often unclear and may not correspond to an easily recognizable concept. This difficulty justifies the use of an alternative approach for interpreting the regression outputs.

Interpretation and Hypotheses Generation. To address the challenge of interpreting these complex results that include words and distribution of words, I employ ChatGPT to mimic the role of a research assistant. ChatGPT synthesizes the regression results, generating testable hypotheses by combining both the top words from the Bag-of-Words model and the key topics identified through LDA. This approach allows me to systematically identify hypotheses about which textual features are likely to drive performance gaps in both directions (widening or reducing them). The exact prompt I provided to GPT is described in Figure A8.

Machine learning (ML) techniques offer the potential to uncover patterns that might be overlooked by human analysis (Ludwig and Mullainathan, 2024; Wang et al., 2023). However, the use of these tools for discovery often comes at the expense of interpretability and clarity, as noted by Messeri and Crockett (2024); Batista and Ross (2024). OpenAI's GPT-4 plays a crucial role in addressing this challenge, providing an efficient way to process text data and generate coherent hypotheses, as demonstrated in recent studies (Charness et al., 2023; Batista and Ross, 2024). By leveraging GPT-4, I am able to synthesize complex regression outputs and translate them into testable hypotheses in a way that balances the advantages of ML with the need for interpretability, avoiding the typical trade-offs that arise when using black-box models for exploratory research.

Flagging the Items. To support the testing of these hypotheses, I used an independent prompt to ask GPT-4 to generate a list of keywords corresponding to each hypothesis after providing it with the exact phrasing of each hypothesis .[28] I then utilized these keyword lists to classify each test item based on the occurrence of at least two keywords within the item, minimizing the risk of false positives. Manual cross-checks of the tagged items suggest that the performance of this automated labeling process is sufficiently accurate for the subsequent analysis.

After completing the tagging process, a total of 1,098 items were categorized across the six defined

[28]For example, for the hypothesis related to widening the gender gap, the generated list includes terms like "transportation," "technology," "speed," "theory," "balance," "volume," and "produce," among others. While these words may not fully capture the contextual nuances and could also reflect domain-specific characteristics (such as motivation or ability in areas where these words are more prevalent), the identified topics still emerge after controlling for competencies fixed effects. To address concerns that these topics may reflect domain differences rather than contextual effects, I conduct heterogeneity analyses, demonstrating that the contextual channel remains relevant even in competencies where gender gaps are smaller, suggesting that the impact of context extends beyond specific domain characteristics.

categories. However, there is some overlap, with 40% of the items being labeled in two or more categories. Table A2 presents the distribution of items by tag. While the majority of items align with theoretical expectations—such as widening gender gaps in science and mathematics—a significant portion also appears in other areas. Additionally, Table A4 illustrates how item features predict their labeling. It is observed that these flagged items tend to be lengthier and exhibit worse readability. However, this trend does not differ between items predicting a widening gap and those predicting a reduction. Importantly, there are no statistically significant relationships with the IRT parameters. When such relationships exist, they are associated with less difficult items (p.B). This suggest that these items are not harder and with a similar discriminatory capacity.

4.2 Testing the Hypotheses within-individual performance

Having described the general relationship between performance gaps and item characteristics, the next step is to examine how the patterns discovered in the previous section affect individual performance in a more causal way. As I previously described, the analysis of performance gaps was conducted using all observations in each year that meet the inclusion criteria (i.e., individuals in high-stakes situations). However, given that hundreds of thousands of test-takers take the ENEM each year and each test-taker faces 180 items, the "curse of dimensionality" becomes a constraint, making it computationally infeasible to analyze the full dataset. Therefore, to proceed efficiently, I draw a random sample of 10,000 individuals per year from those meeting the inclusion criteria, resulting in a total sample size of 130,000 individuals.[29] This sample size is large enough to ensure sufficient statistical power to detect even small effect sizes. Table A5 provides the summary statistics of the individuals sampled.

To exploit within-individual differences in performance when a flagged item appears, I leverage the panel structure of the data. Each year, test-takers encounter 180 items sequentially, divided into blocks and by subject. The order in which the items are presented is determined by the random assignment of the booklets. Assuming that test-takers progress through the items sequentially—a reasonable assumption given the paper format—the appearance of a flagged item can be treated as quasi-random. This setup effectively creates a natural experiment, where flagged items may appear earlier for some individuals and later for others, allowing me to examine the causal effect of flagged items on performance.

In addition, the empirical strategy controls for a range of fixed effects: order-item fixed effects, competence fixed effects (based on the reference matrix), regional fixed effects to account for geographic disparities, and year fixed effects to capture cohort differences such as changes in curriculum or test coverage. I also include several item-level controls, such as readability, length, presence of figures or tables, sentiment score, grounding, and difficulty quartile fixed effects, to refine the analysis further. The following is the main specification used for this analysis:

$$Y_{iq} = \gamma_1^r \cdot (Widen_q^r \times D_i^r) + \gamma_2^r \cdot (Reduce_q^r \times D_i^r) + \gamma_3^r \cdot Widen_q^r + \gamma_4^r \cdot Reduce_q^r + \gamma_5^r \cdot X_q + \phi_i^r + \lambda_c^r + \theta_k^r + \psi_p^r + \eta_e^r + \delta_y^r + \varepsilon_{iq}^r \quad (2)$$

[29] The seed for each year's random draw was selected using www.random.org.

where Y_{iq} represents whether individual i correctly answered item q, and $Widen_q^r$ and $Reduce_q^r$ are indicators for whether the flagged item is predicted to widen or reduce the gap in a given dimension r, respectively. D_i^r is an indicator for whether the individual belongs to the group of interest (e.g., low SES, female, non-white), and X_q is a vector of item-specific characteristics (e.g., readability, number of words, figures, tables). The terms ϕ_i, λ_c, θ_k, and ψ_p represent individual-subject fixed effects, item competence fixed effects, item difficulty fixed effects, and item position fixed effects, respectively. Additionally, η_e accounts for regional differences, and δ_y controls for year-specific cohort effects. Standard errors are clustered at the individual level to account for within-individual correlation in responses.

The primary parameters of interest are γ_1 and γ_2, which measure the impact of flagged items on performance gaps for the group of interest. Specifically, γ_1 captures the average performance gap induced by flagged items expected to widen gaps, while γ_2 captures the effect for flagged items expected to reduce gaps. Importantly, the identification of these effects relies on the assumption that flagged items are distributed randomly across individuals due to the randomized booklet allocation, allowing for quasi-experimental conditions.

This approach allows for a robust estimation of the impact of flagged items on individual performance while controlling for a wide array of potential confounders. The inclusion of multiple fixed effects and item-level controls helps to ensure that the observed effects are not driven by unobserved factors related to the test-taker, item characteristics, or test administration.

5 Results

5.1 Unexplained Variance and the Role of Contextual Effects

In this sub-section, I examine the performance gaps across SES, gender, and conditional ethnicity by progressively incorporating various explanatory variables. This exploratory funnel approach begins by documenting gaps across dimensions and subjects, and then narrows down to analyze how competencies, IRT parameters, and non-contextual features contribute to explaining these gaps.

As shown in Table A6, when accounting for all regressors to explain the question-by-question gaps, they explain 41% of the SES gap variance, 35% of the gender gap variance, and 37% of the conditional ethnic gap variance. Gender gaps appear to be particularly sensitive to competencies and non-contextual features, while SES and ethnic gaps are more influenced by the structure of the test, with IRT parameters alone explaining 23.7% and 28.1% of these gaps, respectively while only 9.8% for gender.

Despite being informative, these factors cannot account for all the variation in the performance gaps, suggesting that there is room for other factors, such as contextual elements, to drive the gaps.

5.1.1 Raw Gaps by Dimension and Subject

Across dimensions and subjects, there is significant variation in the distribution of raw gaps. Figure A9 presents the kernel density estimates for each of the three dimensions by subject. Consistent with previous literature, low-SES students underperform overall, though some questions show them outperforming

others, with similar patterns observed for ethnic minorities. Female test-takers lag behind their male peers in all subjects except language, where the gaps are slightly negative and centered around zero. On average, SES gaps are 6.8 percentage points (p.p.), gender gaps are 3.2 p.p., and conditional ethnicity gaps are 1.7 p.p., with an overall correct-answer rate of 32%.

For the gender gap, the distribution in language is centered around zero, showing no substantial bias. However, in social sciences, natural sciences, and math, males tend to outperform females, with the skew particularly strong in social sciences. This may be due to the presence of more abstract or technical questions in social sciences, as indicated by the reference matrix.

For SES gaps, the left-skewed distribution indicates that low-SES test-takers generally underperform compared to their higher-SES peers. However, in 11% of items, particularly in language (31%), low-SES students outperform others. The variation across subjects suggests that accessibility to content may differ due to factors like exposure or cultural capital, but there is still something to discover about the specific characteristics of the questions where low-SES students outperform.

For conditional ethnic gaps, the distribution is narrower, reflecting the influence of SES. However, disparities persist, especially in social sciences and math, suggesting that ethnic differences may be driven by cultural relevance or representation in test content, even after controlling for SES.

5.1.2 Competencies Fixed Effects and Raw Gaps

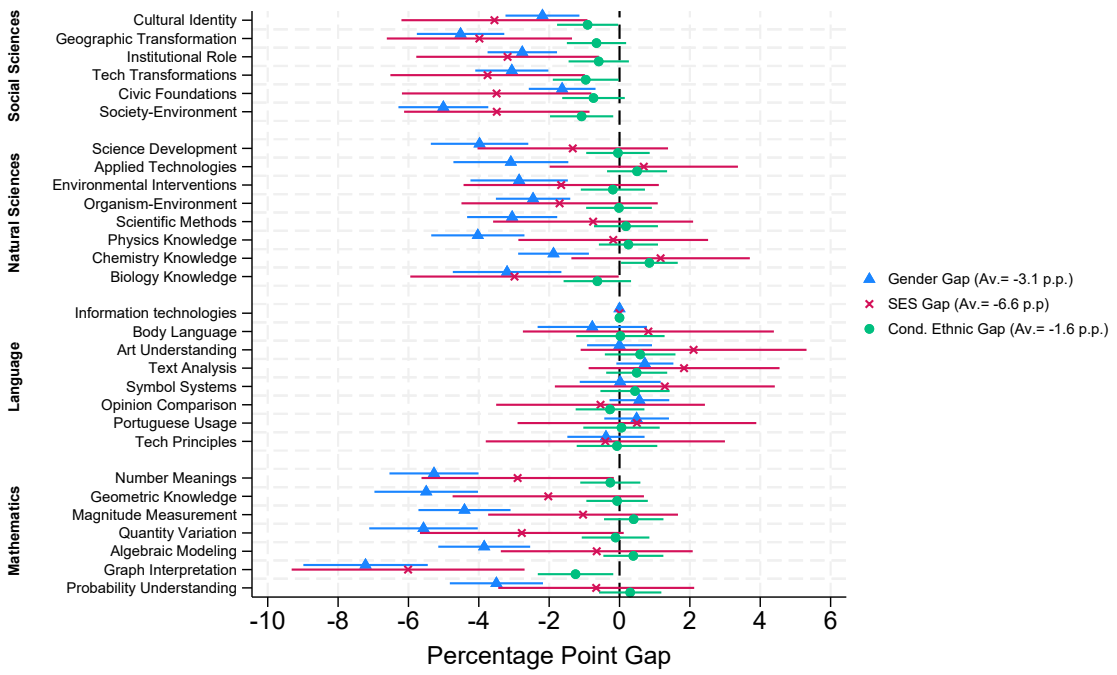
The next layer of analysis examines the extent to which performance gaps can be explained by the competence to which each question belongs. Questions are categorized into a single competence based on the criteria set by the test designers, which reflect the specific aspect or ability the question is intended to measure.

Figure 1 shows how each competence fixed effect impacts gender gaps (blue triangles), socioeconomic gaps (red crosses), and conditional ethnic gaps (green dots) across a wide range of competences in Social Science (SC), Natural Science (NS), Language (LG), and Math (MT). Competencies fixed effects alone account for 24% of the gender gap, 11% of the SES gap variance, and 7.5% of the conditional ethnic gap variance.

Gender gaps (blue triangles) are more sensitive to certain competencies. For instance, in Social Sciences, “Society-Environment” contributes the most, widening the gap by around 5 percentage points, followed by “Geographic Transformation.” In Natural Sciences, “Science Development” and “Physics” have the largest impact, and in Math, ”Graph Interpretation” contributes significantly. This pattern may reflect the susceptibility of societal stereotypes on gender performance.

For SES gaps (red crosses), the effect of competencies varies, but wide confidence intervals suggest significant heterogeneity within competencies. This could indicate that SES gaps are more influenced by question-specific characteristics, such as difficulty or discrimination, than by cross-competence differences.

Conditional ethnic gaps (green dots) cluster near zero, with most not statistically significant, suggesting that ethnic disparities are not primarily driven by differences in competence content.



Note: This figure displays the estimated coefficients for each competence fixed effect when regressing the gender, socioeconomic, and conditional ethnic performance gaps. Not other co-factor is considered. The reference category for the estimated gaps is Language, Information Technologies, with all gaps calculated relative to this baseline. Competences widening a gap have a negative sign. As a benchmark, on average, the gender performance gap is -3.1 percentage points, the SES gap is -6.6 percentage points, and the conditional ethnic gap is -1.6 percentage points. These estimates follow equation 1, and a detailed description of each competence and its topics is provided in section A.4.Competences alone explain 24% of the variance in the gender gap, 11% of the variance in the SES gap, and 8% of the variance in the conditional ethnicity gaps, as measured by R^2. Confidence intervals are estimated at the 99% level.

Figure 1: Performance Gaps by groups in each specific competence domain

5.1.3 IRT Parameters and Non-Contextual Features explaining Raw Gaps

Proceeding with the funnel approach, I now examine the extent to which IRT parameters (e.g., item difficulty and discrimination) and non-contextual features (e.g., readability, number of words) contribute to explaining the gaps, conditional on competencies fixed effects. These measures are estimated directly from the text and focus on how the item is presented and how demanding it is to process. IRT parameters capture inherent design factors like difficulty and discrimination, while non-contextual features serve as proxies for engagement (e.g., longer text or sentiment score) and effort (e.g., tables, figures, readability).

Figure A10 shows the effects over the gaps of standardized coefficients for these features. While most non-contextual features have little impact and wide confidence intervals, IRT parameters are more explanatory. The *discrimination parameter*, which measures how well an item distinguishes between high- and low-ability students, widens both SES and gender gaps. In contrast, the *difficulty parameter* reduces gaps, particularly for SES, suggesting that harder items suppress group differences. This may reflect content mastery deficiencies among certain groups, likely due to differences in the quality of secondary education. These findings suggest that test structure—captured by IRT parameters—significantly influences SES and ethnic disparities, while gender gaps are less affected.

Finally, as seen in Table A6, IRT parameters and non-contextual features explain the gaps differently across dimensions, but a substantial portion of the variance remains unexplained, leaving significant room for other factors. They account for 23.72% of the SES gap and 28.1% of the conditional ethnic gap, indicating that test structure plays a significant role in these areas. However, they provide limited insight into the gender gap (9.8%), suggesting that gender disparities may arise from different underlying factors than those driving SES and ethnic gaps.

5.2 Textual Predictors of the Gaps and Hypotheses generated

5.2.1 Top predicting Words alone

Figure A11 presents the most predictive words for each of the three gaps. For the SES gap, a suggestive pattern emerges, with words like "price," "land", "product," "urban," and "consumption" contributing to wider gaps. The effect sizes for these words range between 0.005 and 0.007 percentage points. On the other hand, items that focus on practical, everyday scenarios, physical sciences, and historical or political content (e.g., "work," "acid," "political") tend to reduce the gap. As shown in Figure A12, the model exhibits a reasonably good fit for the SES gap.

For the gender gap, items emphasizing language and reading comprehension, emotional and social contexts, as well as cultural or creative topics (e.g., "text," "family," "music") are associated with a widening of the gap. In contrast, items related to STEM subjects, quantitative reasoning, abstract analytical contexts, and physical geography (e.g., "energy," "rate," "graph") are likely to reduce the gap. The model for gender gaps also shows a decent fit, although it is slightly weaker compared to the SES model.

For the conditional ethnic gap, questions involving political, social, and cultural contexts, as well as scientific and technical topics, seem to reduce the gap (e.g., "political," "cultural," "mass"). Conversely, items focused on physical and biological sciences, urban environments, socioeconomic issues, and text interpretation (e.g., "cell," "urban," "text") tend to widen the gap. Notably, the effect sizes for words associated with this gap are an order of magnitude smaller than those observed for SES and gender gaps, suggesting a more subtle relationship between textual content and ethnic disparities.

However, these results should be interpreted with caution. The regressions capture the average effect of each word independently of the context in which it is used, and it is possible that the same word could widen or reduce the gap depending on the context. For example, a reference to a "price discount" might have a positive effect, while a mention of a "price" that feels unreachable for the test-taker could have a negative impact.

5.2.2 Top predicting topics

The next step is to explore the extent to which the occurrence and co-occurrence of words, grouped as topics or clusters, affect the observed gaps. As explained in the methodology section, the most appropriate and natural way to proceed is by using Hierarchical LDA (Latent Dirichlet Allocation) for topic modeling.

This allows the algorithm to define the number of topics in a data-driven way by maximizing the coherence score, a common metric to guide these algorithms.

Figure A14 shows the coherence score for different numbers of topics, with the maximum coherence achieved at 114 topics.The coherence score is a commonly used metric to evaluate how interpretable or meaningful the topics are. A higher coherence score indicates that the topics are more semantically coherent and easier to interpret. It helps guide the selection of the optimal number of topics for the model. It is important to note that the topic clusters are not directly related to the estimated gaps, but rather reflect the distribution of words that tend to occur together.

I then proceed by running an L1-regularization (LASSO) model that includes all item features, competence fixed effects, and the 114 discovered topics to predict the gaps. The advantage of using L1 regularization is that it shrinks irrelevant information to zero, highlighting the explanatory power of the selected factors. This approach allows me to identify the top topics that explain the variation in the gaps, both widening and reducing them.

Tables A10, A11 and A12 provide the coefficients for each of the topics and other elements for SES, Gender and Ethnicity Dimension. It can be observed that LASSO is capturing competencies fixed effects as well as several non-contextual features, aligning with the evidence presented in the previous section. Additionally, certain topics—derived solely from the distribution of text and independent of the performance gaps—emerge as highly predictive in both directions across all dimensions. However, for instance, simply knowing that Topics 24, 58, and 47 are associated with the widening of the gender gap does not provide much interpretive insight. Figure 2 shows the vectorized distribution of words for each of these topics along with their corresponding weights.

As shown, it is challenging to identify systematic patterns across these vectors to effectively characterize the content of each topic. Additionally, since each question's text is represented by a distribution of topics, linking an item to a single dominant topic risks losing valuable information. Therefore, independent assistance is required to synthesize these results and provide a meaningful explanation of how the topics relate to the observed performance gaps.

5.2.3 Hypotheses generated

The final step is to document the output generated by ChatGPT when it was prompted to provide hypotheses explaining the patterns behind the top 50 words and key topics, along with their composition, in relation to the variation in performance gaps. The rationale for using this tool is that ChatGPT can help identify patterns that may be challenging for a researcher to detect. Additionally, it serves as an independent resource for generating hypotheses, ensuring that the analysis is not influenced by the author's potential biases or preconceived notions.

The six hypotheses provided are the following:

- **SES Gap:**
  - **Widening:** Test questions focused on financial situations, business-related decision making,

Topic 24:

$$\begin{bmatrix} 0.046 \times \text{“this”} + 0.040 \times \text{“of”} + 0.028 \times \text{“wave”} + 0.018 \times \text{“month”} + 0.016 \times \text{“year”} \\ +0.016 \times \text{“are”} + 0.016 \times \text{“increase”} + 0.014 \times \text{“present”} + 0.014 \times \text{“brazil”} + 0.013 \times \text{“occur”} \\ +0.012 \times \text{“object”} + 0.012 \times \text{“represent”} + 0.011 \times \text{“departure”} + 0.011 \times \text{“large”} + 0.011 \times \text{“brazilian”} \\ +0.010 \times \text{“height”} + 0.010 \times \text{“period”} + 0.010 \times \text{“produce”} + 0.010 \times \text{“development”} + 0.010 \times \text{“national”} \\ +0.010 \times \text{“follow”} + 0.010 \times \text{“know”} + 0.010 \times \text{“region”} + 0.009 \times \text{“relationship”} + 0.008 \times \text{“country”} \\ +0.007 \times \text{“be”} + 0.007 \times \text{“length”} + 0.007 \times \text{“plane”} + 0.007 \times \text{“analyze”} + 0.007 \times \text{“power”} \end{bmatrix}$$

Topic 58:

$$\begin{bmatrix} 0.069 \times \text{“player”} + 0.064 \times \text{“transport”} + 0.057 \times \text{“game”} + 0.054 \times \text{“pass”} + 0.026 \times \text{“has”} \\ +0.025 \times \text{“result”} + 0.022 \times \text{“currently”} + 0.020 \times \text{“international”} + 0.020 \times \text{“greek”} + 0.020 \times \text{“relate”} \\ +0.018 \times \text{“are”} + 0.017 \times \text{“possibility”} + 0.016 \times \text{“technology”} + 0.015 \times \text{“activity”} + 0.015 \times \text{“good”} \\ +0.014 \times \text{“of”} + 0.014 \times \text{“sector”} + 0.013 \times \text{“form”} + 0.013 \times \text{“still”} + 0.012 \times \text{“part”} \\ +0.012 \times \text{“phenomenon”} + 0.011 \times \text{“social”} + 0.010 \times \text{“this”} + 0.010 \times \text{“born”} + 0.009 \times \text{“five”} \\ +0.009 \times \text{“measure”} + 0.009 \times \text{“contribute”} + 0.009 \times \text{“group”} + 0.008 \times \text{“build”} + 0.008 \times \text{“world”} \end{bmatrix}$$

Topic 47:

$$\begin{bmatrix} 0.031 \times \text{“mark”} + 0.027 \times \text{“each”} + 0.017 \times \text{“km”} + 0.016 \times \text{“this”} + 0.014 \times \text{“of”} \\ +0.014 \times \text{“sell”} + 0.012 \times \text{“disease”} + 0.011 \times \text{“present”} + 0.011 \times \text{“in”} + 0.011 \times \text{“must”} \\ +0.010 \times \text{“can”} + 0.010 \times \text{“test”} + 0.010 \times \text{“earth”} + 0.010 \times \text{“some”} + 0.010 \times \text{“water”} \\ +0.010 \times \text{“fuel”} + 0.009 \times \text{“fish”} + 0.009 \times \text{“total”} + 0.008 \times \text{“user”} + 0.008 \times \text{“define”} \\ +0.008 \times \text{“probability”} + 0.008 \times \text{“following”} + 0.007 \times \text{“wire”} + 0.007 \times \text{“radius”} + 0.007 \times \text{“large”} \\ +0.007 \times \text{“previous”} + 0.007 \times \text{“graph”} + 0.007 \times \text{“environment”} + 0.006 \times \text{“trash”} + 0.006 \times \text{“future”} \end{bmatrix}$$

Note: This figure illustrates the composition of the three topics most predictive of the gender gap widening, topic 24, 58 and 47, as identified by the LASSO estimations. Each topic is a weighted combination of words clustered by the LDA model, based on word occurrence and co-occurrence in the same text. Words were translated using Deepl API. It is difficult to pinpoint the specific drivers of the gender gap solely from the presence of individual words within a topic, and even more challenging when considering all topics together. This complexity underscores the necessity of tools like ChatGPT to generate hypotheses and interpret the results effectively.

Figure 2: Top Words and Weights for Topics 24, 58, and 47

and abstract economic concepts increase the SES gap.

  - **Reducing:** Test questions that utilize real-world applications, familiar settings, or culturally relevant contexts reduce the SES performance gap.

- **Gender Gap:**

  - **Widening:** Test questions involving environmental and physical science contexts, particularly those utilizing complex abstract scientific terminology, contribute to widening the gender gap in performance
  - **Reducing:** Test questions involving real-world, practical, or familiar contexts across societal, cultural, or technical domains contribute to reduce the gender performance gap.

- **Ethnic Gap:**

- **Widening:** Test questions involving abstract social phenomena, legal concepts, and technical terminologies increase the conditional ethnic gap.
- **Reducing:** Test questions focused on historical, political, and cultural contexts reduce the conditional ethnic gap.

These hypotheses are testable using panel data, which also allows for the observation of potential spillover effects. Furthermore, they are broad enough to identify flagged items across various competencies and subjects, enabling general claims that are not confined to specific domains. Importantly, these hypotheses are closely aligned with the theory of Stereotype Threat, ensuring they are well-grounded in established theoretical frameworks.

5.3 Cueing and within individual performance

In this section, I describe the results of the final step of the analysis, where the hypotheses are tested on the individual-level random sample, considering the complete set of 180 questions that participants faced. To determine whether specific items exhibit the properties described in each hypothesis, I employed a keyword-based tagging approach to classify the items accordingly. This method allows for a systematic identification of items that align with the hypothesized characteristics, ensuring a rigorous testing of the relationships between item content and the performance gaps.

5.3.1 Hypotheses testing overall population

Table 1 presents the results from estimating equation 2. Each column displays the estimates of the factors influencing the probability of correctly answering an item (coded as 0/100), including the interaction terms for each of the proposed hypotheses.

Before interpreting the results related to flagged items, it is worth noting the average gaps for non-labeled items, which serve as a baseline in this analysis. For non-labeled items, the mean SES gap is 6.1 percentage points, the gender gap is 3.6 percentage points, and the ethnic gap is 4.2 percentage points. These baseline gaps highlight the disparities observed in the data even in the absence of specific flagged characteristics.

In the case of the SES gap, the presence of a question in the widening channel, that is flagged as containing financial terms, widens the gap by 1.4 percentage points. This represents a 22% increase relative to the average SES gap of 6.1 percentage points, indicating strong support for the hypothesis that abstract financial and business-related terms disadvantage lower-SES test-takers. On the other hand, items that involve real-world, practical contexts, as outlined for the reducing channel in SES, help reduce the SES gap by 0.8 percentage points. This result suggests that test items grounded in everyday scenarios can mitigate disparities.

For the gender gap, in turn, the results for the widening effect across the entire population are not statistically supported by the testing procedure. Therefore, abstract scientific and technical terms do not appear to trigger larger gender gaps when considering all test-takers, conditional on . However, items

Table 1: Hypotheses and their impact on Performance (p.p)

	SES Gap	Gender Gap	Ethnic Gap
	Correct	Correct	Correct
	(0/100)	(0/100)	(0/100)
Widening	1.092***	-0.078*	-0.181***
	(0.045)	(0.043)	(0.042)
Widening × Dimension	-1.424***	0.226***	-0.154***
	(0.081)	(0.054)	(0.055)
Reducing	-0.320***	-0.967***	-0.097
	(0.031)	(0.044)	(0.114)
Reducing × Dimension	0.810***	0.644***	-0.058
	(0.055)	(0.056)	(0.251)
Mean Gap	-6.141	-3.583	-4.192
Item's Controls	Yes	Yes	Yes
Indiv X Subject FE	Yes	Yes	Yes
Other FEs	Yes	Yes	Yes
Obs.	22,006,964	22,006,964	22,006,964
Indiv.	129,982	129,982	129,982

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

Note: This table presents the results for the relevant coefficients when regressing equation 2. Each dimension—socioeconomic status (SES), gender, and ethnicity—is tested independently using the same set of observations. The dependent variable is whether the test-taker answered the specific question correctly. The variables Widening and Reducing are dummies set to 1 if the question corresponds to the widening or reducing channel for a given dimension. The variable Dimension is a dummy set to 1 if the test-taker belongs to a disadvantaged group—low-SES, female, or non-white—depending on the specification. Mean gaps are provided as benchmarks, estimated using the same specification but without Individual × Subject fixed effects. Standard errors are clustered at the individual level and are reported in parentheses.

associated with reducing the gender gap, such as those centered around practical or social contexts, show a reduction of 0.6 percentage points.

Regarding the ethnic gap, the widening channel shows a statistically significant effect at the 99% confidence level, but the size of the impact is only 0.15 percentage points, which is relatively small and may not be substantively meaningful. The reduction channel, however, does not receive support from the whole sample, as no significant effects are observed for items aimed at reducing the ethnic gap.

5.3.2 Hypotheses by ability quartile

As the consequences of these disparities vary across ability quartiles—particularly because test-takers must exceed the selective cut-off of their desired program through the SISU clearinghouse system—the next analysis proposes splitting the sample into four quartiles based on ability. Historical data on previous cut-offs indicate that even the least demanding programs (those around the 10th percentile of cut-offs) align with scores at the upper limit of the fourth quartile. This reflects the high level of competition, where test-takers in the top quartile must meet or exceed these cut-offs. Within each quartile, I compare individuals with similar abilities to estimate the effects of the proposed hypotheses.

Ability is estimated using the IRT (Item Response Theory) grading model, the θ parameter, which represents the latent trait explaining the pattern of responses across item characteristic curves. This is a value that is provided with the dataset as it is easily identified from the score in each subject. The intuition behind this model is that if a test-taker answers only the easier items correctly, but also answers some very difficult items correctly without performing consistently in the intermediate range, there is a high likelihood that these correct responses were due to guessing. Based on the calibrated parameters for each item—calibrations conducted by the test designers in a pilot study outside the actual ENEM application—an ability estimate can be performed for each test-taker in a given subject by observing their responses to the 45 items they faced. An important feature of this ability estimate is that it is comparable across years, even though test-takers face different sets of items with varying levels of difficulty. This ability measure is independent of group composition or characteristics.

Table 2 presents the impact of interaction terms on performance across ability quartiles, from Q1 (Low Ability) to Q4 (High Ability). Each column corresponds to a different quartile, and the estimates reflect the factors influencing the probability of correctly answering an item (in percentage points) for various subgroups of test-takers.

Performance gaps are most pronounced in the fourth quartile, where interaction effects become significantly relevant. In the lower ability quartiles (Q1–Q3), the gaps are relatively small, suggesting a possible "floor effect." This effect likely occurs because the items are too difficult for most test-takers in these quartiles, leaving little room for performance disparities to emerge. In contrast, within the highest ability quartile (Q4), gaps become more evident as test-takers' abilities match the item difficulty, allowing disparities to surface more clearly.

The widening effect of flagged items is also more pronounced in Q4. For example, the SES gap widens substantially among high-ability low-income students by 0.593 percentage points, while high-ability female

students experience a widening gender gap of 1.064 percentage points. This indicates that while higher quartiles expose more significant gaps, the flagged items exacerbate these disparities, particularly for top performers, suggesting that test content can disproportionately impact high-ability students from disadvantaged groups.

By contrast, the reduction channels show no strong ability-related effects. The interaction terms across ability quartiles for reducing the gaps—whether SES, gender, or ethnicity—remain relatively consistent, indicating that the mitigating effects of items focused on real-world, practical contexts are not significantly influenced by test-takers' abilities. The only dimension that shows consistent reduction effects across abilities is gender, with the reduction channel remaining fairly constant across ability quartiles.

Table 2: Impact of Interaction Terms on Performance (p.p) by Ability Quartile

	Splitting by Ability Quartiles (p.p.)			
	Q1 (Low)	Q2	Q3	Q4 (High)
	Correct (0/100)	Correct (0/100)	Correct (0/100)	Correct (0/100)
Widening × Low Income	0.092	-0.162	-0.436**	-0.593**
	(0.133)	(0.149)	(0.175)	(0.251)
Reducing × Low Income	0.187**	-0.011	0.353***	-0.016
	(0.093)	(0.105)	(0.121)	(0.168)
Widening × Girl	-0.027	-0.341***	-0.479***	-1.064***
	(0.102)	(0.106)	(0.108)	(0.115)
Reducing × Girl	0.486***	0.785***	0.668***	0.897***
	(0.104)	(0.114)	(0.115)	(0.111)
Widening × Non-White	0.315***	0.319***	0.117	-0.200*
	(0.104)	(0.108)	(0.111)	(0.114)
Reducing × Non-White	0.065	-0.579**	0.144	0.082
	(0.219)	(0.229)	(0.234)	(0.240)
SES Gap (p.p.)	-0.177	-0.154	-0.204	-3.670
Gender Gap (p.p.)	-0.153	-0.156	-0.198	-1.731
Ethnic Gap (p.p.)	-0.148	-0.075	-0.058	-2.482
Item's Controls	Yes	Yes	Yes	Yes
Indiv X Area FE	Yes	Yes	Yes	Yes
Other FEs	Yes	Yes	Yes	Yes
Obs.	5,508,994	5,501,288	5,497,670	5,499,012

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

Note: This table presents the results from estimating equation 2, including all possible interactions simultaneously, with the sample split into four quartiles based on test-takers' ability. Each column compares test-takers within the same quartile, with the fourth quartile representing those most likely to exceed the required cutoffs for admission to a public university. Intraquartile gaps are provided as benchmarks. The focus of the analysis is on the interacted coefficients, which are the only ones reported. Standard errors are clustered at the individual level and shown in parentheses.

5.3.3 Other Heterogeneity analyses

Splitting by Subjects: The first natural analysis is to examine how these effects vary across subjects, as stereotype levels are known to differ by subject area (Bordalo et al., 2019; Guiso et al., 2008). Table A13 presents the results. For the full sample, the widening effect on the SES gap is present across all subjects, with a stable increase of around 1 percentage point, slightly more severe in mathematics and science. Interestingly, the negative effects found in Language suggest that the widening is driven by cognitive effects evoked by the context, rather than a lack of specific competencies related to operations or unit handling. In contrast, the widening effect for gender is only observed in Natural Sciences and Language, while in mathematics, it unexpectedly moves in the opposite direction. For ethnicity, the widening channel emerges in both Language and Mathematics. The reduction channels, on the other hand, are evident in Natural Sciences and Mathematics for SES, across all subjects for gender (most pronounced in Mathematics), and in Language and Mathematics for ethnicity.

When focusing on high performers in the top quartile, the widening effect for SES persists only in Mathematics and Social Sciences, disappearing in Language and Natural Sciences, with no evidence of a reduction effect. For gender, however, the widening effect appears across all subjects, reinforcing the idea that stereotype threats are more likely to affect individuals who are highly identified with the domain. The reduction channel is present in all subjects except Mathematics, with an effect centered around 1 percentage point. Finally, for ethnicity, both widening and reducing effects only emerge in Language.

Splitting by ability-adjusted difficulty: Another important source of heterogeneity is the relationship between these effects and the difficulty of the question. Difficulty has been identified as a significant moderator of stereotype threat effects (Spencer et al., 2016; Schmader et al., 2015), which may also be explained by floor effects, where difficult questions leave no room for disparities to emerge since most test-takers answer them incorrectly. I address this by leveraging the difference between each test-taker's latent ability and the difficulty parameter in the IRT model, providing a straightforward measure of difficulty. The higher the adjusted difficulty, the farther away the difficulty parameter is with respect to the θ estimation for individual ability. I focus my attention only on high-ability test-takers, as for them, the difficulty of questions seems to be better calibrated (see Figure A15). The results, shown in Table A14, indicate that widening effects diminish towards zero for SES and gender, while they only emerge for ethnicity as the question becomes harder, suggesting that these channels only emerge when the ability-difficulty matching is accounted for. In turn, reducing channels are only present for the Gender dimension, shrinking towards zero as the question becomes harder.

Role of Stereotyped Items: Another interesting source of variation is whether, within competencies, the content of the sub-competence is stereotyped and how this interacts with the effects described so far. Following the approach of Coffman (2014) and Bordalo et al. (2019), I categorize questions based on the specific content of the sub-competence to which they belong. This analysis is performed exclusively for gender effects, as stereotypes are already present across competencies related to SES and ethnic gaps, and the gender dimension has proven to be the most sensitive to these effects in this paper.

I define a stereotyped item as one belonging to a sub-competence (one of the 30 that describe the possible objectives of each question) where its contribution to widening the gender performance gap exceeds the observed average. The results show that 47% (1,035) of the items are flagged as stereotyped based on their sub-competence descriptions.

Table A15 presents the results. First, stereotyped items interacted with the female indicator show a negative and statistically significant effect across all ability quartiles, demonstrating that stereotyped domains significantly harm female performance, regardless of ability. Second, both the widening and reduction effects remain robust to the inclusion of this dummy. Widening effects emerge even in items outside stereotyped competencies, and the inclusion of this dummy now brings the coefficient for the whole sample in line with the hypothesis, which was not the case before. Third, the triple interaction term suggests that widening effects persist in stereotyped domains, but only for test-takers in the top quartile, indicating that high-ability individuals might be more susceptible to stereotype threats as the coefficient shifts from positive to zero. Lastly, the reduction channels are robust and consistent across ability levels for non-stereotyped content but disappear entirely when the competence is stereotyped.

Role of Mirroring Effect: Building on the insights from the first section, I test whether the presence of a character mirroring the test-taker's group introduces heterogeneity in the results. To do this, I manually tag items based on whether they depict a female character (inferred through the name or gendered pronouns) and whether they depict a character associated with an underprivileged situation (determined primarily by context and name). I find that 13.7% (301 items) depict female characters, and 5.4% (118 items) depict individuals from low-SES backgrounds.

Table A16 summarizes the results for the gender gap, while Table A17 does the same for the SES gap. For the gender gap, the presence of female characters in an item completely offsets the widening effect across all quartiles, though it does not influence the reduction effect in any of them. In contrast, for the SES gap, this mirroring effect does not act as a protective factor. In fact, the widening effect increases when a character from a low-SES background is present, suggesting that the attention channel is at play. Additionally, in the reduction channel, the presence of these characters reverses the effect, now causing a widening of the gap.

These findings suggest that the presence of female characters may help mitigate the negative impact of items that hold for the widening channel, thereby reducing gender gaps. However, this mirroring effect does not appear to be effective in addressing the widening effects associated with SES gaps.

5.3.4 Are there negative spillovers?

A final step is to examine whether there are spillover effects from one tagged item to those that follow. This is important to consider because if spillovers occur, the overall impact on the test might be greater than anticipated. For instance, Duquennois (2022) documents in the low-stakes context of the TIMSS that an item framed financially can impair the performance of test-takers from the bottom half of the income distribution for up to four subsequent questions. Therefore, I will adopt an approach similar to that suggested by Duquennois (2022). The idea is to isolate the items within windows where they appear

only once. To achieve this, I conduct a panel event study, defining a window of minus four to plus four positions relative to a flagged item, ensuring that no other flagged item appears within this window.[30]

$$Correct_{iq} = \alpha + \sum_{p=-4, p\neq 1}^{4} \pi_p (P_q = p) \times Dimension_i + \vartheta_q + \chi_i + \eta_p + \epsilon_{iq} \tag{3}$$

Where, apart from the previously described items, P_q is the item position relative to the flagged item, interacted with the group of interest, χ_i is an individual fixed effect, and ϑ_q is the item-specific fixed effect. Given the evidence so far, I focus my analyses in the top half performers.

Figure A16 presents the results of the event study. It can be observed that only the widening channel has a negative and statistically significant effect on the flagged items, with around a 2 percentage point decline in both cases. The effect is local and only affects the treated item. This aligns partially with the evidence from Duquennois (2022), suggesting that while the underlying cause may be similar, the spillover effects differ significantly in a high-stakes context with highly motivated students. Conversely, the reduction channels show null effects for the SES gap and produce noisy, erratic estimations with pre-trends, so not much can be concluded from these results.

6 Discussion and Conclusion

This paper shows that the contextual features of test questions significantly influence performance disparities in high-stakes standardized tests, based on 13 years of data from 3.8 million test-takers. Through a detailed analysis of question-by-question performance gaps across gender, socioeconomic status (SES), and ethnicity, the study tests six hypotheses on how these gaps are affected by different types of content. The findings support four out of six hypotheses, showing that SES gaps widen by 1.4 percentage points with items involving abstract and financially related content, while items related to grounded, daily activities reduce the gap by 0.8 percentage points, particularly among lower-ability test-takers. For gender gaps, abstract scientific concepts and measurements increase the gap by 1.06 percentage points, especially among high-ability female test-takers, while practical problem-solving scenarios, creativity, emotions, and social interactions consistently reduce the gap by 0.8 percentage points across all ability levels.

The policy implications of these findings are profound. Simple adjustments in test design—such as reducing stereotyped content and incorporating more neutral textual contexts—can significantly narrow performance gaps. This approach is particularly relevant as centralized admission systems, which rely heavily on standardized tests, are increasingly adopted for their perceived fairness and cost-effectiveness.

This is highly consequential in the context of Brazil. For instance, Duryea et al. (2023) shows that the marginal low-SES student who benefits from attending a high-quality, free public university increases their earnings by 26% ten years later, with no significant impact on high-SES students. The effect is driven by low-SES students who, in the absence of access to these universities, enroll in colleges with lower returns, and have little opportunity to close the earnings gap later in their careers.

[30] The choice of this position window is based on the spillover effects documented in Duquennois (2022). The goal is to have a window long enough to capture spillover effects, but not so large that it results in identifying only a few items.

Furthermore, the evidence suggests that targeted interventions, such as pairing gender-related widening questions including female characters, could effectively mitigate disparities. Policymakers and educational institutions can use these insights to refine test design, making access to higher education more equitable, especially in high-stakes environments like Brazil's ENEM.

This research underscores the importance of expanding access to test questionnaires, as Brazil has done, to allow for continuous improvements in test fairness. By leveraging NLP and data-driven methods, policymakers can better understand and address the sources of performance disparities, ultimately creating more inclusive educational opportunities.

References

Afridi, F., Li, S. X., and Ren, Y. (2015). Social identity and inequality: The impact of china's hukou system. *Journal of public economics*, 123:17–29.

Altmejd, A., Bizopoulou, A., Kaila, M., Barrios-Fernández, A., Neilson, C., Otero, S., and Ye, X. (2022). Inequality in college applications: Evidence from three continents.

Anaya, L., Iriberri, N., Rey-Biel, P., and Zamarro, G. (2022). Understanding performance in test taking: The role of question difficulty order. *Economics of Education Review*, 90:102293.

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American statistical Association*, 103(484):1481–1495.

Attali, Y., Neeman, Z., and Schlosser, A. (2018). Differential performance in high vs. low stakes tests: evidence from the gre test.

Baldiga, K. (2014). Gender differences in willingness to guess. *Management Science*, 60(2):434–448.

Batista, R. M. and Ross, J. (2024). Words that work: Using language to generate hypotheses. Working paper. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4926398.

Beilock, S. L. and McConnell, A. R. (2004). Stereotype threat and sport: Can athletic performance be threatened? *Journal of Sport and Exercise Psychology*, 26(4):597–609.

Binelli, C. and Menezes-Filho, N. (2019). Why brazil fell behind in college education? *Economics of Education Review*, 72:80–106.

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2016). Stereotypes. *The Quarterly Journal of Economics*, 131(4):1753–1794.

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2019). Beliefs about gender. *American Economic Review*, 109(3):739–773.

Brown, C. L., Kaur, S., Kingdon, G., and Schofield, H. (2022). Cognitive endurance as human capital. Technical report, National Bureau of Economic Research.

Buser, T., Niederle, M., and Oosterbeek, H. (2014). Gender, competitiveness, and career choices. *The quarterly journal of economics*, 129(3):1409–1447.

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

Carvalho de Lima Moreno, G., de Souza, M. P., Hein, N., and Kroenke Hein, A. (2022). Alt: A software for readability analysis of portuguese-language texts. *arXiv e-prints*, pages arXiv–2210.

Charness, G., Jabarian, B., and List, J. A. (2023). Generation next: Experimentation with ai. Technical report, National Bureau of Economic Research.

Coffman, K. B. (2014). Evidence on self-stereotyping and the contribution of ideas. *The Quarterly Journal of Economics*, 129(4):1625–1660.

Coffman, K. B., Exley, C. L., and Niederle, M. (2021). The role of beliefs in driving gender discrimination. *Management Science*, 67(6):3551–3569.

Coffman, K. B. and Klinowski, D. (2020). The impact of penalties for wrong answers on the gender gap in test scores. *Proceedings of the National Academy of Sciences*, 117(16):8794–8803.

Cohen, A., Karelitz, T., Kricheli-Katz, T., Pumpian, S., and Regev, T. (2023). Gender-neutral language and gender disparities. Technical report, National Bureau of Economic Research.

Cullen, M. J., Waters, S. D., and Sackett, P. R. (2006). Testing stereotype threat theory predictions for math-identified and non-math-identified students by gender. *Human Performance*, 19(4):421–440.

De Paola, M. and Gioia, F. (2016). Who performs better under time pressure? results from a field experiment. *Journal of Economic Psychology*, 53:37–53.

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

Duryea, S., Ribas, R. P., Sampaio, B., Sampaio, G. R., and Trevisan, G. (2023). Who benefits from tuition-free, top-quality universities? evidence from brazil. *Economics of Education Review*, 95:102423.

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

Ellis, A. P. J. and Ryan, A. M. (2003). Race and cognitive-ability test performance: The mediating effects of test preparation, test-taking strategy use and self-efficacy. *Journal of Applied Social Psychology*, 33(12):2607–2629.

Flore, P. C., Mulder, J., and Wicherts, J. M. (2018). The influence of gender stereotype threat on mathematics test scores of dutch high school students: A registered report. *Comprehensive Results in Social Psychology*, 3(2):140–174.

Franco, C. and Povea, E. (2024). Innocuous exam features? the impact of answer placement on high-stakes test performance and college admissions. *NHH Dept. of Economics Discussion Paper*, (04).

Freedle, R. (2003). Correcting the sat's ethnic and social-class bias: A method for reestimating sat scores. *Harvard Educational Review*, 73(1):1–43.

Fryer, R. G., Levitt, S. D., and List, J. A. (2008). Exploring the impact of financial incentives on stereotype threat: Evidence from a pilot study. *American Economic Review*, 98(2):370–375.

Galasso, V. and Profeta, P. (2024). Gender differences in math tests: The role of time pressure. *The Economic Journal*, page ueae052.

Gomez-Ruiz, M., Cervini-Plá, M., and Ramos, X. (2024). Do women fare worse when men are around? quasi-experimental evidence.

Goodman, J., Gurantz, O., and Smith, J. (2020). Take two! sat retaking and college enrollment gaps. *American Economic Journal: Economic Policy*, 12(2):115–158.

Griselda, S. (2022). The gender gap in math: What are we measuring? *Available at SSRN 4022082*.

Guiso, L., Monte, F., Sapienza, P., and Zingales, L. (2008). Culture, gender, and math. *Science*, 320(5880):1164–1165.

Hickendorff, M. (2013). The effects of presenting multidigit mathematics problems in a realistic context on sixth graders' problem solving. *Cognition and Instruction*, 31(3):314–344.

Hoff, K. and Pandey, P. (2014). Making up people—the effect of identity on performance in a modernizing society. *Journal of Development Economics*, 106:118–131.

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., and Williams, C. C. (2008). Gender similarities characterize math performance. *Science*, 321(5888):494–495.

Kiessling, L., Pinger, P., Seegers, P., and Bergerhoff, J. (2024). Gender differences in wage expectations and negotiation. *Labour Economics*, 87:102505.

Koedinger, K. R., Alibali, M. W., and Nathan, M. J. (2008). Trade-offs between grounded and abstract representations: Evidence from algebra problem solving. *Cognitive Science*, 32(2):366–397.

Koedinger, K. R. and Nathan, M. J. (2004). The real story behind story problems: Effects of representations on quantitative reasoning. *The journal of the learning sciences*, 13(2):129–164.

Landaud, F., Maurin, É., Willage, B., and Willén, A. (2024). The value of a high school gpa. *Review of Economics and Statistics*, pages 1–24.

Lodder, P., Ong, H. H., Grasman, R. P., and Wicherts, J. M. (2019). A comprehensive meta-analysis of money priming. *Journal of Experimental Psychology: General*, 148(4):688.

Ludwig, J. and Mullainathan, S. (2024). Machine Learning as a Tool for Hypothesis Generation*. *The Quarterly Journal of Economics*, page qjad055.

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

Messeri, L. and Crockett, M. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002):49–58.

Miller, D. I. and Halpern, D. F. (2014). The new science of cognitive sex differences. *Trends in cognitive sciences*, 18(1):37–45.

Muskens, M., Frankenhuis, W. E., and Borghans, L. (2024). Math items about real-world content lower test-scores of students from families with low socioeconomic status. *npj Science of Learning*, 9(1):19.

Ofek-Shanny, Y. (2024). Measurements of performance gaps are sensitive to the level of test stakes: Evidence from pisa and a field experiment. *Economics of Education Review*, 98:102490.

Otero, S., Barahona, N., and Dobbin, C. (2021). Affirmative action in centralized college admission systems: Evidence from brazil. *Unpublished manuscript*.

Pinotti, P., Britto, D. G., Fonseca, A., Sampaio, B., and Warwar, L. (2022). Intergenerational mobility in the land of inequality. Technical report, CReAM Discussion Paper Series.

Priest, R., Griebie, A., Zhou, Y., Tomeh, D., and Sackett, P. R. (2024). Stereotype lift and stereotype threat effects on subgroup mean differences for cognitive tests: A meta-analysis of adult samples. *Journal of Applied Psychology*.

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

Reyes, G., Riehl, E., and Xu, R. (2023). Do high stakes muddle the information from standardized tests? evidence from brazil's enem exam.

Reynolds, M. R., Hajovsky, D. B., and Caemmerer, J. M. (2022). The sexes do not differ in general intelligence, but they do in some specifics. *Intelligence*, 92:101651.

Roussille, N. (2024). The role of the ask gap in gender pay inequality. *The Quarterly Journal of Economics*, page qjae004.

Rudner, L. M. (1992). Pre-employment testing and employee productivity. *Public Personnel Management*, 21(2):133–150.

Schmader, T., Hall, W., and Croft, A. (2015). Stereotype threat in intergroup relations. pages 447–471.

Schmidt, F. L. and Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological bulletin*, 124(2):262.

Shewach, O. R., Sackett, P. R., and Quint, S. (2019). Stereotype threat effects in settings with features likely versus unlikely in operational test settings: A meta-analysis. *Journal of Applied Psychology*, 104(12):1514.

Smith, J. L. and White, P. H. (2001). Development of the domain identification measure: A tool for investigating stereotype threat effects. *Educational and Psychological Measurement*, 61(6):1040–1057.

Spencer, S. J., Logel, C., and Davies, P. G. (2016). Stereotype threat. *Annual review of psychology*, 67:415–437.

Steele, C. M. and Aronson, J. (1995). Stereotype threat and the intellectual test performance of african americans. *Journal of personality and social psychology*, 69(5):797.

Stenlund, T., Eklöf, H., and Lyrén, P.-E. (2017). Group differences in test-taking behaviour: An example from a high-stakes testing program. *Assessment in Education: Principles, Policy & Practice*, 24(1):4–20.

Van de Weijer-Bergsma, E. and Van der Ven, S. H. (2021). Why and for whom does personalizing math problems enhance performance? testing the mediation of enjoyment and cognitive load at different ability levels. *Learning and Individual Differences*, 87:101982.

Vaz, D. V. (2020). Background familiar, retornos da educaÇÃo e desigualdade racial no brasil. *Cadernos de Pesquisa*, 50(177):845–864.

Vohs, K. D., Mead, N. L., and Goode, M. R. (2006). The psychological consequences of money. *science*, 314(5802):1154–1156.

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., et al. (2023). Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60.

World Bank (2024). Fy24-28 country partnership framework for the federative republic of brazil. Technical report, World Bank Group.

Zanella, M. (2021). Stereotypical selection. *Job-Market Paper*.

Zhang, H., Zhou, X., Nielsen, M. S., and Klyver, K. (2023). The role of stereotype threat, anxiety, and emotional intelligence in women's opportunity evaluation. *Entrepreneurship Theory and Practice*, 47(5):1699–1730.

A.1 Appendix Tables

Appendix Table A1: Comparison of Score Gaps across Admission Tests in Standardized Score

Admission Test	Country	SES Gap	Gender Gap	Ethnic Gap	Reference Year	Source
Gokao	China	-0.75	0.21	N.A	2018	(Altmejd et al., 2022)
PAES	Chile	-0.36	-0.15	N.A	2024	Author's estimations
ENEM	Brazil	-0.69	-0.31	-0.48	2022	Author's estimations
Saber 11	Colombia	-0.61	-0.17	-0.63	2019	Author's estimations
Ylioppilastutkinto	Finland	-0.67	0.09	N.A	2020	(Altmejd et al., 2022)
Panhellenic Examinations	Greece	-0.28	0.19	N.A	2012	(Altmejd et al., 2022)
EBAU	Spain	-0.66	0.17	N.A	2020	(Altmejd et al., 2022)
Högskoleprovet	Sweden	-0.59	-0.38	N.A	2016	(Altmejd et al., 2022)

Note: This table presents evidence on performance gaps among relevant groups across different admission tests in countries that use centralized clearinghouses for college placement. The aim is to provide benchmarks. All gaps are expressed in terms of standard deviations on the official grades used to apply through Centralized Admission Systems. SES gaps represent the differences between low- and high-income populations, gender gaps reflect differences between females and males, and ethnic gaps show differences between non-white and white test-takers.

Appendix Table A2: Distribution of Flagged Items by Subject Category

	SES_Widen	SES_Reduce	Gender_Widen	Gender_Reduce	Ethnic_Widen	Ethnic_Reduce
Social Science	21.9%	13.1%	8.8%	38.7%	42.5%	40.9%
Natural Science	6.5%	36.6%	42.5%	6.6%	10.8%	0.0%
Language	13.6%	18.3%	7.6%	53.2%	29.2%	56.1%
Math	58.0%	32.0%	41.1%	1.5%	17.5%	3.0%
Total Items	169	366	341	395	325	66
Respect All	7.7%	16.6%	15.5%	17.9%	14.8%	3.0%

Note: This table presents the distribution of tagged items by hypothesis content. It shows that 169 items are tagged as inducing a widening of SES gaps, while 366 are tagged as reducing SES gaps. In turn, 341 items are associated with widening gender gaps and 395 with reducing them. Finally, 325 items are linked to widening ethnic gaps, while 66 are associated with reducing them. All channels are represented across all subjects.

Appendix Table A3: Summary Statistics by Subject Area

	Social Sciences					**Natural Science**				
	Count	Mean	SD	Min	Max	Count	Mean	SD	Min	Max
Readability Score	576	-0.26	1.00	-3.88	3.30	568	-0.39	0.84	-3.17	2.47
Length in Words (SD)	576	-0.37	0.68	-1.82	1.90	568	-0.26	0.71	-1.82	3.01
Figures	576	0.16	0.37	0.00	1.00	568	0.29	0.46	0.00	1.00
Tables	576	0.00	0.06	0.00	1.00	568	0.05	0.22	0.00	1.00
Sentiment Score	576	0.16	0.12	0.00	0.83	568	0.13	0.11	0.00	0.70
Grounded	576	0.01	0.08	0.00	1.00	568	0.03	0.17	0.00	1.00
Man Character	576	0.14	0.35	0.00	1.00	568	0.10	0.30	0.00	1.00
Woman Character	576	0.18	0.38	0.00	1.00	568	0.06	0.24	0.00	1.00
Underprivileged	576	0.08	0.27	0.00	1.00	568	0.02	0.13	0.00	1.00
Privileged	576	0.24	0.43	0.00	1.00	568	0.08	0.27	0.00	1.00
Discrimination P.	576	2.16	0.97	0.34	6.69	568	2.28	1.03	0.30	8.00
Difficulty P.	576	1.14	0.78	-2.11	4.49	568	1.39	0.79	-1.80	4.44
Guessing P.	576	0.17	0.07	0.00	0.50	568	0.16	0.07	0.00	0.48
% of Corrects	576	0.50	0.18	0.06	0.95	568	0.38	0.17	0.08	0.95
SES Widen	576	0.06	0.25	0.00	1.00	568	0.02	0.14	0.00	1.00
SES Reduce	576	0.08	0.28	0.00	1.00	568	0.24	0.42	0.00	1.00
Gender Widen	576	0.05	0.22	0.00	1.00	568	0.26	0.44	0.00	1.00
Gender Reduce	576	0.27	0.44	0.00	1.00	568	0.05	0.21	0.00	1.00
Ethnic Widen	576	0.24	0.43	0.00	1.00	568	0.06	0.24	0.00	1.00
Ethnic Reduce	576	0.05	0.21	0.00	1.00	568	0.00	0.00	0.00	0.00
	Language					**Mathematics**				
	Count	Mean	SD	Min	Max	Count	Mean	SD	Min	Max
Readability Score	492	0.47	1.08	-2.71	4.22	565	0.39	0.72	-1.94	2.23
Length in Words (SD)	492	0.60	1.32	-1.80	5.43	565	-0.03	0.74	-1.65	4.22
Figures	492	0.22	0.41	0.00	1.00	565	0.40	0.49	0.00	1.00
Tables	492	0.00	0.00	0.00	0.00	565	0.10	0.30	0.00	1.00
Sentiment Score	492	0.18	0.14	0.00	0.78	565	0.10	0.08	0.00	0.79
Grounded	492	0.15	0.36	0.00	1.00	565	0.21	0.41	0.00	1.00
Man Character	492	0.22	0.41	0.00	1.00	565	0.14	0.35	0.00	1.00
Woman Character	492	0.24	0.43	0.00	1.00	565	0.08	0.28	0.00	1.00
Underprivileged	492	0.08	0.27	0.00	1.00	565	0.04	0.19	0.00	1.00
Privileged	492	0.21	0.41	0.00	1.00	565	0.13	0.33	0.00	1.00
Discrimination P.	492	2.06	0.82	0.27	6.00	565	2.05	0.78	0.24	6.20
Difficulty P.	492	0.86	0.84	-0.92	4.35	565	1.91	0.96	-1.26	4.97
Guessing P.	492	0.15	0.07	0.00	0.40	565	0.16	0.06	0.00	0.42
% of Corrects	492	0.37	0.24	0.06	0.96	565	0.39	0.17	0.08	0.92
SES Widen	492	0.05	0.21	0.00	1.00	565	0.17	0.38	0.00	1.00
SES Reduce	492	0.14	0.34	0.00	1.00	565	0.21	0.41	0.00	1.00
Gender Widen	492	0.05	0.22	0.00	1.00	565	0.25	0.43	0.00	1.00
Gender Reduce	492	0.43	0.50	0.00	1.00	565	0.01	0.10	0.00	1.00
Ethnic Widen	492	0.19	0.40	0.00	1.00	565	0.10	0.30	0.00	1.00
Ethnic Reduce	492	0.08	0.26	0.00	1.00	565	0.00	0.06	0.00	1.00

Note: This table presents the summary statistics at the question level, separated by subject. The final sample comprises 2,201 questions that appeared between 2010 and 2022. Questions that do not meet the inclusion criteria (i.e., those not considered for grading purposes because their performance did not align with the characteristic curve) were excluded. The Language subject contains fewer questions because the five questions per year assessing foreign language competencies are not included..

Appendix Table A4: Flagged Items and items characteristicas as predictors

	(1) SES Wd (0/1)	(2) SES Rd (0/1)	(3) Gender Wd (0/1)	(4) Gender Rd (0/1)	(5) Ethnic Wd (0/1)	(6) Ethnic Rd (0/1)
Figure (0/1)	-0.027**	0.048**	0.046**	-0.043***	-0.039**	-0.017***
	(0.013)	(0.021)	(0.019)	(0.015)	(0.016)	(0.006)
Table (0/1)	0.033	0.072	0.049	-0.034**	-0.001	0.002
	(0.042)	(0.044)	(0.041)	(0.015)	(0.040)	(0.010)
# Words	0.037***	0.076***	0.076***	0.095***	0.074***	0.017***
	(0.007)	(0.010)	(0.009)	(0.009)	(0.009)	(0.005)
Readability	-0.003	-0.043***	-0.026***	-0.016*	-0.035***	-0.009**
	(0.006)	(0.008)	(0.007)	(0.009)	(0.007)	(0.004)
Positive Sentiment	-0.049	0.131**	0.142**	0.039	0.092	0.024
	(0.037)	(0.064)	(0.057)	(0.068)	(0.062)	(0.035)
Grounded (0/1)	0.046*	0.018	0.022	-0.010	-0.036	-0.006
	(0.026)	(0.029)	(0.026)	(0.023)	(0.023)	(0.010)
Discrimination p.A	-0.005	0.010	0.018**	0.002	0.012	-0.001
	(0.005)	(0.009)	(0.008)	(0.008)	(0.008)	(0.003)
Difficulty p.B	-0.021***	0.008	0.010	-0.021**	-0.009	-0.004
	(0.007)	(0.009)	(0.008)	(0.008)	(0.008)	(0.004)
Guessing p.C	0.101	0.044	0.117	-0.175*	-0.125	-0.004
	(0.072)	(0.110)	(0.096)	(0.105)	(0.101)	(0.051)
Const.	0.046**	-0.010	-0.072***	0.431***	0.108***	0.122***
	(0.020)	(0.033)	(0.027)	(0.053)	(0.036)	(0.035)
Comptence FE	Yes	Yes	Yes	Yes	Yes	Yes
Obs.	2,201	2,201	2,201	2,201	2,201	2,201

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

Note: This table shows the results of a regression between a question being tagged with a specific hypothesis and its non-contextual features, considering competence fixed effects. These fixed effects are aimed to capture differences by domains (i.e.: some type of questions are more likely to include graphs while having shorter text). The sample comprises all included questions (2,201). The aim is to show how these features correlate with the tagged questions. It can be observed that tagged questions tend to have longer texts and slightly lower readability. They also show differences in the presence of figures. Finally, there are no significant differences in the difficulty parameter (p.B) or the discrimination parameter (p.A). Overall, this evidence suggests that the flagged items are likely similar to others in the pool. Robust standard errors are reported in parentheses.

Appendix Table A5: Summary Statistics

	Count	Mean	S.D.	Min	Max
Low Income	129,982	0.28	0.45	0.00	1.00
Girl	129,982	0.57	0.49	0.00	1.00
Bianco	129,982	0.43	0.50	0.00	1.00
Preta	129,982	0.11	0.31	0.00	1.00
Parda	129,982	0.43	0.49	0.00	1.00
Asiatic	129,982	0.02	0.14	0.00	1.00
Indigenous	129,982	0.01	0.08	0.00	1.00
Rural	129,982	0.13	0.33	0.00	1.00
Federal School	129,982	0.04	0.20	0.00	1.00
Private School	129,982	0.23	0.42	0.00	1.00
Public School	129,982	0.73	0.44	0.00	1.00
Father with Higher Education	129,982	0.16	0.37	0.00	1.00
Father with Low Education	129,982	0.24	0.43	0.00	1.00
With access to Internet	129,982	0.75	0.43	0.00	1.00
Spanish Foreign Language	129,982	0.50	0.50	0.00	1.00
Grade in Sciences (IRT)	129,982	486.02	77.51	0.00	833.40
Grade in Social Sciences (IRT)	129,982	528.57	84.78	0.00	873.20
Grade in Language (IRT)	129,982	513.44	72.39	0.00	772.90
Grade in Math (IRT)	129,982	515.97	112.56	0.00	990.70
Total Correct Science	129,982	12.91	5.13	0.00	44.00
Total Correct Social Sciences	129,982	17.47	6.79	0.00	45.00
Total Correct Language	129,982	13.92	7.67	0.00	44.00
Total Correct Math	129,982	12.89	6.00	0.00	45.00

Note: This table presents the summary statistics for the sample in the period 2010–2022. Each year, 10,000 test-takers meeting the inclusion criteria are drawn to study how question-by-question performance varies as the questions reflect the respective hypotheses. Eight observations are missing because their municipality cannot be identified. "Spanish Foreign Language" is a dummy variable with a value of 1 if the test-taker chose to take the Spanish package of questions in the Language test instead of the English one. IRT grades are the official grades that test-takers use to apply through SISU.

Appendix Table A6: Explained variance (R^2) for SES, Gender, and Conditional Ethnic Gaps based on different regressors

When regressing on:	SES Gap	Gender Gap	Cond. Ethnic Gap
Competencies FE Only	0.1057	0.2395	0.0756
IRT Parameters Only	0.2372	0.0238	0.281
Non-Contextual Features Only	0.0271	0.0712	0.0100
Competencies FE & IRT Parameters	0.4128	0.3458	0.3745
Competencies FE & Non-Contextual Features	0.1134	0.2513	0.0800
IRT Parameters & Non-Contextual Features	0.2660	0.0986	0.2897
All Regressors	0.4139	0.3510	0.3762

Note: This Table reports the R^2 values from regressions predicting the SES, gender, and conditional ethnic gaps using different sets of explanatory variables. The rows correspond to the various combinations of predictors used in the regressions: competencies fixed effects (FE), IRT parameters, and non-contextual features. The R^2 values indicate the proportion of the variance in each gap that is explained by the respective predictors. The final row ("All regressors") shows the R^2 when all predictors are included together, reflecting the combined explanatory power of competencies, IRT parameters, and non-contextual features.

Appendix Table A7: Top 20 Words Influencing the SES Gap (Effect in p.p.)

Words Reducing	Effect (p.p.)	Words Widening	Effect (p.p.)
temperature	0.7016	price	0.7482
work	0.7003	land	0.7143
acid	0.5827	work	0.7004
political	0.5495	sense	0.6153
black	0.5478	day	0.5624
car	0.5455	urban	0.5519
music	0.5452	model	0.5284
production	0.4857	demonstrate	0.5177
problem	0.4849	next	0.5097
Paul	0.4799	product	0.5082
industrial	0.4642	right	0.4976
study	0.4621	impact	0.4651
observe	0.4555	consumption	0.4330
height	0.4442	effect	0.4210
axis	0.4424	publish	0.4181
conception	0.4334	occupation	0.4160
consider	0.4315	person	0.4033
take	0.4255	information	0.4003
long	0.3963	walk	0.3933
century	0.3939	service	0.3921

Note: This table reports the top words that contribute to widening or reducing the SES gap when predicting the SES gap based on competences fixed effects and all other non-contextual features of the text, using the RIDGE model reported in Figure A12 panel (a). The words were translated from Portuguese using the Deepl API.

Appendix Table A8: Top 20 Words Influencing the Gender Gap (Effect in p.p.)

Words Reducing	**Effect (p.p.)**	**Words Widening**	**Effect (p.p.)**
life	1.1720	frame	1.3710
consider	1.1208	next	1.2903
acid	1.1016	follow	1.2432
temperature	1.0499	illustrate	1.1840
height	1.0009	center	1.1290
seek	0.9544	minute	1.1131
substance	0.9542	earth	1.1070
maintain	0.9387	day	1.0774
observe	0.9004	energy	1.0452
study	0.8314	rain	1.0030
test	0.8298	graph	0.9722
social	0.8161	rate	0.9404
show	0.8092	agriculture	0.9350
text	0.8044	country	0.9269
reader	0.7911	refer	0.8944
family	0.7639	student	0.8894
love	0.7443	occupation	0.8844
stage	0.7422	data	0.8824
work	0.7402	around	0.8738
production	0.7369	year	0.8653

Note: This table reports the top words that contribute to widening or reducing the gender gap when predicting the gender gap based on competences fixed effects and all other non-contextual features of the text, using the RIDGE model reported in Figure A12 panel (b). The words were translated from Portuguese using the Deepl API.

Appendix Table A9: Top 20 Words Influencing the Conditional Ethnic Gap (Effect in p.p.)

Words Reducing	**Effect (p.p.)**	**Words Widening**	**Effect (p.p.)**
resistance	0.1244	text	0.1535
reaction	0.1223	man	0.1296
political	0.1186	work	0.1277
Brazil	0.1147	walk	0.1163
temperature	0.1118	student	0.1115
mass	0.1054	next	0.1082
respectively	0.1042	manual labor	0.1078
exist	0.1003	impact	0.0985
long	0.1002	density	0.0951
music	0.0983	urban	0.0945
observe	0.0944	contain	0.0894
necessary	0.0914	day	0.0887
various	0.0912	provide	0.0822
cultural	0.0909	occupation	0.0768
night	0.0892	effect	0.0767
work	0.0839	fast	0.0762
regime	0.0822	slave	0.0757
study	0.0820	person	0.0747
metro	0.0806	house	0.0744
fabric	0.0765	consumption	0.0739

Note: This table reports the top words that contribute to widening or reducing the ethnic gap when predicting the ethnic gap based on competences fixed effects and all other non-contextual features of the text, using the RIDGE model reported in Figure A12 panel (c). The words were translated from Portuguese using the Deepl API.

Appendix Table A10: LASSO coefficients for SES Gap (p.p.)

SES Gap			
Features Reducing	**Coeff. (p.p.)**	**Features Widening**	**Coeff. (p.p.)**
LG-Body Language	4.8929	MT-Graph Interpretation	-5.3403
LG-Art Understanding	4.1617	MT-Number Meanings	-2.9924
Difficulty Parameter (B)	3.9257	MT-Quantity Variation	-2.9886
LG-Portuguese Usage	3.2958	MT-Probability Understanding	-2.7317
LG-Symbol Systems	3.0693	MT-Geometric Knowledge	-2.1508
LG-Opinion Comparison	2.8827	MT-Algebraic Modeling	-2.1259
LG-Text Analysis	2.7955	topic_14	-1.7810
LG-Tech Principles	2.6704	MT-Magnitude Measurement	-1.5010
LG-Information technologies	2.2676	SC-Institutional Role	-1.3413
NS-Chemistry Knowledge	1.0970	SC-Civic Foundations	-1.2687
Male Character	0.4625	SC-Cultural Identity	-1.2494
Sentiment Score	0.4540	SC-Geographic Transformation	-1.2268
NS-Physics Knowledge	0.4483	topic_53	-1.2249
Female Character	0.3917	SC-Society-Environment	-0.5973
# of Words	0.3661	SC-Tech Transformations	-0.4101
NS-Applied Technologies	0.3309	Discrimination Parameter (A)	-0.3239
topic_55	0.2219	Privileged Character	-0.2550
NS-Scientific Methods	0.1010	Underprivileged Character	-0.2356
Readability Score	0.0429	# of Words Sqr.	-0.1218

Note: This table reports the top features widening and reducing the SES gap when performing a LASSO, which includes competencies fixed effects, non-contextual features, contextual features, and the topics that were generated following the LDA by maximizing the coherence score. The fitting of the model is shown in Figure A13 panel (a). LG stands for the competencies in Language, MT for the competencies in Mathematics, NS for the competencies in Natural Sciences, and SC for the competencies in Social Sciences. Topics are a weighted vector of words estimated based on the occurrence and co-occurrence of words. An example is shown in Figure 2.

Appendix Table A11: LASSO coefficients for Gender Gap (p.p.)

Gender Gap			
Features Reducing	**Coeff. (p.p.)**	**Features Widening**	**Coeff. (p.p.)**
LG-Opinion Comparison	3.8887	MT-Graph Interpretation	-4.2957
LG-Information technologies	3.2011	MT-Quantity Variation	-3.6621
LG-Art Understanding	3.1607	MT-Geometric Knowledge	-3.4305
LG-Portuguese Usage	2.8879	MT-Number Meanings	-3.0187
LG-Body Language	2.7742	MT-Algebraic Modeling	-2.4318
LG-Text Analysis	2.6098	topic_24	-2.3613
LG-Symbol Systems	2.5439	MT-Probability Understanding	-2.1469
LG-Tech Principles	2.4912	MT-Magnitude Measurement	-2.0881
Female Character	1.0106	topic_58	-1.9921
Guessing Parameter(C)	0.8710	topic_47	-1.5356
SC-Civic Foundations	0.7961	SC-Society-Environment	-1.3823
topic_49	0.6510	topic_53	-1.3548
SC-Cultural Identity	0.5916	NS-Physics Knowledge	-1.2853
# Words	0.4519	SC-Geographic Transformation	-1.2589
Difficulty Parameter (B)	0.4269	topic_23	-1.0691
Difficulty Parameter (B) Sqr.	0.3553	NS-Science Development	-0.8247
topic_1	0.3356	topic_64	-0.5933
NS-Organism-Environment	0.2622	Privileged Character	-0.2780
SC-Tech Transformations	0.1421	NS-Applied Technologies	-0.2409
Male Character	0.1328	Grounded	-0.2401

Note: This table reports the top features widening and reducing the Gender gap when performing a LASSO, which includes competencies fixed effects, non-contextual features, contextual features, and the topics that were generated following the LDA by maximizing the coherence score. The fitting of the model is shown in Figure A13 panel (b). LG stands for the competencies in Language, MT for the competencies in Mathematics, NS for the competencies in Natural Sciences, and SC for the competencies in Social Sciences. Topics are a weighted vector of words estimated based on the occurrence and co-occurrence of words. An example is shown in Figure 2.

Appendix Table A12: LASSO coefficients for Ethnic Gap (p.p.)

Ethnicity (Color) Gap			
Features Reducing	**Coeff. (p.p.)**	**Features Widening**	**Coeff. (p.p.)**
Difficulty Parameter (B)	1.7931	MT-Graph Interpretation	-1.3113
LG-Body Language	1.1134	MT-Probability Understanding	-0.5080
LG-Symbol Systems	0.9114	MT-Number Meanings	-0.5004
LG-Art Understanding	0.8855	topic_14	-0.4644
LG-Tech Principles	0.7314	SC-Society-Environment	-0.4394
LG-Text Analysis	0.6391	SC-Cultural Identity	-0.4036
LG-Portuguese Usage	0.5960	MT-Quantity Variation	-0.3814
LG-Opinion Comparison	0.5229	MT-Geometric Knowledge	-0.3508
LG-Information technologies	0.4978	SC-Institutional Role	-0.3478
NS-Chemistry Knowledge	0.4672	MT-Algebraic Modeling	-0.2612
Male Character	0.1756	SC-Civic Foundations	-0.2473
topic_15	0.1569	SC-Geographic Transformation	-0.2470
NS-Organism-Environment	0.1274	SC-Tech Transformations	-0.2426
Sentiment Score	0.1133	Difficulty Parameter (B) Sqr.	-0.1760
Tables	0.0890	Underprivileged Character	-0.0753
Discrimination Parameter (A)	0.0645	MT-Magnitude Measurement	-0.0627
# of Words	0.0321	Figures	-0.0298
Readability Score	0.0071	# of Words Sqr.	-0.0212
NS-Physics Knowledge	0.0124	Grounded	-0.0191

Note: This table reports the top features widening and reducing the Ethnic gap when performing a LASSO, which includes competencies fixed effects, non-contextual features, contextual features, and the topics that were generated following the LDA by maximizing the coherence score. The fitting of the model is shown in Figure A13 panel (c). LG stands for the competencies in Language, MT for the competencies in Mathematics, NS for the competencies in Natural Sciences, and SC for the competencies in Social Sciences. Topics are a weighted vector of words estimated based on the occurrence and co-occurrence of words. An example is shown in Figure 2

Appendix Table A13: Hypotheses and their impact in Performance (p.p) splitting by subjects

	All Test-takers				Only Q4 Ability			
	SC Correct (0/1)	NS Correct (0/1)	Lang. Correct (0/1)	Math Correct (0/1)	SC Correct (0/1)	NS Correct (0/1)	Lang. Correct (0/1)	Math Correct (0/1)
SES Wd	1.188***	2.589***	4.208***	0.514***	3.890***	2.269***	7.402***	-0.122
	(0.098)	(0.172)	(0.118)	(0.063)	(0.197)	(0.307)	(0.202)	(0.118)
SES Wd × Low Income	-0.629***	-1.382***	-0.807***	-1.695***	-2.012***	2.632***	1.648***	-1.712***
	(0.173)	(0.298)	(0.211)	(0.108)	(0.550)	(0.843)	(0.571)	(0.349)
SES Rd	-1.182***	-0.155***	1.275***	-1.704***	-0.961***	0.784***	1.104***	-3.114***
	(0.086)	(0.055)	(0.074)	(0.058)	(0.160)	(0.105)	(0.132)	(0.113)
SES Rd × Low Income	0.032	0.845***	-0.748***	1.967***	0.118	0.403	-0.945***	-0.362
	(0.152)	(0.096)	(0.128)	(0.099)	(0.442)	(0.291)	(0.355)	(0.324)
Gender Wd	0.125	1.231***	1.441***	-0.635***	0.264	1.592***	3.260***	-0.823***
	(0.134)	(0.069)	(0.139)	(0.071)	(0.249)	(0.131)	(0.270)	(0.136)
Gender Wd × Girl	0.067	-0.446***	-0.566***	1.097***	-0.152	-1.941***	-0.713**	-0.546***
	(0.171)	(0.085)	(0.175)	(0.087)	(0.344)	(0.176)	(0.353)	(0.191)
Gender Rd	-1.604***	-2.112***	0.323***	-1.506***	-0.599***	-3.785***	0.587***	-2.672***
	(0.070)	(0.136)	(0.063)	(0.277)	(0.130)	(0.265)	(0.117)	(0.506)
Gender Rd × Girl	0.926***	0.750***	0.362***	2.264***	1.190***	1.121***	0.556***	-0.584
	(0.089)	(0.175)	(0.078)	(0.355)	(0.177)	(0.382)	(0.149)	(0.765)
Ethnic Wd	-0.227***	-2.012***	0.027	0.865***	-1.630***	-1.588***	0.505***	1.342***
	(0.072)	(0.115)	(0.082)	(0.095)	(0.126)	(0.211)	(0.139)	(0.166)
Ethnic Wd × Non-White	-0.016	0.355**	-0.275***	-0.626***	-0.016	-0.367	-0.414**	-0.241
	(0.089)	(0.146)	(0.099)	(0.122)	(0.182)	(0.314)	(0.197)	(0.255)
Ethnic Rd	1.848***		-0.937***	-7.339***	1.218***		-0.420**	-11.028***
	(0.141)		(0.114)	(0.408)	(0.246)		(0.191)	(0.850)
Ethnic Rd × Non-White	-0.474**		0.458***	3.501***	0.078		0.891***	-0.670
	(0.185)		(0.149)	(0.528)	(0.385)		(0.296)	(1.300)
Const.	55.231***	43.657***	38.736***	30.838***	68.199***	63.281***	43.337***	57.829***
	(0.104)	(0.110)	(0.089)	(0.121)	(0.209)	(0.223)	(0.168)	(0.256)
Item's Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Indv. X Subject FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Other FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Obs.	5,759,203	5,679,215	4,919,323	5,649,223	1,441,176	1,417,656	1,227,881	1,412,299
Indiv.	129,982	129,982	129,982	129,982	32,466	32,466	32,466	32,466

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

Note: This table presents the results from estimating equation 2, with the sample split by subject: SC (Social Sciences), NS (Natural Sciences), Lang. (Language), and Math (Mathematics). The left panel shows results for all test-takers, while the right panel focuses on those in the highest quartile of ability, who are more likely to gain admission through SISU. The item controls include factors such as length, readability, sentiment score, groundedness, and IRT parameters. Additional fixed effects cover competence, region, year, and question order. There are no questions labeled under the reducing channel in Natural Sciences. Standard errors are clustered at the individual level.

Appendix Table A14: Impact of Hypotheses on Performance (p.p) by Distance (only Q4)

	SES Gap	Gender Gap	Ethnic Gap
	Correct (0/100)	Correct (0/100)	Correct (0/100)
Widening Effect	2.836***	-0.372***	0.036
	(0.085)	(0.084)	(0.076)
Adj. Difficulty	-9.726***	-10.047***	-9.861***
	(0.057)	(0.064)	(0.062)
Widening × Dimension	-0.527**	-1.300***	-0.226**
	(0.257)	(0.123)	(0.113)
Widening × Adj. Difficulty	0.325***	0.164**	-0.106
	(0.072)	(0.073)	(0.068)
Adj. Difficulty × Dimension	0.633***	0.146**	0.661***
	(0.108)	(0.065)	(0.067)
Widening × Dimension × Adj. Difficulty	0.198	0.546***	-0.632***
	(0.248)	(0.111)	(0.109)
Reducing Effect	-0.600***	-0.722***	0.111
	(0.059)	(0.083)	(0.157)
Reducing × Dimension	-0.719***	0.790***	0.039
	(0.187)	(0.113)	(0.243)
Reducing X Adj. Difficulty	0.545***	2.554***	0.918***
	(0.055)	(0.086)	(0.158)
Reducing × Dimension × Adj. Difficulty	1.059***	-0.374***	-0.087
	(0.184)	(0.121)	(0.255)
Const.	45.083***	45.607***	45.299***
	(0.072)	(0.073)	(0.072)
Item's Controls	Yes	Yes	Yes
Indiv X Subject FE	Yes	Yes	Yes
Other Fixed Effects	Yes	Yes	Yes
Obs.	5,499,012	5,499,012	5,499,012
Indiv.	32,423	32,423	32,423

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

Note: This table presents the results from estimating equation 2, incorporating the adjusted difficulty parameter. This parameter is measured at the individual-question level and captures the difference between the test-taker's latent ability and the difficulty of the question. It is calculated by subtracting the test-taker's θ value from the item's b parameter, meaning positive values indicate more challenging items from the perspective of the individual. The analysis is restricted to test-takers in the top quartile of abilities. The results are displayed by dimension, following the same structure as Table 1. Standard errors are clustered at the individual level and shown in parentheses

Appendix Table A15: Hypotheses and their impact in Performance (p.p) by Stereotyped Item (only Gender Gap)

	All	Ability Quartiles			
	Correct (0/100)	Correct (0/100)	Correct (0/100)	Correct (0/100)	Correct (0/100)
Gender Wd	-0.099	0.154	-0.073	0.316**	0.121
	(0.065)	(0.130)	(0.133)	(0.131)	(0.124)
Gender Wd × Girl	-0.215**	-0.257	-0.712***	-1.316***	-1.045***
	(0.083)	(0.158)	(0.165)	(0.169)	(0.175)
Stereotyped	1.816***	-0.824***	0.230**	2.022***	4.308***
	(0.046)	(0.088)	(0.092)	(0.090)	(0.088)
Gender Wd × Stereotyped	-0.055	1.168***	0.926***	-0.255	-0.854***
	(0.084)	(0.169)	(0.171)	(0.166)	(0.156)
Girl × Stereotyped	-2.643***	-1.124***	-1.769***	-2.288***	-2.379***
	(0.054)	(0.097)	(0.103)	(0.104)	(0.109)
Gender Wd × Girl × Stereotyped	0.917***	0.398*	0.722***	1.559***	0.170
	(0.108)	(0.205)	(0.213)	(0.218)	(0.226)
Gender Rd	-0.436***	-0.651***	-0.455***	-0.384***	-0.055
	(0.050)	(0.095)	(0.103)	(0.101)	(0.092)
Gender Rd × Girl	0.417***	0.442***	0.687***	0.637***	0.690***
	(0.064)	(0.120)	(0.130)	(0.130)	(0.124)
Gender Rd × Stereotyped	-1.810***	0.425**	-1.607***	-2.307***	-3.247***
	(0.090)	(0.173)	(0.185)	(0.183)	(0.169)
Gender Rd × Girl × Stereotyped	0.580***	0.059	0.225	-0.176	0.076
	(0.116)	(0.215)	(0.234)	(0.240)	(0.235)
Const.	41.762***	24.528***	40.666***	47.055***	55.905***
	(0.057)	(0.105)	(0.113)	(0.113)	(0.110)
Item's Controls	Yes	Yes	Yes	Yes	Yes
Indiv. X Subject FE	Yes	Yes	Yes	Yes	Yes
Other FEs	Yes	Yes	Yes	Yes	Yes
Obs.	22,006,964	5,508,994	5,501,288	5,497,670	5,499,012

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

Note: This table presents the results from estimating equation 2, incorporating the Stereotyped dummy for each question. This hand-constructed dummy takes a value of 1 if the sub-competence (one of the 30 competencies that describe the objectives of questions across the four subjects) is categorized as stereotyped. A sub-competence is considered stereotyped if its effect on widening the gender gap is above the mean performance gap, similar to the analysis shown in Figure 1. Of the total number of questions, 47% (1,035) are classified as stereotyped. Standard errors are clustered at the individual level and shown in parentheses.

Appendix Table A16: Hypotheses and their impact in Performance (p.p) by Mirroring Effect on Gender Gaps

	All	Ability Quartiles			
	Correct (0/100)	Correct (0/100)	Correct (0/100)	Correct (0/100)	Correct (0/100)
Gender Wd	0.096**	0.981***	0.662***	0.346***	-0.047
	(0.045)	(0.090)	(0.091)	(0.090)	(0.085)
Gender Wd × Girl	0.236***	-0.048	-0.300***	-0.426***	-1.069***
	(0.057)	(0.107)	(0.111)	(0.113)	(0.120)
Female Character	0.627***	0.244**	0.646***	0.873***	0.917***
	(0.055)	(0.109)	(0.113)	(0.111)	(0.103)
Gender Wd × Female Character	-2.109***	-1.092***	-1.655***	-1.962***	-4.059***
	(0.139)	(0.271)	(0.290)	(0.279)	(0.252)
Girl× Female Character	0.820***	0.707***	0.944***	1.079***	0.784***
	(0.071)	(0.136)	(0.142)	(0.144)	(0.145)
Gender Wd × Girl × Female Character	-0.051	0.250	-0.318	-0.445	0.376
	(0.182)	(0.336)	(0.366)	(0.369)	(0.366)
Gender Rd	-0.395***	-0.290***	-0.356***	-0.279***	-0.204**
	(0.048)	(0.092)	(0.100)	(0.098)	(0.088)
Gender Rd × Girl	0.625***	0.575***	0.691***	0.699***	0.909***
	(0.062)	(0.115)	(0.126)	(0.127)	(0.121)
Gender Rd × Female Character	-2.624***	-1.317***	-2.565***	-3.503***	-3.621***
	(0.094)	(0.185)	(0.198)	(0.193)	(0.178)
Gender Rd × Girl × Female Character	-0.120	-0.624***	0.152	-0.362	-0.116
	(0.123)	(0.235)	(0.250)	(0.252)	(0.247)
Const.	41.685***	23.594***	39.946***	47.127***	57.345***
	(0.055)	(0.101)	(0.108)	(0.107)	(0.105)
Item's Controls	Yes	Yes	Yes	Yes	Yes
Indiv X Subject FE	Yes	Yes	Yes	Yes	Yes
Other FEs	Yes	Yes	Yes	Yes	Yes
Obs.	22,006,964	5,508,994	5,501,288	5,497,670	5,499,012

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

Note: This table presents the results from estimating equation 2, incorporating the Female Character variable. This is a variable that has been constructed by asking ChatGPT whether the questions depict a female character based on the usage of names, pronouns or descriptions. A total of 301 (13%) items include a female character, distributed in all subjects as 33% in social sciences, 12% in natural sciences, 39% in language and 15% in mathematics. Standard errors are clustered at the individual level and shown in parentheses.

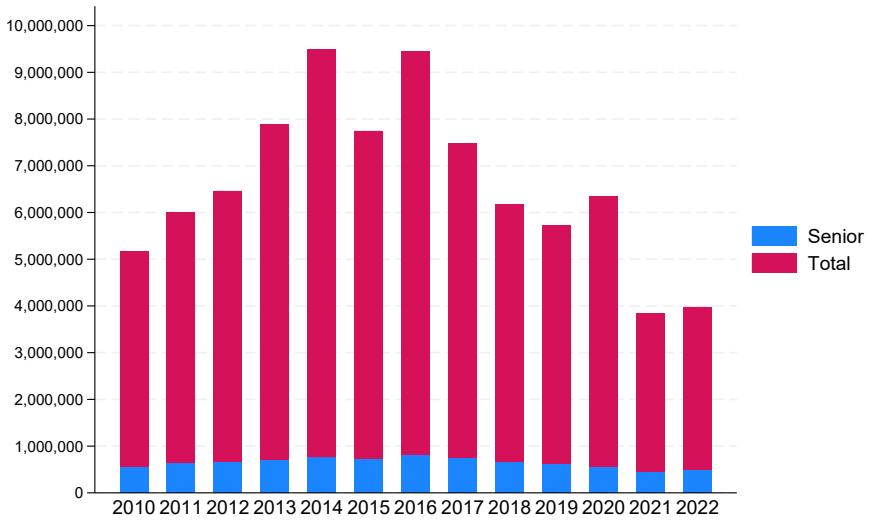
Appendix Table A17: Hypotheses and their impact in Performance (p.p) by Mirroring Effect on SES Gaps

	All	Ability Quartiles			
	Correct (0/100)	Correct (0/100)	Correct (0/100)	Correct (0/100)	Correct (0/100)
SES Wd	0.797***	-1.098***	-0.576***	0.487***	2.634***
	(0.047)	(0.093)	(0.093)	(0.092)	(0.087)
SES Wd × Low Income	-1.236***	0.327**	0.083	-0.271	-0.456*
	(0.084)	(0.138)	(0.154)	(0.180)	(0.258)
Low SES Character	-1.181***	-1.903***	-1.701***	-1.176***	-0.377***
	(0.058)	(0.118)	(0.119)	(0.114)	(0.105)
SES Wd × Low SES Character	3.586***	4.769***	5.037***	3.713***	1.450***
	(0.172)	(0.341)	(0.329)	(0.343)	(0.352)
Low Income × Low SES Character	0.326***	0.544***	0.664***	-0.022	0.606*
	(0.107)	(0.180)	(0.202)	(0.231)	(0.314)
SES Wd × Low Income × Low SES Character	-2.181***	-2.613***	-2.648***	-2.029***	-2.375**
	(0.315)	(0.522)	(0.567)	(0.712)	(1.112)
SES Rd	-0.487***	0.001	0.095	-0.192***	-0.795***
	(0.032)	(0.065)	(0.066)	(0.064)	(0.060)
SES Rd × Low Income	0.952***	0.200**	0.031	0.360***	0.042
	(0.057)	(0.097)	(0.108)	(0.124)	(0.172)
SES Rd × Low SES Character	1.932***	0.841***	0.085	1.230***	4.152***
	(0.122)	(0.249)	(0.246)	(0.238)	(0.229)
SES Rd× Low Income × Low SES Character	-1.721***	-0.009	-0.334	0.375	-0.671
	(0.225)	(0.379)	(0.415)	(0.488)	(0.705)
Const.	41.783***	23.865***	40.241***	47.287***	57.161***
	(0.054)	(0.100)	(0.107)	(0.106)	(0.104)
Item's Controls	Yes	Yes	Yes	Yes	Yes
Indiv X Subject FE	Yes	Yes	Yes	Yes	Yes
Other FEs	Yes	Yes	Yes	Yes	Yes
Obs.	22,006,964	5,508,994	5,501,288	5,497,670	5,499,012

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

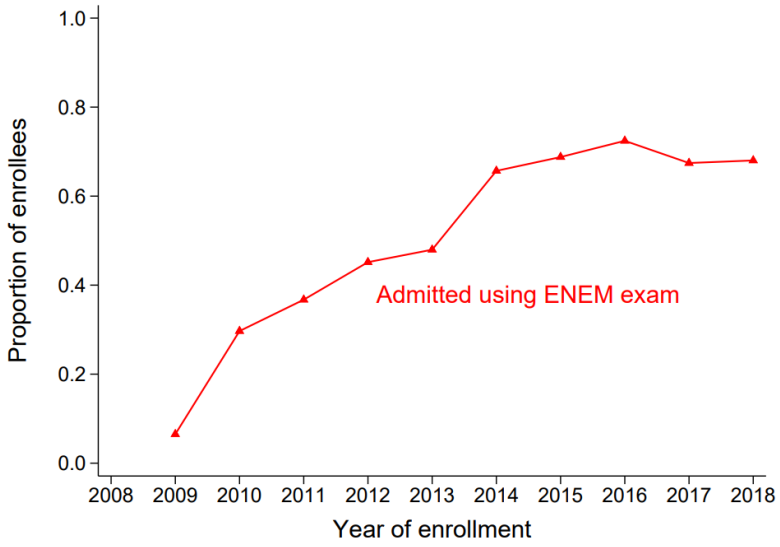
Note: This table presents the results from estimating equation 2, incorporating the Low-SES Character variable. This is a variable that has been hand coded considering whether the questions depict a low-SES character based on the usage of names, the historical backgorund of celebrities or descriptions of the context in which they live. A total of 118 (5.4%) items include a Low-SES character, distributed in all subjects as 39% in social sciences, 8.5% in natural sciences, 34% in language and 18.5% in mathematics. Standard errors are clustered at the individual level and shown in parentheses.

A.2 Appendix Figures



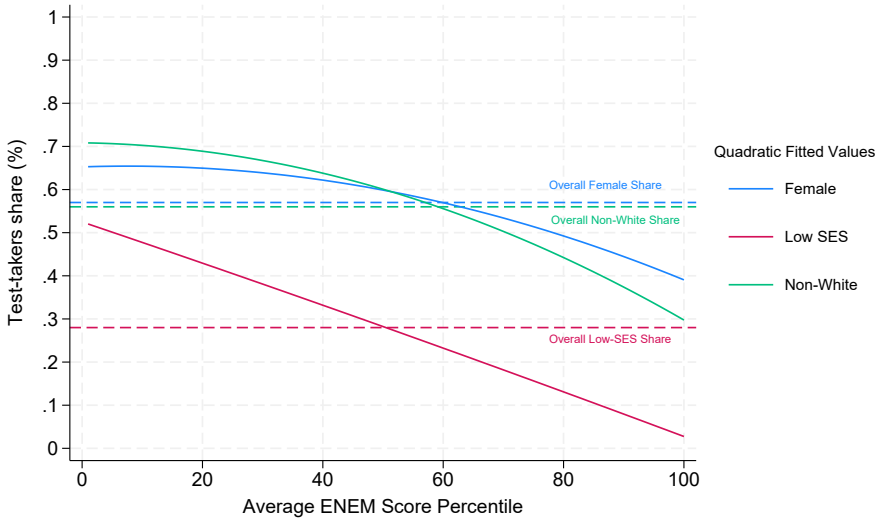
Note: This figure shows the yearly variation in the total number of test-takers who enroll to take the ENEM, as well as the share of the sample corresponding to high-school seniors from the regular track in a given year. While the total number of test-takers varies by year, the share of seniors remains more stable. The rapid increase in total test-takers between 2010 and 2016 is explained by the growing number of universities that voluntarily transitioned to the SISU system. The differences between these measures are due to the presence of a population of already-graduated individuals interested in competing for university slots.

Appendix Figure A1: Proportion of test-takers that are high-school seniors each year relative to the total of test takers



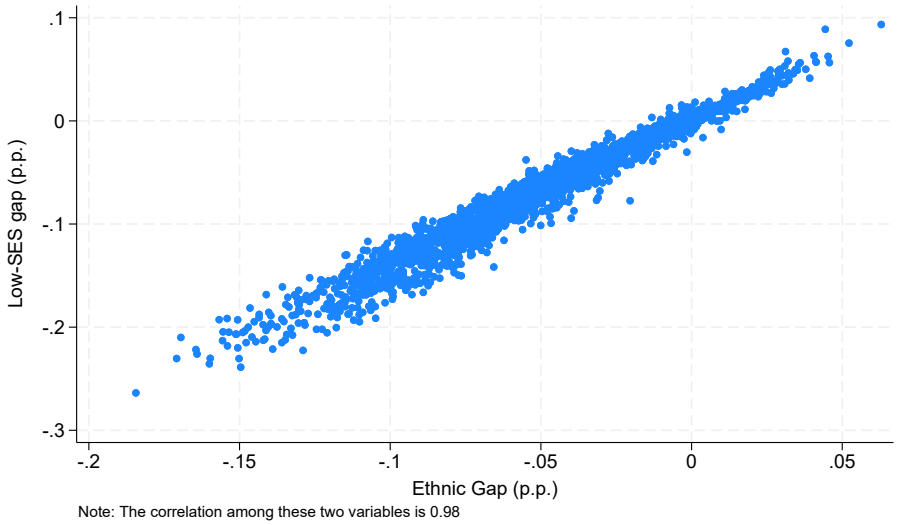
Note: This figure, borrowed from Reyes et al. (2023), shows the increasing share of enrollees in public universities admitted through the ENEM and SISU. An upward trend can be observed from the beginning, which stabilizes around 2017.

Appendix Figure A2: Proportion of test-takers that were admitted in a Federal University by ENEM



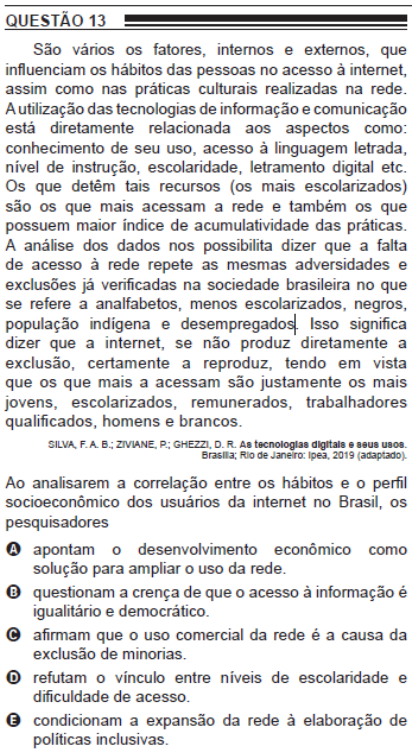
Note: This figure shows how the share of specific test-takers by gender, SES, and ethnicity varies across performance percentiles in the average ENEM. Percentiles are estimated for each year based on the simple average of the four ENEM subjects, with the 100th percentile representing the top performers for that year. Quadratic fittings are provided. The dashed lines represent the sample averages as benchmarks: 27% for low-SES, 56% for non-white, and 57% for female test-takers. As performance percentiles increase, the share of these test-takers declines dramatically, falling to 3% for low-SES, 30% for non-white, and 40% for female test-takers at the top percentile. These gaps have significant implications for group composition and are well aligned with the empirical findings in the literature.

Appendix Figure A3: Distribution of the share of test-takers by performance in the ENEM by dimension



Note: This figure shows the question-by-question (2,201 observations) correlation between SES gaps and ethnicity gaps, as estimated in Equation 1, with a correlation value of 0.98. This high correlation suggests that these two dimensions are closely linked in the Brazilian context. To avoid circularity, I decided to perform the hypothesis generation step by observing which elements increase the ethnicity gaps, conditional on being low-SES.

Appendix Figure A4: Correlation among Low-SES gap and Ethnic Gap



QUESTÃO 13

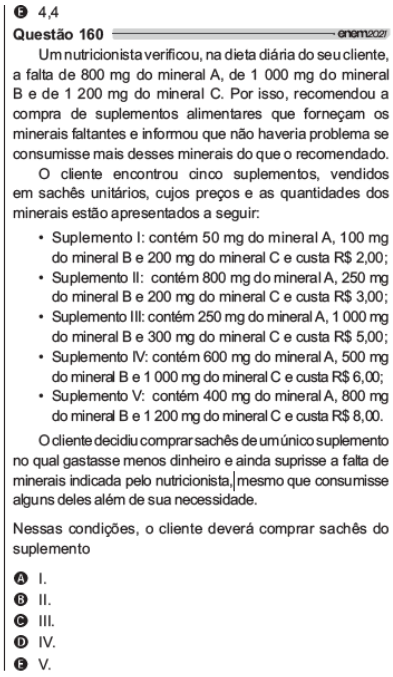
São vários os fatores, internos e externos, que influenciam os hábitos das pessoas no acesso à internet, assim como nas práticas culturais realizadas na rede. A utilização das tecnologias de informação e comunicação está diretamente relacionada aos aspectos como: conhecimento de seu uso, acesso à linguagem letrada, nível de instrução, escolaridade, letramento digital etc. Os que detêm tais recursos (os mais escolarizados) são os que mais acessam a rede e também os que possuem maior índice de acumulatividade das práticas. A análise dos dados nos possibilita dizer que a falta de acesso à rede repete as mesmas adversidades e exclusões já verificadas na sociedade brasileira no que se refere a analfabetos, menos escolarizados, negros, população indígena e desempregados. Isso significa dizer que a internet, se não produz diretamente a exclusão, certamente a reproduz, tendo em vista que os que mais a acessam são justamente os mais jovens, escolarizados, remunerados, trabalhadores qualificados, homens e brancos.

SILVA, F. A. B.; ZIVIANE, P.; GHEZZI, D. R. **As tecnologias digitais e seus usos**. Brasília; Rio de Janeiro: Ipea, 2019 (adaptado).

Ao analisarem a correlação entre os hábitos e o perfil socioeconômico dos usuários da internet no Brasil, os pesquisadores

A apontam o desenvolvimento econômico como solução para ampliar o uso da rede.
B questionam a crença de que o acesso à informação é igualitário e democrático.
C afirmam que o uso comercial da rede é a causa da exclusão de minorias.
D refutam o vínculo entre níveis de escolaridade e dificuldade de acesso.
E condicionam a expansão da rede à elaboração de políticas inclusivas.

(a) Language ID 120622



E 4,4

Questão 160 enem2021

Um nutricionista verificou, na dieta diária do seu cliente, a falta de 800 mg do mineral A, de 1 000 mg do mineral B e de 1 200 mg do mineral C. Por isso, recomendou a compra de suplementos alimentares que forneçam os minerais faltantes e informou que não haveria problema se consumisse mais desses minerais do que o recomendado.

O cliente encontrou cinco suplementos, vendidos em sachês unitários, cujos preços e as quantidades dos minerais estão apresentados a seguir:

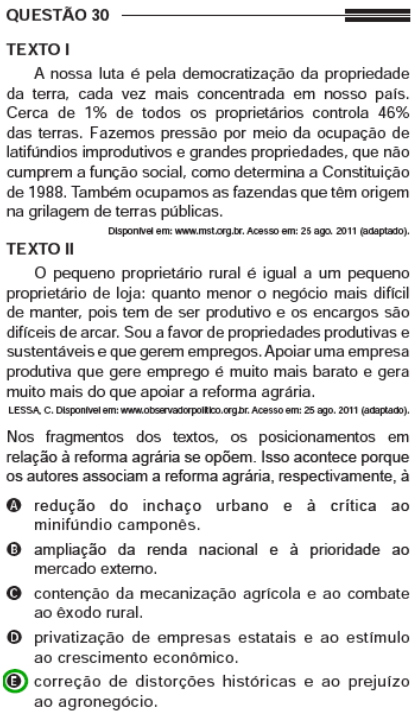
- Suplemento I: contém 50 mg do mineral A, 100 mg do mineral B e 200 mg do mineral C e custa R$ 2,00;
- Suplemento II: contém 800 mg do mineral A, 250 mg do mineral B e 200 mg do mineral C e custa R$ 3,00;
- Suplemento III: contém 250 mg do mineral A, 1 000 mg do mineral B e 300 mg do mineral C e custa R$ 5,00;
- Suplemento IV: contém 600 mg do mineral A, 500 mg do mineral B e 1 000 mg do mineral C e custa R$ 6,00;
- Suplemento V: contém 400 mg do mineral A, 800 mg do mineral B e 1 200 mg do mineral C e custa R$ 8,00.

O cliente decidiu comprar sachês de um único suplemento no qual gastasse menos dinheiro e ainda suprisse a falta de minerais indicada pelo nutricionista, mesmo que consumisse alguns deles além de sua necessidade.

Nessas condições, o cliente deverá comprar sachês do suplemento

A I.
B II.
C III.
D IV.
E V.

(b) Mathematics ID 1179022



QUESTÃO 30

TEXTO I

A nossa luta é pela democratização da propriedade da terra, cada vez mais concentrada em nosso país. Cerca de 1% de todos os proprietários controla 46% das terras. Fazemos pressão por meio da ocupação de latifúndios improdutivos e grandes propriedades, que não cumprem a função social, como determina a Constituição de 1988. Também ocupamos as fazendas que têm origem na grilagem de terras públicas.

Disponível em: www.mst.org.br. Acesso em: 25 ago. 2011 (adaptado).

TEXTO II

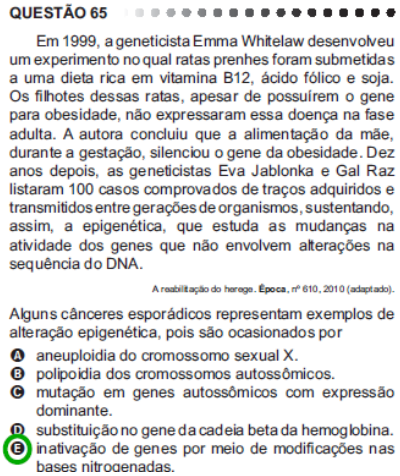
O pequeno proprietário rural é igual a um pequeno proprietário de loja: quanto menor o negócio mais difícil de manter, pois tem de ser produtivo e os encargos são difíceis de arcar. Sou a favor de propriedades produtivas e sustentáveis e que gerem empregos. Apoiar uma empresa produtiva que gere emprego é muito mais barato e gera muito mais do que apoiar a reforma agrária.

LESSA, C. Disponível em: www.observadorpolitico.org.br. Acesso em: 25 ago. 2011 (adaptado).

Nos fragmentos dos textos, os posicionamentos em relação à reforma agrária se opõem. Isso acontece porque os autores associam a reforma agrária, respectivamente, à

A redução do inchaço urbano e à crítica ao minifúndio camponês.
B ampliação da renda nacional e à prioridade ao mercado externo.
C contenção da mecanização agrícola e ao combate ao êxodo rural.
D privatização de empresas estatais e ao estímulo ao crescimento econômico.
E correção de distorções históricas e ao prejuízo ao agronegócio.

(c) Social Sciences ID 51402



QUESTÃO 65

Em 1999, a geneticista Emma Whitelaw desenvolveu um experimento no qual ratas prenhes foram submetidas a uma dieta rica em vitamina B12, ácido fólico e soja. Os filhotes dessas ratas, apesar de possuírem o gene para obesidade, não expressaram essa doença na fase adulta. A autora concluiu que a alimentação da mãe, durante a gestação, silenciou o gene da obesidade. Dez anos depois, as geneticistas Eva Jablonka e Gal Raz listaram 100 casos comprovados de traços adquiridos e transmitidos entre gerações de organismos, sustentando, assim, a epigenética, que estuda as mudanças na atividade dos genes que não envolvem alterações na sequência do DNA.

A reabilitação do herege. **Época**, nº 610, 2010 (adaptado).

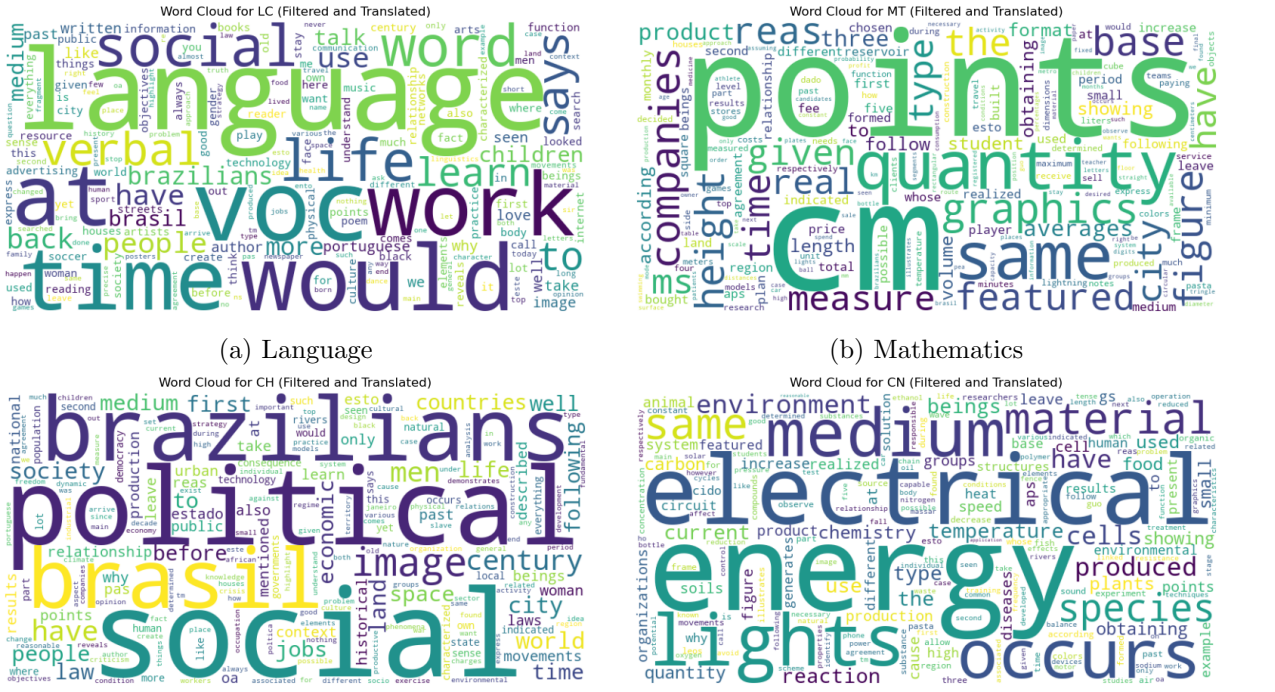
Alguns cânceres esporádicos representam exemplos de alteração epigenética, pois são ocasionados por

A aneuploidia do cromossomo sexual X.
B polipoidia dos cromossomos autossômicos.
C mutação em genes autossômicos com expressão dominante.
D substituição no gene da cadeia beta da hemoglobina.
E inativação de genes por meio de modificações nas bases nitrogenadas.

(d) Natural Sciences ID 706160

Note: These figures display examples of questions as displayed in the booklet for all of the subjects.

Appendix Figure A5: Some Items examples by Subjects

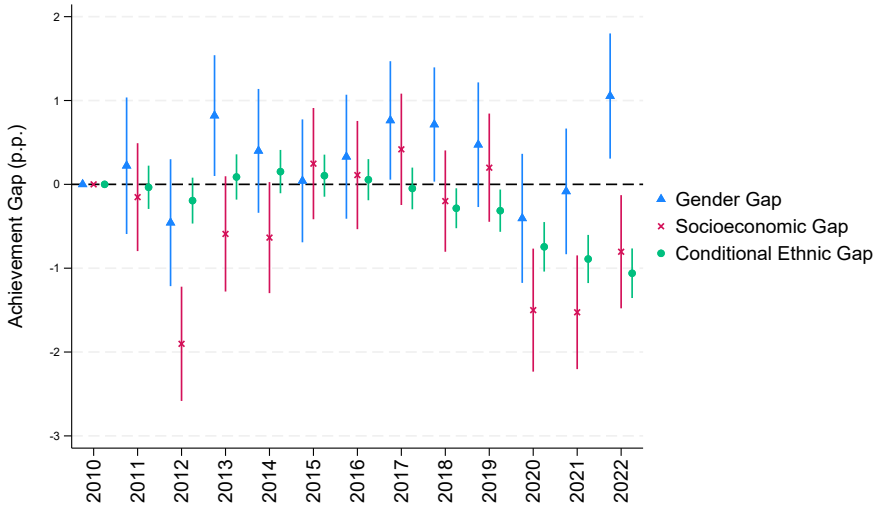


(a) Language (b) Mathematics

(c) Social Sciences (d) Natural Sciences

Note: These figures display the frequency of the most common words used to contextualize questions by subject. The estimation excludes the 50 most frequent words and applies TF-IDF weighting. The original language is Portuguese, and all words were translated using the DeepL API.

Appendix Figure A6: Words clouds by Subject



Note: This graph shows the coefficients for the year fixed effects when regressing each of the gaps for the 2,201 questions in my sample. It can be observed that no consistent yearly trends emerge. Only some statistically significant gaps appear for low-SES in 2012 and during the pandemic in 2020 and 2021. This evidence is important as it shows that test designers are not intentionally attempting to reduce these gaps, for instance, by altering the yearly composition of the questions. Such changes would pose a threat to my identification strategy. Confidence intervals are shown at the 95% level.

Appendix Figure A7: Yearly fixed effects on achievement gap by dimension

Prompt for Contextual Driver Analysis:
I am working on a project to identify the contextual drivers behind performance gaps in test scores. I have results from two independent methods: topic modeling and unigrams. The goal is to explore the convergences between these methods and understand what contextual elements are driving variations in performance, focusing on the topics and words that predict the highest variation in the [insert channel] of the [insert dimension] gap.
I will provide the following:

- The top 100 words for each topic that, using Lasso regression, are identified as predicting the highest variation in the [insert channel] of the [insert dimension] gap.
- The top 50 unigrams (individual words) that, using Ridge regression, predict the most variation in the [insert channel] of the [insert dimension] gap.

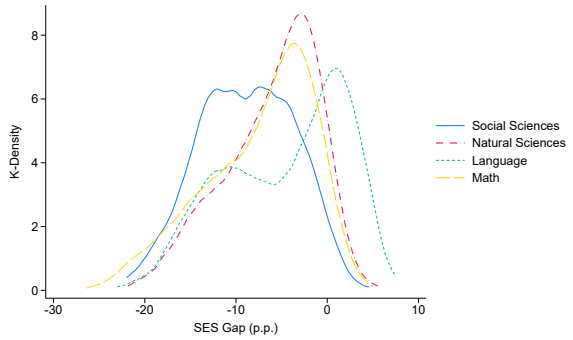
Your task: Interpret these word distributions to identify potential contextual drivers. The challenge is to distinguish between "context" and "content." For instance, a word like "atom" might represent a specific context (e.g., scientific framing) while also signaling content related to chemistry, where performance differences are expected.
Considerations:

- Separate content from contextual drivers as much as possible.
- Look for elements like framing, terminology, emotional triggers, or specific references that could affect test-takers' understanding or approach.
- Compare the top 50 unigrams predicting the [insert channel] of the [insert dimension] gap alone with the top 100 words composing each topic, identifying commonalities or unique drivers across both sets.
- Provide hypotheses for each topic and word set, suggesting what contextual elements could explain the observed [insert dimension] gap in the [insert channel]. These hypotheses should be testable statistically.

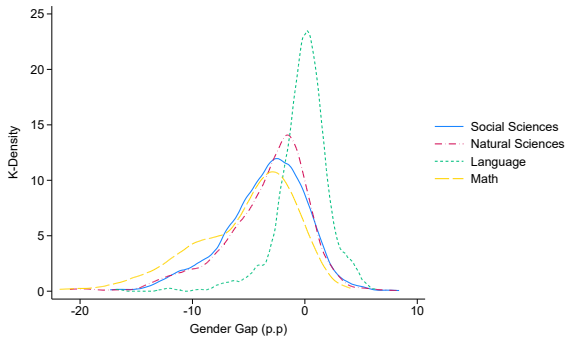
I aim to test these hypotheses statistically by linking them to performance differences across demographic groups (e.g., gender, socioeconomic status, ethnicity). Therefore, the interpretations should focus on potential contextual factors that could vary in impact across these groups.

Note: This figure shows the exact text provided to ChatGPT as a prompt. Each prompt is run independently by dimension and by channel (i.e., six separate prompts in total). I provide both the results of the bag-of-words exercise and the topic modeling. The bag-of-words includes the top 50 unigrams from the RIDGE model, along with their coefficients predicting the gap in the respective dimension and channel. The topic model includes the composition of topics, specifically the 100 most important words and their weights, from the topics that showed high predictive power in the Lasso model. Based on this combined input, ChatGPT generates a single hypothesis that captures the common contextual driver behind the performance gaps.

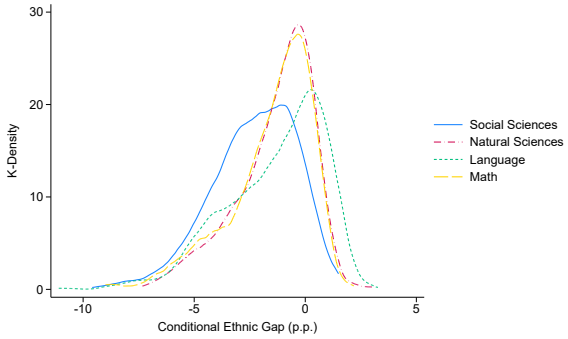
Appendix Figure A8: ChatGPT prompt for Hypothesis Generation



(a) K-Density of Socioeconomic Performance Gaps (p.p.)



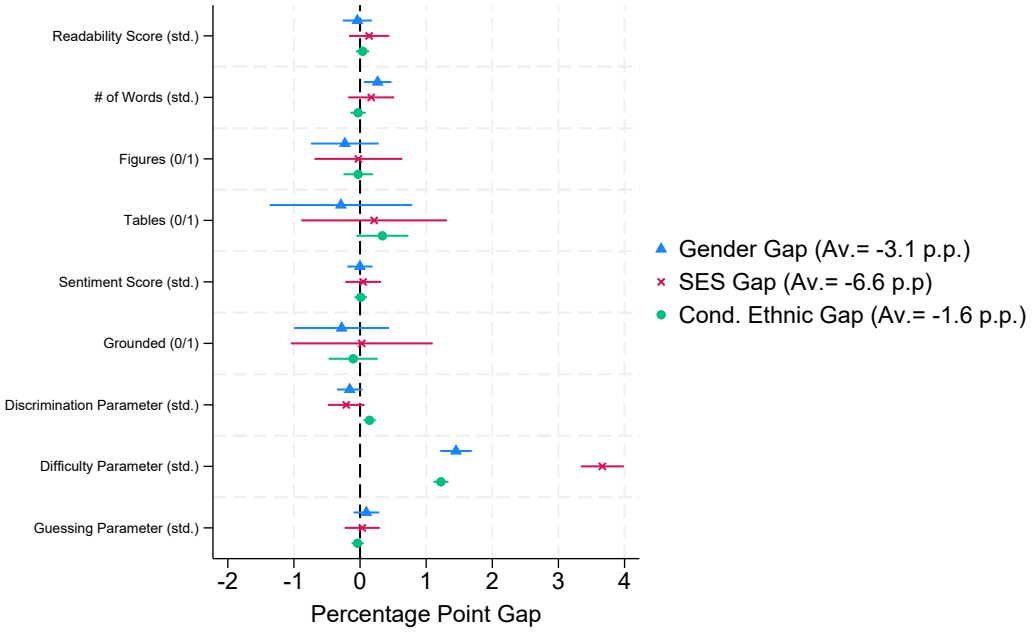
(b) K-Density of Gender Performance Gaps (p.p.)



(c) K-Density of Ethnic Performance Gap conditional on LowSES (p.p)

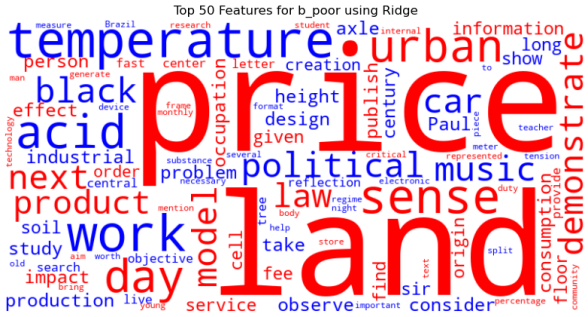
Note: These plots show the k-density estimations for the distribution of gaps at the question-by-question level, as outlined in Equation 1 for each dimension. Panel (a) shows the distributions for the SES gap by subject, Panel (b) shows the distributions for the gender gap by subject, and Panel (c) shows the distributions for the conditional ethnicity gap (conditional on being low-SES). Significant variability can be observed across all dimensions and subjects. On average, SES gaps are 6.8 p.p., gender gaps are 3.2 p.p., and conditional ethnicity gaps are 1.7 p.p., with a 32% correct-answer rate overall.

Appendix Figure A9: performance Gaps density by different tagging

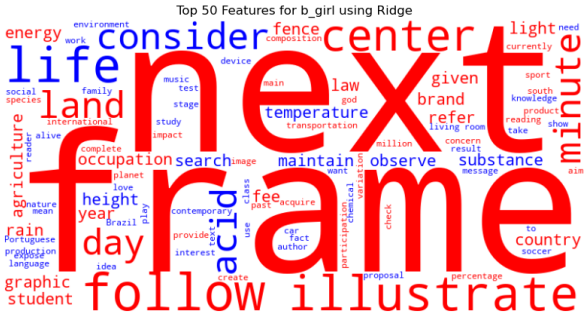


Note: This figure presents the estimated coefficients for each non-contextual measure and IRT parameters when regressing the gender, socioeconomic, and conditional ethnic performance gaps, and controlling for the competence fixed effects outlined in figure 1. "Std." indicates standardized measures, applied to continuous variables. Readability scores are calculated using a composite index based on Carvalho de Lima Moreno et al. (2022), and sentiment scores are estimated using the multilingual BERT model, measuring the probability of a very positive sentiment. IRT parameters are included, and other variables are hand-coded. As a benchmark, the average gender gap is -3.1 percentage points, the SES gap is -6.6 percentage points, and the conditional ethnic gap is -1.6 percentage points, all estimated using equation 1. Confidence intervals are estimated at the 99% level. These factors account for 9.8% of the gender gap variance, 27% of the SES gap variance and 29% of the cond. ethnic gap variance.

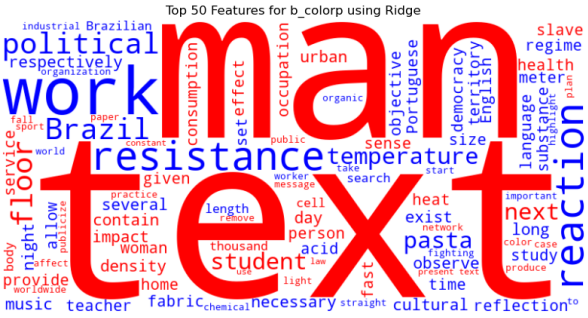
Appendix Figure A10: Performance Gaps and text-based objective measures by groups



(a) SES Gap



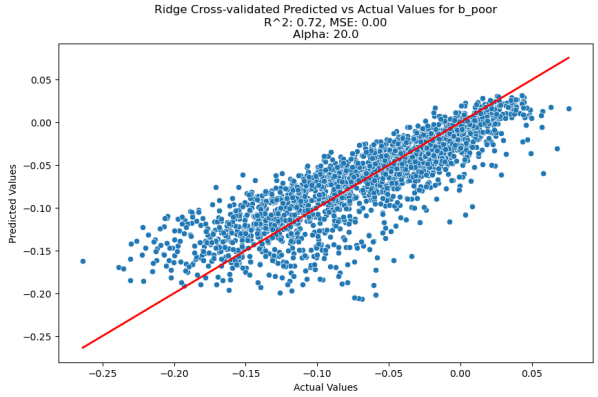
(b) Gender Gap



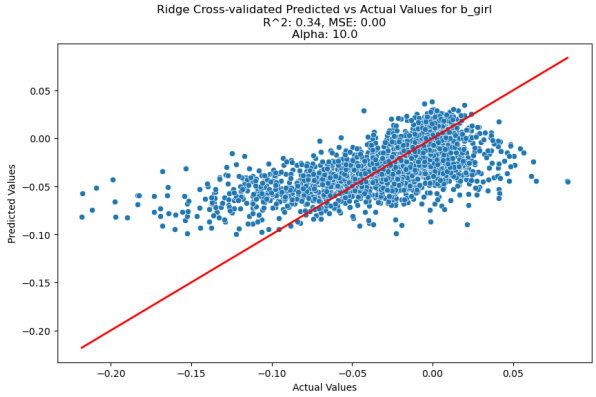
(c) Cond. Ethnic Gap

Note: These panels display the words in the text with the highest predictive power in explaining the observed gaps. Words in blue reduce the gaps, while words in red widen them. To obtain these words, the text is first vectorized into uni-grams and bi-grams. The 2,201 gap values for each question are then analyzed, accounting for competence fixed effects and non-contextual factors, to isolate the variation attributable to context. Panel (a) shows the results for the words that predict the SES gap, Panel (b) shows the words for the gender gap, and Panel (c) shows the words for the conditional ethnicity gap. The words were estimated using a Ridge model, with the alpha parameter selected via grid search, controlling for item features and competence fixed effects. Figure A12 displays the model fits. And Tables A7, A8 and A9 display the size of the estimations word by word.

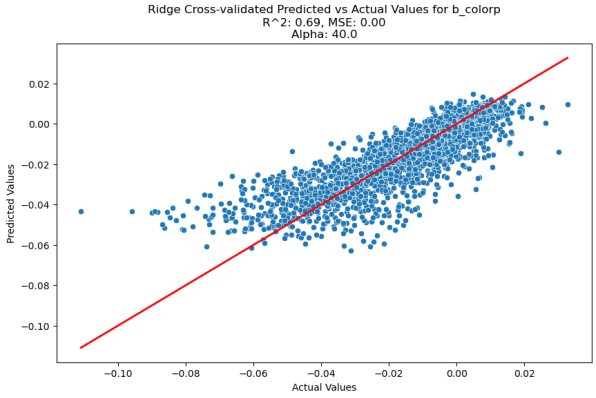
Appendix Figure A11: Words with the highest predicting power by type of performance gap



(a) SES gap



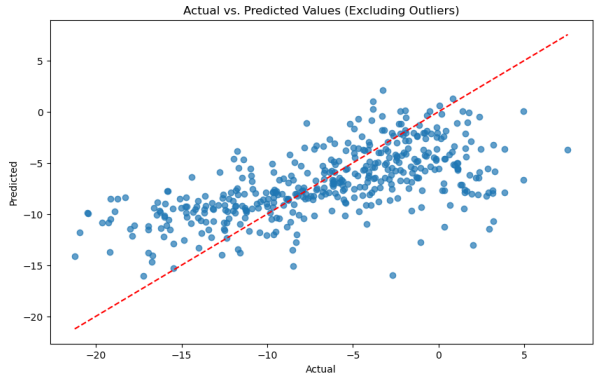
(b) Gender Gap



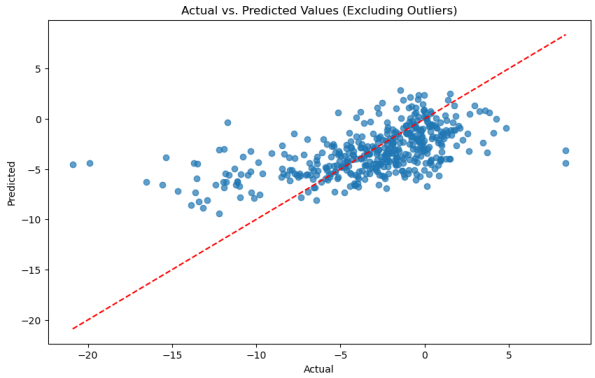
(c) Cond. Ethnic Gap

Note: These panels show the model fitting for each of the estimations based on the words in the RIDGE model, including also competence fixed effects and non-contextual features of each question. The estimations are performed for each of the three gaps for all 2,201 questions. The y-axis plots the predicted values, and the x-axis plots the actual values. The red line depicts the 45-degree line. Additionally, the top of each panel provides information about the α value selected by the grid search and the R^2.

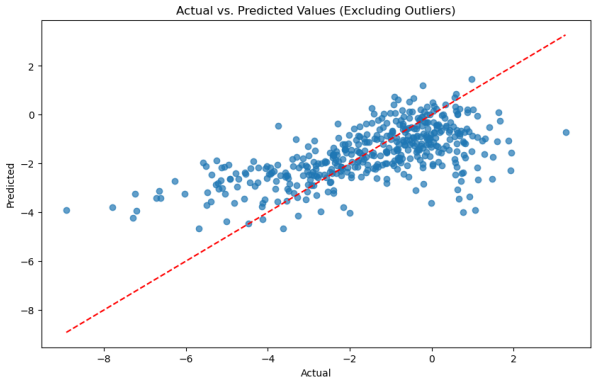
Appendix Figure A12: Fitting in different prediction models-RIDGE uni-grams and bi-grams



(a) SES gap



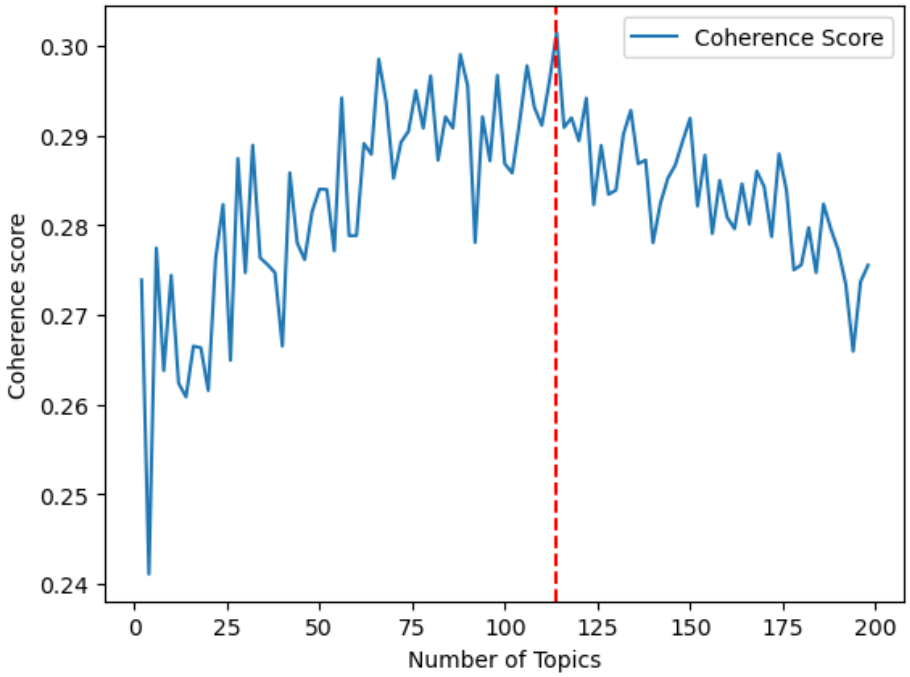
(b) Gender Gap



(c) Cond. Ethnic Gap

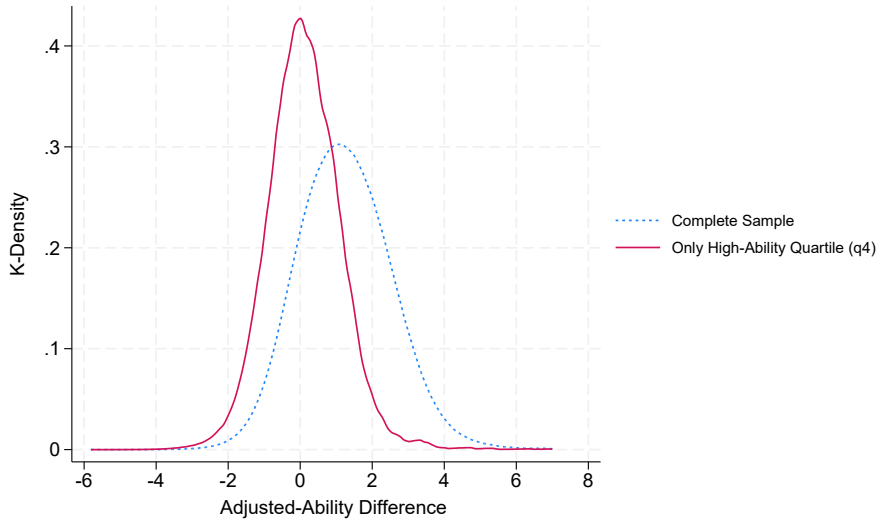
Note: These panels show the model fitting for each of the estimations based on the words in the LASSO model, which also includes competence fixed effects, topics, and non-contextual features of each question. The estimations are performed for each of the three gaps across all 2,201 questions. Topics are generated using the Latent Dirichlet Allocation (LDA) algorithm, in a data-driven manner by maximizing the coherence score. There are 114 topics. The y-axis plots the predicted values, and the x-axis plots the actual values. The red line depicts the 45-degree line.

Appendix Figure A13: Fitting in different prediction models-LASSO with topics



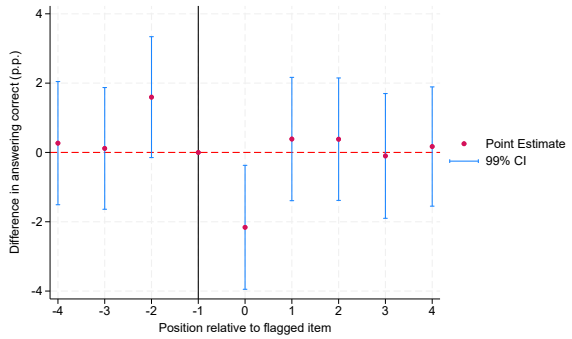
Note: This graph shows the process of maximizing the coherence score—a metric of the degree of semantic similarity between high-scoring words in the topic—on the text corpus. The corpus consists of all the text that appears in any question, and consist on a total of 130,215 words. With each iteration, the algorithm assesses the coherence of the splitting and stops when it is maximized. Topics are distributions of words, and each questions' text is a distribution of topics. Importantly, this clustering process into topics is independent of the gaps, as it only considers the distribution of words for classification.

Appendix Figure A14: Coherence Score by Number of Topics

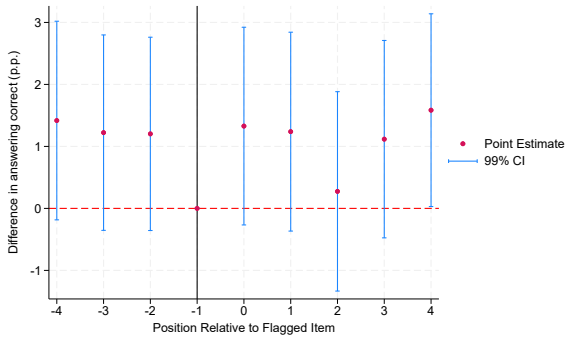


Note: This figure shows the distributions of adjusted ability for the complete sample and for test-takers in the highest performing quartile. Adjusted ability is a measure estimated by subtracting the difficulty parameter B from the estimation of each test-taker's latent ability, θ. Positive values mean that the question is harder than the test-taker's ability. The x-axis is measured in standardized units of ability. A distribution not centered around zero, but around a positive value, indicates that the set of questions is harder than the ability of the test-takers. However, we observe that for the small proportion of test-takers who secure a position in selected public universities, the set of questions is well-calibrated.

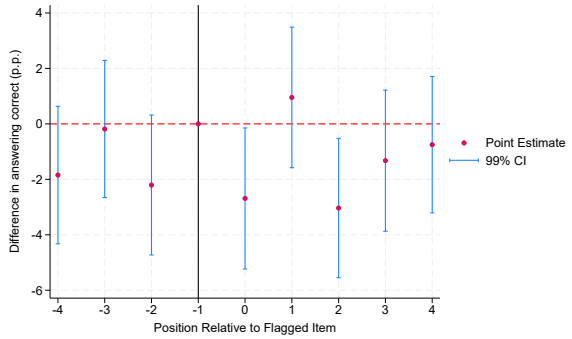
Appendix Figure A15: Adjusted-Ability Distributions



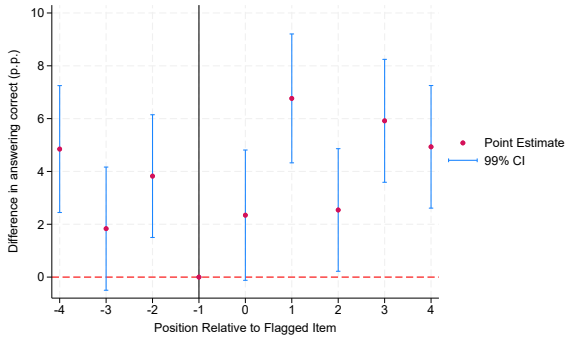
(a) SES relative gap in the presence of a flagged item with a widening effect



(b) SES relative gap in the presence of a flagged item with a reducing effect



(c) Gender relative gap in the presence of a flagged item with a widening effect



(d) Gender relative gap in the presence of a flagged item with a reduce effect

Note: These graphs show the coefficients from executing Equation 3. The time window considered is from -4 items to +4 items, with the tagged question containing one of the testable hypotheses in the middle. The election of this time window is informed by the results of Duquennois (2022). The baseline corresponds to the question immediately preceding the tagged question. The estimation excludes sequences of questions where one of the items belongs to the opposite channel (i.e., a question tagged as reducing when the graph aims to test the widening channel).

Appendix Figure A16: Performance Gaps dynamics by relative position to flagged item

A.3 How the way test are performed affect gaps

Most of the literature has concentrated on interventions that create disparities among groups by altering the settings or environmental conditions in which tests are conducted. In the following, I will summarize this evidence. It is important to note that much of this evidence comes from low-stakes contexts, particularly PISA and TIMMS, due to their accessibility and high-quality data. Additionally, many studies focus on very specific and controlled interventions that examine only one mechanism at a time. There is limited knowledge about the effects in high-stakes situations.

Effects of Gendered Language: Cohen et al. (2023), exploiting a policy change in the Israeli Admission test[31], documents that transitioning from a male-centered language to one that is more neutral benefits women's performance in domains where stereotype threat is present, such as mathematics, reducing the gender gap by 20% with no effects on domains without such threat.

Multiple Choice and wrong penalization: The format in which questions are displayed might also affect the performance across groups. Multiple-choice formatting is widely used because it is easier to mark at large scale. But the specific design opens rooms to other factors that are no active in alternative formatting like open-ended questions. For instance, if an specific group is able to discard some irrelevant choices of is more prone to guess, then performance disparities might arise only due to design factors. Griselda (2022) shows that question that try to measure the same ability, when are displayed as multiple-choice as opposed to open ended questions, they trigger significant gender performance gaps. As indicated in Baldiga (2014) using the SAT find that this effect can be ever strong when there is a penalization for wrong answers. Analogously, Coffman and Klinowski (2020) show that when penalization for wrong responses are removed in the Chilean admission system, the gender gap in test scores widens, particularly on those domains that are necessary to apply to STEM programs.

Content based differences: The content of a test may induce some performance disparities. For instance, females, no matter the type of test or the age in which the test is performed, tend to show and advantage in reading and writing, while males show and advantage in mathematics. But there are others within domain differences. Males tend to perform better in questions asking about rotation, geometry or statistical analyses, while females, tend to be advantaged in algebra and short-answer problems (Reynolds et al., 2022; Miller and Halpern, 2014).

Abstract versus grounded: There is evidence showing that the use of contextual tools that make some question grounded, as oppose to abstract, affect performance because the strategy followed might change (Koedinger et al., 2008; Koedinger and Nathan, 2004). However, the evidence of these effect inducing groups disparities has not been extensible study. One exception is Hickendorff (2013) find no gender difference due to a manipulation in this dimension in primary school students in the Netherlands. However, Van de Weijer-Bergsma and Van der Ven (2021) show that, even though there are no demographic groups difference, grounded questions help the performance of low-ability students through a motivation channel.

Cognitive Endurance and item's order: The level of difficulty of the earlier questions in a test may affect performance in later questions (Anaya et al., 2022) find that ordering the questions from easiest to most difficult yields the lowest probability to abandon the test, as well as the highest number of correct answers. It could be either because tiredness play a role, or because facing a more difficult question early updates the expectation about the overall difficulty on the test, creating anxiety or other responses that impair performance. Reyes (2023) explores the role of cognitive endurance –the ability to sustain effortful mental activity over a continuous stretch of time–

[31] Hebrew, like Portuguese, is a grammatical gender language in which nouns generally have a gender assigned to them, and the noun's gender affects the form of the verb used with it and the form of the pronoun used to refer to it.

on later on outcomes, and being an important generator of disparities[32]. This paper finds a one-standard-deviation higher endurance predicts a 5.4% wage increase. In turn, Brown et al. (2022) exploiting a field experiment, show that cognitive endurance can be enhanced, but only in high-quality schools, suggesting that this may further disadvantage poor children.

Share of same group members: Other environmental condition that might create performance gaps across groups is the observed share of peers performing the test in the same room. Gomez-Ruiz et al. (2024) exploit an exogenous variation in the gender composition in an admission test for a coding program in Uruguay. They find that the absence of male applicants leads to a 0.1 standard deviation increase in women's test scores in mathematics and logical reasoning compared to women in mixed-gender editions, with no effects in the verbal part.

Time constraints: Other dimension related with the environment of the test is the time constraint. De Paola and Gioia (2016) show that on math test, gender gaps emerge when stringent time constraints apply, a common feature in test such as ENEM. Cai et al. (2019) show that these patterns emerge also in real-world high stakes context like the *Gaokao*, exploiting within differences between a mock exam and the real examination. While Galasso and Profeta (2024) using an RCT show that these effects emerge even in low-stakes context.

Answering strategies: There is also evidence highlighting gender differences in test-taking behavior. For instance, Stenlund et al. (2017) found that female test-takers tend to rely on random guessing more than their male counterparts. They also observed that low-achieving test-takers are less likely to follow effective time management strategies during tests. Similarly, Ellis and Ryan (2003) reported that African-American test-takers are more prone to using ineffective test-taking strategies, such as randomly selecting an answer when guessing, reading through the entire test before starting, or reviewing the answer options before reading the questions. Their findings suggest that part of the performance gap can be explained by the use of these ineffective strategies.

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

A.4 Competence definitions and the reference Matrix

The Reference Matrix (accesible at https://download.inep.gov.br/download/enem/matriz_referencia.pdf) defines the cognitive axes that the Ministry is looking at when designing each question. Each question belongs to one of the 30 competencies described for each text. In turn, this competencies can be grouped in 6 to 9 categories per subject. The following figures describe each of the specific competencies by subject.

Area	Competence	Description	Dimension
Language	**I-Apply communication and information technologies in school, work and other contexts relevant to your life**		
	1	Identify the different languages and their expressive resources as elements of characterization of communication systems	Abstract
	2	Use knowledge about the languages of communication and information systems to solve social problems	Grounded
	3	Relate information generated in communication and information systems, considering the social function of these systems	Abstract
	4	Recognize critical positions regarding the social uses made of languages and communication and information systems	Abstract
	II-Know and use modern foreign language(s) as instrument for accessing information and other cultures and social groups		
	5	Associate words and expressions from a Modern Foreign Language (MFL) text with their theme	Dropped
	6	Use MFL knowledge and its mechanisms as a means of expanding possibilities of access to information, technologies and cultures	Dropped
	7	Relate a text in MFL, linguistic structures, its function and social use	Dropped
	8	Recognize the importance of cultural production in MFL as a representation of cultural and linguistic diversity	Dropped
	III-Understand and use body language as relevant for life itself, social integrator and identity-forming		
	9	Recognize bodily manifestations of movement as originating from the daily needs of a social group	Grounded
	10	Recognize the need to transform body habits based on kinesthetic needs	Abstract
	11	Recognize body language as a means of social interaction, considering performance limits and adaptation alternatives for different individuals	Grounded
	IV-Understand art as generating cultural and aesthetic knowledge of meaning and integrator of the organization of the world and of one's own identity		
	12	Recognize different functions of art, the work of artists' production in their cultural environments	Abstract
	13	Analyze the various artistic productions as a means of explaining different cultures, beauty standards and prejudices	Abstract
	14	Recognize the value of artistic diversity and the interrelationships of elements that appear in the manifestations of various social and ethnic groups	Abstract
	V-Analyze, interpret and apply expressive resources of languages, relating texts to their contexts, through the nature, function, organization, structure of demonstrations, according to production conditions and reception		
	15	Establish relationships between the literary text and the moment of its production, situating aspects of the historical, social and political context	Abstract
	16	Relate information about artistic conceptions and literary text construction procedures	Abstract
	17	Recognize the presence of updatable and permanent social and human values in the national literary heritage	Abstract
	VI-Understand and use the symbolic systems of different languages as means of cognitive organization of reality through the constitution of meanings, expression, communication and information		
	18	Identify the elements that contribute to thematic progression and the organization and structuring of texts of different genres and types	Abstract
	19	Analyze the function of the predominant language in texts in specific interlocution situations	Abstract
	20	Recognize the importance of linguistic heritage for the preservation of memory and national identity	Abstract
	VII-Compare opinions and points of view on different languages and their specific manifestations		
	21	Recognize, in texts of different genres, verbal and non-verbal resources used to create and change behaviors and habits	Abstract
	22	Relate, in different texts, opinions, themes, subjects and linguistic resources	Abstract
	23	Infer from a text what its producer's objectives are and who its target audience is, by analyzing the argumentative procedures used	Abstract
	24	Recognize in the text argumentative strategies used to convince the public, such as intimidation, seduction, commotion, blackmail, among others	Abstract
	VIII-Understand and use Portuguese as a language maternal, generator of meaning and integrator of the organization of the world and of one's own identity		
	25	Identify, in texts of different genres, the linguistic marks that distinguish social, regional and register linguistic varieties	Abstract
	26	Relate linguistic varieties to specific situations of social use	Grounded
	27	Recognize the uses of the standard norm of the Portuguese language in different communication situations	Abstract
	IX-Understand the principles, nature, function and impact communication and information technologies in your personal and social life, in development of knowledge		
	28	Recognize the function and social impact of different communication and information technologies	Grounded
	29	Identify communication and information technologies by analyzing their languages	Abstract
	30	Relate communication and information technologies to the development of societies and the knowledge they produce	Grounded

Note: This table shows the translation of each of the competencies and sub-competencies comprising the Language section of the ENEM. The translation from Portuguese was done using DeepL. The "grounded" indicator shows whether the competence aims to measure daily situations as opposed to abstract concepts. Only the competencies in bold (those that group the sub-competencies) are considered in the fixed effects specifications.

Appendix Figure A17: Competencies descriptions for Language

Area	Competence	Description	Dimension
Math	**I-Build meanings for natural numbers, integers, rational and real**		
	1	Recognize, in the social context, different meanings and representations of numbers and operations - natural, integers, rational or real	Grounded
	2	Identify numerical patterns or counting principles	Abstract
	3	Solve problem situations involving numerical knowledge	Grounded
	4	Evaluate the reasonableness of a numerical result when constructing arguments about quantitative statements	Abstract
	5	Evaluate intervention proposals in reality using numerical knowledge	Grounded
	II-Use geometric knowledge to read and the representation of reality and acting on it		
	6	Interpret the location and movement of people/objects in three-dimensional space and their representation in two-dimensional space	Grounded
	7	Identify characteristics of flat or spatial figures	Abstract
	8	Solve problem situations that involve geometric knowledge of space and shape	Grounded
	9	Use geometric knowledge of space and shape in the selection of arguments proposed as solutions to everyday problems	Grounded
	III-Build notions of magnitudes and measurements for understanding reality and solving everyday problems		
	10	Identify relationships between quantities and units of measurement	Grounded
	11	Use the notion of scales when reading representations of everyday situations	Grounded
	12	Solve problem situations involving measurements of quantities	Grounded
	13	Evaluate the result of a measurement in the construction of a consistent argument	Abstract
	14	Evaluate intervention proposals in reality using geometric knowledge related to quantities and measurements	Abstract
	IV-Build notions of variation of quantities for the understanding reality and solving everyday problems		
	15	Identify the dependency relationship between quantities	Abstract
	16	Solve a problem situation involving the variation of quantities, directly or inversely proportional	Grounded
	17	Analyze information involving variation in magnitudes as a resource for building arguments	Abstract
	18	Evaluate intervention proposals in reality involving variation in magnitudes	Abstract
	V-Model and solve problems involving variables socioeconomic or technical-scientific, using algebraic representations		
	19	Identify algebraic representations that express the relationship between quantities	Abstract
	20	Interpret Cartesian graph that represents relationships between quantities	Abstract
	21	Solve problem situations whose modeling involves algebraic knowledge	Grounded
	22	Use algebraic/geometric knowledge as a resource for building arguments	Abstract
	23	Evaluate intervention proposals in reality using algebraic knowledge	Grounded
	VI-Interpret information of a scientific and social nature obtained from reading graphs and tables, forecasting trends, extrapolation, interpolation and interpretation		
	24	Use information expressed in graphs or tables to make inferences	Abstract
	25	Solve problems with data presented in tables or graphs	Grounded
	26	Analyze information expressed in graphs or tables as a resource for constructing arguments	Abstract
	VII-Understand the random and non-deterministic nature of natural and social phenomena and use appropriate instruments for measurements, sample determination and probability calculations to interpret information		
	27	Calculate measures of central tendency or dispersion of a set of data expressed in a table of frequencies of grouped data (not in classes) or in graphs	Abstract
	28	Solve problem situations that involve knowledge of statistics and probability	Grounded
	29	Use knowledge of statistics and probability as a resource for building arguments	Abstract
	30	Evaluate intervention proposals in reality using knowledge of statistics and probability	Abstract

Note: This table shows the translation of each of the competencies and sub-competencies comprising the Math section of the ENEM. The translation from Portuguese was done using DeepL. The "grounded" indicator shows whether the competence aims to measure daily situations as opposed to abstract concepts. Only the competencies in bold (those that group the sub-competencies) are considered in the fixed effects specifications.

Appendix Figure A18: Competencies descriptions for Math

Area	Competence	Description	Dimension
Science	**I-Understand natural sciences and related technologies associated as human constructions, realizing their roles in the processes of production and the economic and social development of humanity**		
	1	Recognize characteristics or properties of wave or oscillatory phenomena, relating them to their uses in different contexts	Abstract
	2	Associate the solution of communication, transport, health or other problems with the corresponding scientific and technological development	Abstract
	3	Confront scientific interpretations with interpretations based on common sense, over time or across different cultures	Abstract
	4	Evaluate proposals for intervention in the environment, considering the quality of human life or measures for the conservation, recovery or sustainable use of biodiversity	Abstract
	II-Identify presence and apply associated technologies to natural sciences in different contexts		
	5	Dimension circuits or electrical devices for everyday use	Abstract
	6	Relate information to understand installation or use manuals for devices or technological systems in common use	Abstract
	7	Select control tests, parameters or criteria for comparing materials and products, with a view to consumer protection, worker health or quality of life	Abstract
	III-Associate interventions that result in degradation or environmental conservation to productive and social processes and instruments or actions scientific-technological		
	8	Identify stages in processes of obtaining, transforming, using or recycling natural, energy or raw materials resources, considering biological, chemical or physical processes involved in them	Abstract
	9	Understand the importance of biogeochemical cycles or energy flow for life, or the action of agents or phenomena that can cause changes in these processes	Abstract
	10	Analyze environmental disturbances, identifying sources, transport and/or destination of pollutants or predicting effects on natural, productive or social systems	Abstract
	11	Recognize benefits, limitations and ethical aspects of biotechnology, considering biological structures and processes involved in biotechnological products	Abstract
	12	Evaluate impacts on natural environments resulting from social or economic activities, considering contradictory interests	Abstract
	IV-Understand interactions between organisms and the environment, in particularly those related to human health, relating knowledge scientific, cultural aspects and individual characteristics		
	13	Recognize life transmission mechanisms, predicting or explaining the manifestation of characteristics of living beings	Abstract
	14	Identify patterns in phenomena and vital processes of organisms, such as maintaining internal balance, defense, relationships with the environment, sexuality, among others	Abstract
	15	Interpret models and experiments to explain biological phenomena or processes at any level of organization of biological systems	Abstract
	16	Understand the role of evolution in the production of patterns, biological processes or in the taxonomic organization of living beings	Abstract
	V-Understand scientific methods and procedures natural resources and apply them in different contexts		
	17	Relate information presented in different forms of language and representation used in physical, chemical or biological sciences, such as discursive text, graphs, tables, mathematical relationships or symbolic language	Abstract
	18	Relate physical, chemical or biological properties of products, systems or technological procedures to the purposes for which they are intended	Abstract
	19	Evaluate methods, processes or procedures from natural sciences that contribute to diagnosing or solving social, economic or environmental problems	Abstract
	VI-Appropriate knowledge of physics to, in problem situations, interpret, evaluate or plan scientific and technological interventions		
	20	Characterize causes or effects of the movements of particles, substances, objects or celestial bodies	Abstract
	21	Use physical and (or) chemical laws to interpret natural or technological processes within the context of thermodynamics and (or) electromagnetism	Abstract
	22	Understand phenomena arising from the interaction between radiation and matter in their manifestations in natural or technological processes, or in their biological, social, economic or environmental implications	Abstract
	23	Evaluate possibilities for generating, using or transforming energy in specific environments, considering ethical, environmental, social and/or economic implications	Abstract
	VII-Appropriate knowledge of chemistry to, in problem situations, interpret, evaluate or plan scientific and technological interventions		
	24	Use chemistry codes and nomenclature to characterize materials, substances or chemical transformations	Abstract
	25	Characterize materials or substances, identifying stages, yields or biological, social, economic or environmental implications of their obtaining or production	Abstract
	26	Evaluate social, environmental and/or economic implications in the production or consumption of energy or mineral resources, identifying chemical or energy transformations involved in these processes	Abstract
	27	Evaluate proposals for intervention in the environment applying chemical knowledge, observing risks or benefits	Abstract
	VIII-Appropriate knowledge of biology to, in problem situations, interpret, evaluate or plan scientific and technological interventions		
	28	Associate adaptive characteristics of organisms with their way of life or their distribution limits in different environments, especially in Brazilian environments	Abstract
	29	Interpret experiments or techniques that use living beings, analyzing implications for the environment, health, food production, raw materials or industrial products	Abstract
	30	Evaluate proposals of individual or collective scope, identifying those that aim to preserve and implement individual, collective or environmental health	Abstract

Note: This table shows the translation of each of the competencies and sub-competencies comprising the Natural Sciences section of the ENEM. The translation from Portuguese was done using DeepL. The "grounded" indicator shows whether the competence aims to measure daily situations as opposed to abstract concepts. Only the competencies in bold (those that group the sub-competencies) are considered in the fixed effects specifications.

Appendix Figure A19: Competencies descriptions for Science

Area	Competence	Description	Dimension
Social Sciences	**I-Understand the cultural elements that make up identities**		
	1	Interpret historically and/or geographically documentary sources about aspects of culture	Abstract
	2	Analyze the production of memory by human societies	Abstract
	3	Associate current cultural manifestations with their historical processes	Abstract
	4	Compare points of view expressed in different sources on a certain aspect of culture	Abstract
	5	Identify the manifestations or representations of the diversity of cultural and artistic heritage in different societies	Abstract
	II-Understanding the transformations of geographic spaces as a product of socioeconomic and cultural power relations		
	6	Interpret different graphic and cartographic representations of geographic spaces	Abstract
	7	Identify the historical-geographical meanings of power relations between nations	Abstract
	8	Analyze the action of national states with regard to the dynamics of population flows and in facing economic-social problems	Abstract
	9	Compare the historical-geographical significance of political and socioeconomic organizations on a local, regional or global scale	Abstract
	10	Recognize the dynamics of the organization of social movements and the importance of community participation in the transformation of historical-geographic reality	Abstract
	III-Understand the production and historical role of social, political and economic institutions, associating them with different groups, conflicts and social movements		
	11	Identify records of social group practices in time and space	Abstract
	12	Analyze the role of justice as an institution in the organization of societies	Abstract
	13	Analyze the actions of social movements that contributed to changes or ruptures in processes of dispute for power	Abstract
	14	Compare different points of view, present in analytical and interpretative texts, on situations or facts of a historical-geographical nature regarding social, political and economic institutions	Abstract
	15	Critically evaluate cultural, social, political, economic or environmental conflicts throughout history	Abstract
	IV-Understand technical and technological transformations and their impact on production processes, knowledge development and social life		
	16	Identify records about the role of techniques and technologies in the organization of work and/or social life	Grounded
	17	Analyze factors that explain the impact of new technologies on the process of territorialization of production	Abstract
	18	Analyze different processes of production or circulation of wealth and their socio-spatial implications	Abstract
	19	Recognize the technical and technological transformations that determine the various forms of use and appropriation of rural and urban spaces	Abstract
	20	Select arguments for or against the changes imposed by new technologies on social life and the world of work	Abstract
	V-Use historical knowledge to understand and value the foundations of citizenship and democracy, favoring a consciousness of the individual in society		
	21	Identify the role of the media in the construction of social life	Abstract
	22	Analyze the social struggles and achievements obtained with regard to changes in legislation or public policies	Abstract
	23	Analyze the importance of ethical values in the political structuring of societies	Abstract
	24	Relate citizenship and democracy in the organization of societies	Abstract
	25	Identify strategies that promote forms of social inclusion	Abstract
	VI-Understanding society and nature, recognizing their interactions in space in different historical and geographic contexts		
	26	Identify in different sources the process of occupation of physical environments and the relationships between human life and the landscape	Abstract
	27	Critically analyze society's interactions with the physical environment, taking into account historical and/or geographic aspects	Abstract
	28	Relate the use of technologies with socio-environmental impacts in different historical-geographical contexts	Abstract
	29	Recognize the role of natural resources in the production of geographic space, relating them to the changes caused by human actions	Abstract
	30	Evaluate the relationships between preservation and degradation of life on the planet at different scales	Abstract

Note: This table shows the translation of each of the competencies and sub-competencies comprising the Social Sciences section of the ENEM. The translation from Portuguese was done using DeepL. The "grounded" indicator shows whether the competence aims to measure daily situations as opposed to abstract concepts. Only the competencies in bold (those that group the sub-competencies) are considered in the fixed effects specifications.

Appendix Figure A20: Competencies descriptions for Social Sciences

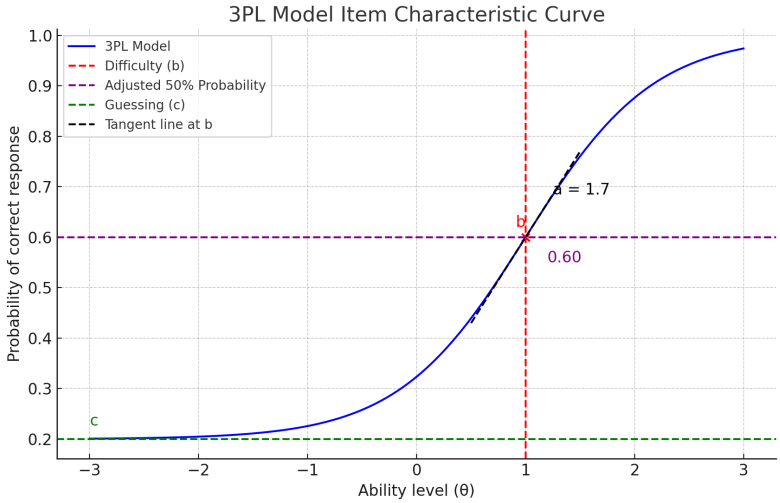
A.5 IRT and the ENEM scoring

The ENEM test relies on a item bank to select the question each year compose the test. This bank has been made with the help of experts, combining high school teachers and college faculties. The idea is to have a broad set of question aiming to assess the different set of skills necessary for the college continuation.

Each individual item is tested in pre-test session, where a representative sample of population face these items and where their performance is assessed. Based on the performance of this training sample, the parameters of each items are calibrated. The model used to calibrate is the so called 3PL model.

$$\Pr(y_{ij} = 1 \mid \theta_j) = c_i + (1 - c_i)\frac{\exp\{a_i(\theta_j - b_i)\}}{1 + \exp\{a_i(\theta_j - b_i)\}} \quad \theta_j \sim N(0, 1) \tag{4}$$

Based on this model, the item probability of answering correct can be characterized by the use of three parameters; a, the discrimination parameter which approximates the effectiveness of the item to differentiate between examinees with a high ability level and a low ability level; b, the difficulty parameter, which indicates the ability point in which the examinee has 50 percent probability of responding correctly (without considering guessing) ; and c, the guessing parameter indicates the likelihood that a student with an infinitely negative ability has to correctly respond to the question. Figure A21 shows the so called characteristic curve for some general item.



Note: This figure describes the elements of the "Characteristic Curve" in the 3PL IRT model. The guessing parameter C represents the probability that a test-taker with extremely low ability (approaching negative infinity) can still guess the correct answer. The discrimination parameter A reflects how sharply the item distinguishes between test-takers of different abilities, corresponding to the slope of the curve at the difficulty parameter. A higher value indicates better discrimination. Finally, the difficulty parameter B represents the ability level at which a test-taker has a 50% probability of answering the item correctly, assuming no guessing.

Appendix Figure A21: 3PLM Characteristic Curve

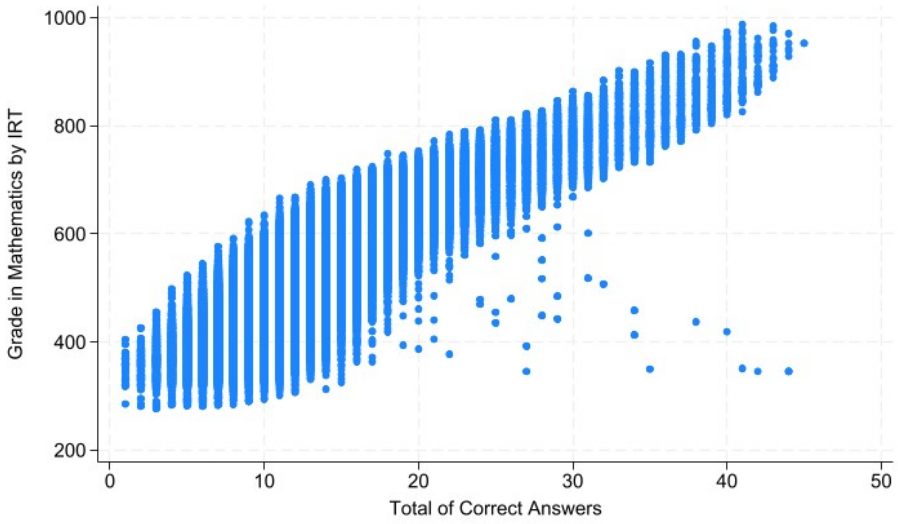
The intuition behind the Item Response Theory (IRT) is that a sequence of items can provide information about the latent variable representing ability if some assumptions hold. Individuals with low ability should only be able to respond easier questions. If they actually respond correctly a series of harder items, then, the model assumes that this is due to chance. And therefore they do not imput complete weight to those items. So based on the complete string of empirical responses, by using Maximum Likelihood, the θ_j parameter is pin down. It can be note in Figure A22 that this produces a high variability of scores, conditional on having the same number of raw

correct answers (particularly in the neighborhood 10 to 20 correct answers) .

Note that these parameters have been calibrated using the whole pilot sample, so they do not consider important factor such as the order of the item (for instance, Reyes (2023) documents important differences in performances due to different endurance abiliy) nor the nuances that are associated with the contextual cuing effect studied in this paper. Because of this, there are some critics of the use of this model in grading the scores of the ENEM.

The information provided by the IRT scores might be helpful to control when estimating the gaps per question, as provide one way of include a proxy of ability. The latent ability estimation (parameter θ) is the result of the most consistent pattern across different set of question, calibrated in a sample that is not the one that I am observing. The intuition if that a test-taker with low ability does well in a set of hard items, then, it is likely due to chance, and therefore the model discounts the score based on this inference. This is the reason why the profiles of grades versus the raw total number of correct items looks as depicted in Figure A22, where it can be seen a huge dispersion in terms of the score granted, even conditional on the same number of overall raw correct responses.

And as these features might also affect the performance on *an specific* question, it can not be the case that the theta estimation is totally biased, since most of the questions do not suffer from this biasing structure. That is why I consider that this theta is a good proxy of ability.



Note: This figure illustrates the differences between grading schemes when using IRT compared to traditional grading based on the number of questions answered correctly. One of the advantages of IRT is that it increases the number of possible scores, thereby enhancing the granularity of the metric. For instance, in 2012, with 45 questions available in mathematics, the total possible scores in traditional marking is 46, whereas when using IRT, this number increases to 6,406. However, this comes with a cost, as the dispersion for the same number of correct responses is high, allowing disparities to emerge if there are contextual differential responses by groups.

Appendix Figure A22: IRT v.s Actual Number of Corrects-Mathematics Test