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

Cristóbal Ruiz-Tagle

August, 2024.
Preliminary and Incomplete.
Please do not cite or circulate.

Abstract

Standardized tests are widely used to screen candidates, operating under the assumption that they accurately assess innate abilities, often with high-stakes consequences. While significant research has explored how testing procedures can induce performance gaps across groups, little attention has been given to the role of textual content in test items as a potential trigger for these disparities. This paper examines how performance gaps related to socioeconomic status (SES), gender, and ethnicity are influenced by the contextual features of items in the Brazilian admission test, ENEM. The analysis is conducted in two steps. First, I investigate the aggregate performance gaps observed for each group over a 13-year period and over 3.8 million test-takers, analyzing how these gaps relate to the textual content of each item. Using text-analysis techniques, regularization regressions, and assistance from ChatGPT, I develop six testable hypotheses—two for each group: one predicting a widening of gaps and the other predicting a reduction, based solely on the text content. Interestingly, although these hypotheses were generated through a data-driven approach, they align with the well-established Stereotype Threat theory. In the second part, I test these hypotheses using the within-individual data structure. The findings provide strong evidence that the SES gap widens when items include financial concepts, with this effect being particularly pronounced among top performers. For gender gaps, the widening effect is driven by items featuring abstract scientific concepts and measurements, though this negative impact can be mitigated when the item also includes a female character. The evidence for gap reduction is weaker in both cases, and no text-induced biases are found for ethnic disparities. These results offer valuable insights for test design, providing a data-driven approach that can be extended to other contexts.

Keywords: standardized test; item design; text analysis; performance gaps
JEL Codes: I23; I24; I31; D63.

Acknowledgments: I am deeply thankful to Pamela Giustinelli and Sarah Eichmeyer for their invaluable mentorship and guidance throughout this project. I also acknowledge the financial support provided by ANID-Chile. As always, the usual disclaimers apply.
Affiliations: Bocconi University. Mail to cristobal.ruiztagle@unibocconi.it.
First version: August, 2024.

1 Introduction

Standardized test are widely used to estimate the ability of college applicants, to monitor schools achievement and to screen the most suitable candidate for a job positions.[1] In most cases, standardized test are at high-stakes, being the open door –to some– to better opportunities. Despite of their extended use, increasingly, groups oppose the use of these systems, considering that they are not able to screen underlying ability and that they only replicate the inequalities observed in education systems.

While there is an extensive literature in economics that studies to what extent the way in which a standardized test *is applied*—such as whether it is open-ended, multiple-choice, the length and order of items, the type of competencies measured, the level of abstraction, or whether the paper or virtual format generates differences between groups—triggers group disparities, little has been studied about the effect that *the way the question is asked* has in generating these disparities.[2]. This paper aims to contribute to closing the gap in understanding how the contextual text used to frame a question influences performance disparities across socioeconomic, gender, and ethnic groups. By "contextual text," I refer to the elements that provide context or set the stage for a question, such as hypothetical scenarios, examples, descriptions rooted in everyday experiences, names, places, and similar contextual tools.

The setting for this research is the Brazilian college admission process, specifically focusing on the ENEM standardized test. The ENEM is the second-largest admission test in the world, after the Chinese *Gaokao*. Each year, millions of Brazilians take this test, which comprises 180 questions divided equally among four subjects: mathematics, language, social science, and natural sciences. The test is administered on paper over two days. The items are multiple-choice, with five possible alternatives, and follow a standardized format regardless of the subject, consisting of three elements: a heading, an assignment, and the choices. Each item is independent and self-contained. All test-takers in a given year face the same set of items, but the order varies based on the random assignment of one of the eight available booklets. The ENEM scores are the input through which candidates apply to free public universities using the centralized clearinghouse system available. Therefore, the test is a high-stakes context.

The Brazilian context provides an unique setting to study this research question. Inequalities

[1] Currently, several countries adopt this practice, such as 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 internatonal graduates students by the Graduate Record Examination (GRE) or to asses foreing language proeficiency (TOEFL or IELTS). Similarly, standardized assessments are used to license professionals in fields such as medicine and law. Some examples are the bar exams in law, the United States Medical Licensing Examination (USMLE), or the Medico Interno Residente (MIR) in Spain. In addition, standardized tests are used to compare educational systems across countries. Among the most prominent examples is the National Assessment of Educational Progress (NAEP) in the United States, while the Program for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS). Finally, standardized tests also are commonly used for hiring and screening candidates in competitive labor markets (Rudner, 1992; Schmidt and Hunter, 1998).

[2] There is another strand of the literature related to this paper that shows how seemingly minor and random factors—unrelated to academic ability—can affect standardized test performance, such as temperature (Park, 2022), pollen levels (Bensnes, 2016) or pollutants (Ebenstein et al., 2016).

in Brazilian education are pervasive, with specific underprivileged groups showing lower rates of enrollment and great performance gap in the admission test. Secondly, the stakes that are at play when taking the test are high due to the high returns higher education provides (Duryea et al., 2023; Binelli and Menezes-Filho, 2019). This is driven mainly by the fact that the most prestige university do not charge fees, affirmative action policies have been active since early 2000 and because application is centralized and based on preferences and ENEM results, giving no room to discriminatory practices at the enrollment/selection stage.[3] On top of this, data quality and availability are comparatively good. All information is publicly available, which is useful for replication, with rich individual-level information and the performance on each question faced, including the text and the order of the item in the booklet for randomly assigned booklets. These characteristics strengthen the identification strategy and allow for the use of individual fixed effects, order-of-the-item fixed effects, subject fixed effects, or the ability that is being measured.

In this paper, I adopt a deductive approach consisting of two steps. First, I explore general associations between item-specific groups and their contextual features. Using data spanning 13 years from 2010 to 2022 and a population of 3.8 million test-takers, I collect all the items included in the regular booklets of the ENEM test. I identify 2,201 items that are informative in terms of their behavior (i.e., a reasonable share of test-takers answer correctly, and the distribution of responses converges into a well-behaved characteristic curve). Leveraging text analysis techniques, regularization regressions, and the assistance of ChatGPT, I uncover six hypotheses based on the analysis results—two for each gap (SES, gender, and ethnicity)—with one hypothesis predicting a widening effect and the other a reduction effect. I refer to these as the *widening* channel and the *reduction* channel. For the SES gap, the widening channel is associated with items that include financial contexts, while the reduction channel is linked to practical problem-solving scenarios rooted in real-life situations. For the gender gap, the widening channel is tied to items involving abstract scientific concepts, shapes, and measurements, while the reduction channel involves practical problem-solving in the context of narratives, social interactions, creative expression, or emotions. Finally, for the ethnicity gap, the widening effect is connected to abstract social phenomena and terminologies, while the reduction channel focuses on historical, political, and cultural contexts.

The approach I follow here is similar to those used in other settings (Batista and Ross, 2024; Wang et al., 2023). The use of large language models combined with machine learning techniques has recently been highlighted as a way to expand the set of testable hypotheses, generate novel theories, and reduce researcher-driven biases (Ludwig and Mullainathan, 2024; Charness et al.,

[3]Prior to 2010, each higher education institution had its own admission exams to select students, who, in turn, could apply to multiple institutions by taking their institution-specific admission exams. Some students received multiple offers, while many others did not get any offer. In 2010, the Ministry of Education created the SISU system, a centralized clearinghouse that allocates students to tuition-free and prestigious public higher education institutions. Using scores from a nationwide exam called ENEM, candidates can submit up to two program choices among those available in the system, and a deferred acceptance algorithm is used to assign students to seats. Not surprisingly, the clearinghouse rapidly expanded, becoming the main instrument for college admission in the country. Between 2010 and 2017, the fraction of programs from public institutions using the SISU platform to select their students increased from 20 to 75 percent Machado and Szerman (2021).

2023). In this paper, I advance this methodology by adapting it to the educational field, introducing a novel approach that is not only inspired by cutting-edge research but also easily replicable with promising potential for further applications.

In the second step, I leverage the panel structure of the data and the random allocation of test booklets to individuals to test how items, classified according to the generated hypotheses, affect within-individual performance. To manage computational constraints, I randomly sample 10,000 individuals each year. The analysis finds supporting evidence for four out of the six hypotheses, with the ethnicity-related hypotheses not supported by the data. For the SES gap, the *widening* channel results in a gap increase of 1.4 percentage points (23% relative to the SES gap). This results is very close to the result in Duquennois (2022) where she finds that on monetary themed items, low-SES test-takers are 1.2 percentage points less likely to answer the flagged item correctly. The *reduction* channel is associated with a 0.8 percentage point effect, particularly among lower-ability test-takers. For gender gaps, the *widening* effect is concentrated in the top half of the ability distribution, with an effect size of 1.06 percentage points (30% of the average gender gap). The *reduction* channel has a consistent effect across quartiles, reducing the gap by 0.8 percentage points. All statistically significant effects are impacting locally, with no spillover to subsequent questions. This finding contrasts with the results of Duquennois (2022), highlighting how differently text-induced distortions can manifest in a high-stakes context.

Heterogeneity analyses reveal several important insights. First, the *widening* effects are concentrated among top performers, which aligns with stereotype threat theory, as these individuals face the greatest pressures and expectations. In contrast, the reduction effects are relatively consistent across ability quartiles. Second, in the case of gender gaps where they are easy to identify, items associated with stereotyped competences exacerbate the *widening* effect, except among females in the top quartile of ability. There is no heterogeneous effect on the *reduction* channel based on whether an item is stereotyped. Third, for gender gaps, a positive mirroring effect is observed: when female test-takers encounter a female character in the item, the *widening* effect is completely offset. Consistent with stereotype threat theory, the *reduction* channel remains unaffected. However, for SES gaps, the mirroring effect actually amplifies the widening channel. This suggests that items featuring concepts linked to gender-related widening effects should be paired with female characters to mitigate the impact.

The results can be interpreted through the well-established Stereotype Threat theory (Spencer et al., 2016). This theory suggests that test-takers who are aware of belonging to a group that is stereotypically expected to perform poorly in certain areas (such as women in mathematics or low-SES individuals in various domains) may underperform due to the pressure of these expectations. The primary mechanism behind this theory is the cueing effect, which impacts working memory as stereotype-related concerns interfere with attentional resources (Beilock et al., 2007; Schmader and Johns, 2003).

This paper contributes to three strands of the literature. First, it adds to the growing body of

research on the effectiveness of standardized assessments in measuring competencies across different groups and on identifying design elements that trigger performance disparities among relevant groups (Duquennois, 2022; Griselda, 2022; Brown et al., 2022; Cai et al., 2019; Cohen et al., 2023; Baldiga, 2014). The key contribution of this paper is providing relevant evidence from a large-scale, high-stakes setting, demonstrating that the framing of questions, particularly the textual resources used to facilitate comprehension, can have significant consequences for specific groups. Policymakers and test designers should consider these differential impacts when deciding on the content and presentation of examination questions. Methodologically, the paper highlights the richness of insights that can be gained from analyzing the text included in test booklets. This layer of analysis has been largely neglected in studies of test features affecting performance disparities, likely due to the lack of available data.[4] The approach proven effective in other settings (Batista and Ross, 2024; Ludwig and Mullainathan, 2024) is easily extendable to the education domain, with promising potential for future research.

Secondly, this paper contributes to the literature on how priming certain situations can affect productivity in relevant groups, particularly in financially themed contexts (Vohs et al., 2006; Duquennois, 2022; Muskens et al., 2024) and under gendered stereotype threat (Schmader et al., 2015). The findings demonstrate that these impacts extend to real-world situations with significant consequences for disadvantaged groups. Furthermore, the paper reveals that the distortive effects of such items are not confined to mathematical domains but can emerge in any domain, affecting multiple relevant groups. A key contribution of this paper is its exploration of how these induced disparities vary by achievement level, showing that they are particularly pronounced in highly stereotyped contexts and most significantly impact high-ability test-takers. Additionally, the paper provides evidence that some of these negative effects can be mitigated through a mirroring effect, as seen when an item associated with gender-related widening effects also includes a female character. In doing so, this research introduces a novel intervention to the policymaker's toolkit—targeting a behavioral constraint rather than a financial or informational one—that can effectively reduce gender and social aspiration gaps in a manner that is low-cost, easy to implement, and scalable.

Finally, this paper contributes to the literature on education systems by examining how testing, as the gateway to improved opportunities, can perpetuate inequalities, particularly in high-stakes contexts (Zwick, 2019; Atkinson and Geiser, 2009; Galla et al., 2019; Rothstein, 2004). In the Brazilian admission system, the findings suggest that simple measures can significantly enhance fairness in college access. By reducing the share of items that exploit the factors studied in this paper and replacing them with equivalent items in terms of difficulty and discriminatory power, access gaps to high-quality, free college education can be substantially narrowed. This has profound implications for social equity in Brazil. For instance, Landaud et al. (2024) find that students in Norway who are fortunate enough to be randomized into exam subjects aligned with their academic strengths have higher high school graduation rates and earn higher wages later in life. This is also

[4]Somewhat related is the work of Adukia et al. (2021).

related to literature showing how environmental variation during testing can generate long-lasting disparities (Gomez-Ruiz et al., 2024; Ebenstein et al., 2016; Bensnes, 2016; Park, 2022).

The paper is structured as follows. Section 2 provides an overview of the ENEM admission test and the Brazilian context, discussing why both theoretical and empirical evidence do not offer a clear prediction of the effect of textual features on performance. Section 3 describes the dataset. Section 4 presents the results of analyzing the mapping between item-specific gaps and their textual features and outlines the process for generating hypotheses from the data. Section 5 tests the hypotheses by leveraging the panel data structure to identify within-individual effects and presents some heterogeneity analyses. Finally, Section 6 concludes.

2 Background

Brazil, home to 203 million people with a real GDP per capita of USD 8,802 in 2022, is a large federal country with 26 states and over 5,500 municipalities. Despite its diversity, systemic racial and gender discrimination limit opportunities for many, perpetuating inter-generational poverty. Afro-Brazilians and Indigenous Peoples have less access to quality education and healthcare compared to whites, while women face significant job discrimination, limiting their earning potential. Rural areas suffer from pervasive inequalities in accessing public services. 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 underscores Brazil's status as one of the most unequal countries globally (Bank, 2024).

2.1 Brazilian ENEM: a high-stakes setting

ENEM is a test available since 1998. It used to be an exam to grant the high school diploma, but in 2009 it also became the examination for accessing college.[5] Figure A1 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) backgrounds.

[5]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.

The 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 A2 shows the share of senior high school students who participate each year in relation to the total number of test-takers.

The ENEM covers four subjects: language, mathematics, social sciences, and natural sciences, plus a handwritten short essay on a proposed topic. 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.[6] 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.[7] 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.

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, requiring about 20% of the allotted time to read. 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.[8]

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

[6]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 F describes the grading model.

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

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

ticipate 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.

High-stakes contexts themselves can be a source of disparities due to differences in managing stress, exerting effort, and changing preparation. For instance, Cai et al. (2019) show that varying the stakes in the *Gaokao* by comparing the same individual's performance in a mock and actual examination leads to significant gender gaps. Similarly, Attali et al. (2018), comparing performance in voluntary GRE preparation tests with the actual GRE exam, show that performance gaps among gender and racial groups are larger in the high-stakes context and provide evidence that this is due to differences in the effort exerted. In turn, Reyes et al. (2023) study how variations on the stakes in the ENEM test in Brazil by measuring the staggered participation of some public universities in some later states, affected performance gaps across groups. Finally, Ofek-Shanny (2024), in line with this evidence, highlights that performance gaps are sensitive to the stakes. This point is crucial because most of the evidence we have regarding how test features (see Section C) affect performance comes from low-stakes contexts such as international tests like PISA or TIMSS, primarily because of their great availability and fine-grained data. Here is where the ENEM context can provide insights into how these features affect performance in a high-stakes context.

2.2 Theoretical Background

There is a significant body of literature examining which elements of test design can trigger performance gaps. Much of the evidence focuses on environmental factors or design elements that influence performance. Additionally, most of this evidence comes from low-stakes settings such as TIMMS, PISA, or national standardized tests aimed at categorizing schools nationwide, with particular emphasis on the role of quantitative/mathematics items. Section C discusses some of the key mechanisms studied in the literature.

Despite this extensive research, less is known about the role of contextual mental cues in affecting

performance. A theoretical framework that helps to characterize the phenomena studied in this paper is the well-known Stereotype Threat theory, initially proposed by Steele and Aronson (1995). Stereotype threat theory predicts that members of negatively stereotyped groups will underperform when the relevant stereotype is made salient or relevant to the task at hand.[9] The underlying mechanism is that when a stereotype about a group to which one belongs becomes prominent, concerns about being judged according to that stereotype arise, which can inhibit performance. There are prevalent stereotypes about women's performance in math-related domains, as well as stereotypes about ethnic and low-income groups' performance across all domains of cognitive tests (Charness and Chen, 2020).[10] This phenomenon is not limited to cognitive domains, as it has also been documented in sports (Beilock and McConnell, 2004) and entrepreneurship (Zhang et al., 2023).

Typically, this literature examines random or quasi-random variations in the salience of group identity, either by asking for gender, income, or ethnicity at the beginning of the test in a pre-test questionnaire, by making explicit statements about the goal of the examinations, or by including explicit statements associated with the stereotype (e.g., "women are poorer at math than men"), or by altering the order of different subjects within the test—placing the domain more exposed to the threat first—and then observing whether there are spillover effects due to this manipulation.[11] The outcome of interest is usually the overall test performance. Meta-analyses suggest that this is a robust effect, though some moderators, such as test difficulty and when negatively stereotyped test-takers hold a minority status, have been identified. It is noteworthy, however, that in high-stakes contexts, the effect may range from negligible to small Shewach et al. (2019).[12]

Interestingly, stereotype salience does not always impair performance. For instance, Levy (1996) shows that memory performance among elders can be improved when positive stereotypes are primed, and Shih et al. (1999) found that Asian American women performed better on a mathematics test when their Asian identity was cued, as there is a positive stereotype of Asians being better at math. The converse happened when Shih et al. (2002) flipped the analysis and studied the effect on verbal tests. This evidence suggests that individuals are not only "threatened" by stereotypes but are more generally "susceptible" to stereotypes (Shih et al., 2002).

Duquennois (2022), building on this effect, studies the impact of varying the content of test items on item-by-item differential performance. She finds, using the mathematics module of the TIMSS, that test-takers from the lowest 50% in terms of parental education perform worse when

[9] The term "stereotype" has Greek origins, meaning "stereo"—rigid and "typos"—impression.

[10] For a general discussion of the mechanisms behind this effect and possible policies to address it, see Schmader et al. (2015). In a nutshell, the most likely explanation points to a reduction in working memory because the added stereotype-related concerns interfere with attentional resources (Beilock et al., 2007; Schmader and Johns, 2003).

[11] Although these manipulations are common in the literature, the causal channel is not clear, as one would expect the effect to be diluted over time, especially in a long-duration test. However, this is difficult to test when only overall performance is observed, rather than question-by-question variation.

[12] Cullen et al. (2006) argue that high motivation in high-stakes contexts might help individuals overcome the effects of stereotype threat. Additionally, it is plausible that pressures in high-stakes tests are already so intense that relatively minor stereotype threat manipulations result in no differences in performance across experimental conditions.

they encounter items with a monetary theme. Moreover, this effect is not only local but also spills over to the subsequent four to five items.[13] Drawing on Manning et al. (2024), she rationalizes these results through the concept of "attention capture," triggered by the context depicted in the flagged questions. Reminders of scarcity crowd out attention, impairing performance temporarily. This attention loss persists for several items before eventually diminishing. To the best of my knowledge, there is no existing evidence that examines the effect of varying threats within items on gender or ethnicity.

Despite this effect being well-known, there is significant scrutiny regarding the external validity of these findings. Most of the evidence comes from underpowered studies or from low-stakes contexts where lack of motivation or engagement might be a significant confounding factor (Priest et al., 2024; Shewach et al., 2019; Flore et al., 2018).

Theoretical perspectives on the ex-ante effect of different priming on within-individual performance remain unclear. Some evidence from underpowered samples, using manipulations that are applied once, before starting the test, suggests that the effects are relevant. However, this evidence comes from general performance data rather than item-by-item variation. By utilizing real examination data and examining the effect of item-by-item contextual features, I address this gap in the literature and alleviate concerns of experimenter demand effects and sensitivity to specifically designed wording of priming statements that may not be reflective of typical examination conditions. Confirming this evidence is important for building upon this theory, particularly in real-world and high-stakes applications (Levitt and List, 2007).

3 Data

My data spans 13 years, from 2010 to 2022.[14] It is composed by 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, racial 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

[13] This is particularly interesting because Walker and Bridgeman (2008) found no spillover effect on the overall gender gap in SAT scores when randomly changing the order of the different subject tests.

[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 A1, 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.

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 A18 shows a general example).[17] Finally, the third data set is based on scrapped text from the booklets [18], from where I distinguish three elements: the heading, the assignment and the alternatives. I also labeled items that included figures or graphs and tables.

In addition to this information, each year test-takers are asked about year-specific issues in companion questionnaires. When the information available there is useful, I take advantage of them, for instance by identifying who took preparatory courses, those that felt unprepared to perform the test (but at the end they submitted it completely) or those who report having felt discriminated against for different reasons (economic, gender, racial, sexual preference, religious, etc.).

Note that this sample is a highly selected sample that is justified under the goal of measuring the role of context cues in high-stakes context. Students that want to pursue a college degree in a high-quality public college in Brazil, that in top of that showed up at the test and submitted the whole package might not be representative of their group, and possible associated with more resilient samples of their group. Also note, that I have no access to any ability measure from previous observation, such as performance in other standardized test or high school grades, that would allow me to control for ability outside the ENEM itself.

3.1 Groups definitions

I decided to focus my analysis on three groups, based on their relevance and the existing literature showing that these groups are prone to disparities in standardized tests.[19] These three groups are defined by gender, socioeconomic status, and self-reported ethnicity. In my dataset, 57% of the respondents are women, 28% are classified as low-SES, and 55% self-identify as *pardo* (mestizo) or *preta* (black). However, since socioeconomic and ethnic gaps are highly correlated (as shown in the Figure A3), suggesting that they may be different measures of the same dimension, I decided to focus on gaps observed across race, but conditional on being low-SES. Among this group, 26% still identify as white. This approach allows me to identify textual triggers specifically associated with ethnicity gaps, as an additional layer of burden in an already underprivileged group.

In order to focus my analysis on individuals facing high-stakes situations, I impose several sample restrictions. First, I retain only high school seniors. This allows me to compare equivalent profiles belonging to the same cohort, thus eliminating problems of self-selection into the test by

[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 A6, the differences are plotted by relevant group.

older individuals, and also controls for curriculum structure. Second, I retain individuals who studied in the traditional track, excluding those in other tracks focused on special needs or adult education. Third, I drop all individuals who did not show up for one of the tests or did not submit their handwritten exam. If one of the tests is missing, applications are not possible, so the fact of dropping could be a signal of private information about performance that is better not to consider.[20]. I also exclude those booklets that were assigned to test-takers with some special disability (for example, with different font sizes) or condition (for example, special rules and conditions for pregnant women) in order to maintain comparability.

3.2 Booklets and typical item design

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.

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. In tables A14, A15, A17 and A16 there is a description of the content they seek to measure by type. 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.

Each item follows a common structure: it starts with a heading, followed by a clear assignment, and then five alternatives are presented. Regardless of the booklet, the order of the alternatives remains the same. Figure A5 shows examples of 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.

4 Item-by-item gaps and textual features

4.1 Measuring gaps

Group differences in performance are difficult to study from individual data. Therefore, I propose a "funnel approach" to study which features considered in the text of the items trigger differences between groups. The intuition is to proceed in an inductive way. First, for each item, I predict the magnitude of the achievement gap it induces for four relevant groups: gender, income, race, and rural achievement gap. The choice of these groups is based on the literature and the specific context of Brazil.

[20]This is also important because, as highlighted in Reyes et al. (2023), there are high school seniors who take the test regardless of their goal of going to college.

In this way, it is possible to relate the characteristics of the items to the achievement gaps I observe. I begin by estimating only the gaps caused by each item in the following simple way:

$$Correct_{i|Q=q_i} = \alpha_0 + \alpha_1 \cdot Group_{gi} + \tau_p + \epsilon_i \tag{1}$$

From where I retrieve, for each item, the value $\hat{\alpha_{1,qg}}$, which corresponds to the observed performance gap, conditional on an specific item $Q = q_i$, between the reference group and the group included in the dummy, g, when just the position along the test of the item p is absorbed.

Accounting for the order of the item in the specification is crucial because, as documented by Reyes (2023), there are significant differences in performance depending on whether the item appears early or later in the test, due to fatigue. This fatigue may also have differential effects across groups Brown et al. (2022). One advantage of my data is that there is no item that shows up in two or more different year-applications. This make that the only source of variability that I need to address is the other within the booklet.

Then I merge these gaps with item's characteristics that likely might explain groups disparities such as: readability[21], length of the text measured as number of words, the use of figures, tables or chemical, the sentiment score[22] each item generates and if the text is grounded or not.[23]

These characteristics are relevant to consider as a first candidate to account for systematic differences between groups based on the literature review. Table A2 shows how these characteristics vary by subject.[24]

Then, the next step will be to map the performance gaps estimated for each item to the text itself that each individual faced. To do this, I will proceed by transforming the text into numerical vectors in order to incorporate this information into a L2 regularization model in a manner similar to Iaria et al. (2022). I do so taking the text, I clean it removing special characters, stop words and I proceed with the lemmatization of it using the Portuguese corpora in SpaCy. This allows me to predict with words that have the most predictive power explaining the gaps estimated following equation (1).

[21] 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).

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

[23] 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.

[24] Over 2,275 items available in the 13-year period (excluding those that ask about), I have a total of 2,201 that are considered in the analyses. I exclude 74 because they have characteristics that could bias my estimates. In particular, because they are outside the 99% of the observations in terms of the parameter b of difficulty, in terms of the proportion of individuals whose parents have a college education, and in terms of whether the item did not reach the convergence to the characteristic curve according to the IRT models.

4.2 Gaps and abilities

The first relevant dimension to document is how the gaps vary across the ability distribution. To explore this, I analyze the gaps by splitting the sample according to ability quartiles, comparing groups within the same ability quartile. This analysis is crucial because the stakes differ by quartile, with the highest stakes for top performers, as their performance is most likely to affect their chances of being admitted to college programs. Figure A8 illustrates the differences in the distribution of each gap when comparing the first quartile (bottom performers) with the fourth quartile (top performers). The distribution suggests that most of the effect is concentrated at the top of the distribution. This may indicate that pressure and expectations regarding performance play a significant role.

4.3 Gaps by subjects and competences

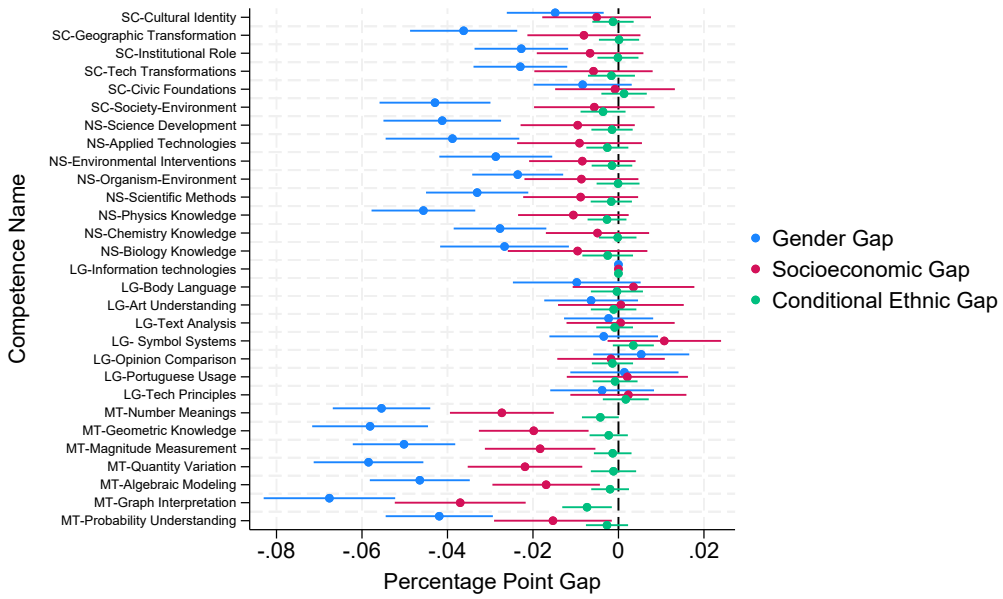
The first stylized fact is to document how these gaps vary by subject. Is well established in the literature that the subjects more prone to trigger gaps are mathematics and sciences. Figure A9 shows that this pattern also holds for the ENEM test-takers. It can be seen that natural sciences and math tend to behave similarly, regardless of the measure used. In terms of gender gaps, Language shows a distinctive behavior with a distribution centered around zero. The other gaps are skewed to the left, exhibiting similar behaviors. For socioeconomic and ethnic gaps, social sciences tend to be more skewed to the left compared to the other distributions. Interestingly, in all cases, a non-negligible portion of the items produce non-negative gaps, with an average of 20% of the items for gender, 11% for socioeconomic status, and 21% for ethnicity, conditional on being low-SES.

Not all competences behave homogeneously within subjects. Figure 1 plots the fixed effects for each competence, controlling for item characteristics, including IRT parameters and quartile difficulty. The results reveal that the gaps studied vary across different competences. While most gaps do not show significant variation across subjects and competences (with the exception of math), the gender gap exhibits considerable variation. This pattern can be explained by stereotype threat theory. Competences like Physics in Natural Science, or Graph Interpretation and Measurement in Math, are those that induce the greatest differences in the gender gap. A similar pattern is observed in Social Sciences, where more stereotyped competences trigger the largest differences.

I investigated trends over the years that could indicate specific actions by the test designers, but no significant patterns were found. This suggests that the test designers may not be actively taking steps to reduce these gaps, which reduces the risk to my identification strategy.

4.4 Gaps and objective text-based measures

An additional step is to document the extent to which the gaps depend on objective measures. In this analysis, I include factors such as readability score, sentiment score, number of words, the presence of auxiliary elements like figures or tables, whether the text is grounded in real-



Note: This figure shows coefficient associated with each of the competence fixed effects when regressing the gender, socioeconomic and conditional ethnic performance gaps and controlling for the item's characteristics. The reference corresponds to Language, Information Technologies. So all gaps are estimated relative to this one. Confidence Intervals at 99% level.

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

world situations, whether it depicts identifiable characters, and the IRT parameters. In section E I describe the literature showing evidence regarding the possible impact of these features on performance.

Table 1 presents the coefficient estimates when the three achievement gaps (in percentage points) are regressed against several item factors. A general pattern emerges, showing that the discrimination and guessing parameters tend to reduce gaps, while the difficulty parameter tends to increase them, which aligns with IRT theory. Regarding SES gaps, no statistically significant features are associated. For gender gaps, it is notable that the presence of female characters reduces the gap by nearly 50% compared to the constant. Additionally, the use of figures increases the gap by about one-third, whereas lengthier text tends to reduce the gap. In the case of ethnic gaps, the opposite effect is observed concerning the length of the text.

4.5 Gaps and relevant words

The R-squared values in the above table indicate that there is still room to uncover meaningful explanations for the observed gaps. Therefore, I proceed to analyze the text as data to determine the extent to which specific words or bi-grams can explain these gaps. I employ an L2-regularization model (RIDGE)[25] with a grid search for the optimal alpha parameter. As is standard practice, I

[25]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

Table 1: Performance gaps based on items' characteristics (p.p)

	(1) SES gap	(2) Gender gap	(3) Ethinc Gap
Readability Score	-0.018	-0.104	-0.010
	(0.075)	(0.080)	(0.028)
Lenght in Words (SD)	0.048	0.222***	-0.088***
	(0.076)	(0.078)	(0.031)
Figures	-0.168	-0.499***	-0.064
	(0.155)	(0.174)	(0.057)
Tables	-0.047	-0.488	0.191*
	(0.290)	(0.400)	(0.104)
Sentiment Score	0.677	0.445	0.156
	(0.556)	(0.609)	(0.228)
Grounded	-0.044	-0.321	-0.076
	(0.226)	(0.256)	(0.086)
Man Character	0.109	0.002	0.127*
	(0.187)	(0.203)	(0.074)
Woman Character	0.309	0.796***	0.042
	(0.194)	(0.202)	(0.080)
Underprivileged	0.040	-0.073	0.043
	(0.256)	(0.258)	(0.108)
Privileged	-0.127	-0.183	0.036
	(0.178)	(0.197)	(0.068)
Discrimination P.	-0.991***	-0.390***	-0.058**
	(0.083)	(0.080)	(0.028)
Difficulty P.	0.213*	0.519***	0.138***
	(0.110)	(0.108)	(0.039)
Guessing P.	14.482***	5.634***	3.646***
	(1.050)	(1.119)	(0.393)
Const.	-0.361	-1.538***	-0.394***
	(0.428)	(0.408)	(0.147)
Area FE	Yes	Yes	Yes
Diff Quartile FE	Yes	Yes	Yes
Obs.	2,201	2,201	2,201
R-squared	0.75	0.38	0.66

Standard errors in parentheses

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

exclude stopwords and words that appear either too infrequently or too frequently in the text, as these are unlikely to provide meaningful information. I do also perform the lemmatization of the words in order to reduce dimensionality. I then focus on the top 50 words predicted to reduce the gap the most (highlighted in blue) and the 50 words that are predicted to exacerbate the gap the most (highlighted in red). The figure below presents the word cloud for each of the gaps considered.

Figure 2 shows the most predictive words for each of the three gaps. For the SES gap, the pattern is clear, with words like "price," "land," "product," "urban," and "consumption" inducing wider gaps. The effect sizes for these words range from 0.007 to 0.005 percentage points. While items that focus on practical, everyday scenarios, physical sciences, and historical or political content (e.g., "work," "acid," "political") might reduce the gap.. Additionally, as shown in Figure A11, the model has a decent fit for this gap.

For the gender gap, certain items might widen the gap by favoring those that emphasize language and reading comprehension, emotional and social contexts, and cultural or creative topics (e.g., "text," "family," "music"). Conversely, items related to STEM and quantitative reasoning, abstract analytical contexts, and physical geography (e.g., "energy," "rate," "graph") might reduce the gap.

Finally, for the conditional ethnic gap, questions related to political, social, and cultural contexts, as well as scientific and technical topics, might reduce the gap (e.g., "political," "cultural," "mass"). On the other hand, items focusing on physical and biological sciences, urban environments, socioeconomic issues, and text interpretation might widen the gap (e.g., "cell," "urban," "text"). It is also worth noting that the effect sizes for the words associated with this gap are an order of magnitude smaller than those observed for the other gaps.

4.6 Gaps by Hierarchical LDA

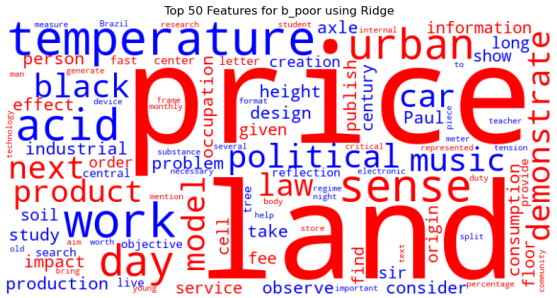
Next, I move into an unsupervised approach. To do so, I proceed using topic modeling based on Latent Dirichlet Allocation (LDA) algorithms. LDA views each item as consisting of different topics. In other words, each document is a mix of different topics. In the same way, topics are mixed from words. This model exploits the fact that when talking about the same topic authors tend to use similar words. Hence, in texts about the same topic, similar words tend to co-occur. The co-occurrences of words are used by LDA to describe each topic as a probability distribution over words and each document as a probability distribution over topics. The use of these algorithms allow us to uncover the latent structure of the text.

Using Hierarchical LDA I can select the number of topics in an unsupervised fashion, based on the coherence score. Figure A12shows the coherence score for different number of topics. The coherence is maximized at 66 topics.

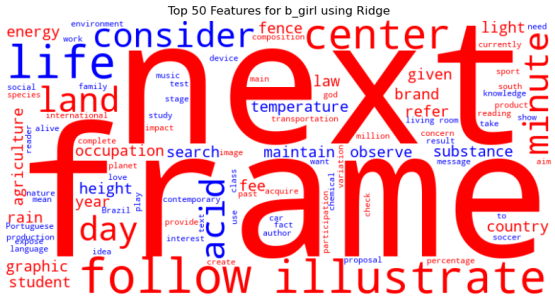
Then I proceed running a L1-regularization (LASSO) model that includes all the item's features, the competence fixed effects and the 66 topics discovered to predict the gaps. The advantage of using L1 is that it will shrink to zero irrelevant information, highlighting the explanatory power of

them entirely.

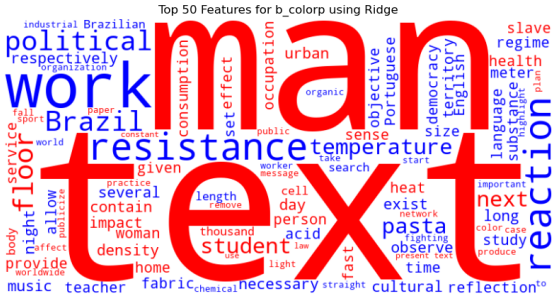
Figure 2: Words with the highest predicting power by type of performance gap



(a) SES gap



(b) Gender Gap



(c) Cond. Ethnic Gap

Note: These images show the words in the text that have the highest predictive power in explaining the observed gaps. Those in blue are words that influence the gaps by reducing them, while those in red widen the gap. The words were estimated using a Ridge model with an alpha defined by grid search and including the item features and competences fixed effects as controls. Figure A11 shows the fit of the models.

the selected factors.

Table A6 provides the coefficients for each of the elements. Then I take those topics and its distribution of words, and aks GPT how to conceptualize the topic based on its elements.

4.7 Hypotheses discovering

This section examines the extent to which the textual features of an item can explain the performance gaps observed across different groups. The distribution of these gaps reveals a high level of heterogeneity that cannot be fully accounted for by subject matter or the specific competence being assessed. Therefore, it is possible that the text itself, along with the various cues it triggers in different test-takers, contributes to these disparities.

Machine learning (ML) offers the potential to uncover patterns that might be overlooked by human analysis (Ludwig and Mullainathan, 2024; Wang et al., 2023), but using these tools for discovery often comes at the expense of interpretability and clarity (Messeri and Crockett, 2024). OpenAI's GPT-4 plays a crucial role in processing text data and generating coherent hypotheses, as demonstrated in recent studies (Charness et al., 2023).

To delve deeper into the observed gaps, I employ machine learning techniques while controlling for competence fixed effects and all item features. This analysis uncovers specific patterns based

on certain words and topics composed of related terms. In a manner similar to Batista and Ross (2024), I leverage OpenAI's GPT-4 to generate hypotheses for each gap, using the 50 most relevant words and top predictive topics as inputs. These words and topics are analyzed to understand how they may contribute to either the widening or narrowing of the gaps, providing a solid foundation for the hypotheses generated by GPT-4.

The six hypotheses provided are the following:

1. "Items that focus on financial situations, business-related decision making, and abstract economic concepts (e.g., value, company, client) might widen the SES gap."

2. "Items that involve practical problem-solving scenarios rooted in real-life contexts, particularly those that relate to material conditions, everyday challenges, and concrete tasks (e.g., water, medicine, temperature, production, work) might reduce the SES gap."

3. "Items that involve abstract scientific concepts, measurements, and dynamic processes (e.g., wave, height, increase, produce, transportation, and technology) might widen the gender gap."

4. "Items that involve practical problem-solving scenarios in the context of narratives, social interactions, creative expression, emotions, or real-life decision-making (e.g., text, family, music, medicine, purchase, response) might reduce the gender gap."

5. "Items that involve abstract social phenomena, legal concepts, and technical terminologies (e.g., value, company, law, message, and impact) might widen the (conditional) ethnic gap."

6. "Items that focus on historical, political, and cultural contexts, as well as tangible scientific concepts (e.g., political, Brazil, music, culture) might 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 competences 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 Testing the hypotheses using Panel Data

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. To do this, I take advantage of the panel structure of the data by exploiting within-individual differences in performance when a flagged item appears. The intuition is as follows: each year, each test-taker faces 180 items sequentially, divided into blocks and by subjects. The order in which the items are presented depends on the random assignment

of the booklets. Assuming that test-takers proceed through the items sequentially—a reasonable assumption given the paper format— the appearance of a flagged item can be assumed to be quasi-random. Thus, the progression through the test for each individual provides an interesting natural experiment, with some flagged items appearing earlier for some and later for others.

Examining within-individual variation due to exposure to flagged items offers a more insightful approach that allows for the identification of effects in a more demanding setting, where identification comes from *within* individual variation. This step is crucial because the relationships observed in the previous analysis might arise due to *between* individual differences, potentially driven by unobserved factors that are not consistent across individuals. To better capture these dynamics, I construct an individual-by-subject Fixed Effect, as the same individual may excel in one area (e.g., math) while performing less well in another (e.g., social science).

Furthermore, the empirical strategy includes order-item fixed effects, competence fixed effects (which have proven to be relevant), and region of address fixed effects to account for geographic differences, as well as year fixed effects to control for cohort differences, such as changes in curriculum, coverage, or other factors. To enhance precision, I also include several item controls to account for differences in readability, length, use of figures or tables, sentiment score, grounding, and difficulty quartile fixed effects. With all these considerations, the main specification is:

$$Correct_{iq} = \gamma_0 + \gamma_1 \cdot Flagged_q \times Group_i + \gamma_3 \cdot Flagged_q + \gamma_4 \cdot Group_i + \gamma_5 \cdot X'_q + \phi_{is} + \varphi_y + \nu_c + \kappa_r + \eta_p + \varepsilon_{iq} \quad (2)$$

Where $Correct_{iq}$ corresponds to an indicator of the individual's correct marking of the item[26], $Flagged_q$ is an indicator for the item representing the trait in its text, $Group_i$ is an indicator for the individual belonging to the group of interest (low SES, female, non-white, or rural), X'_q is a vector of item characteristics (readability, number of words, figures, and tables), ϕ_{is} corresponds to the individual times subject fixed effect, φ_y is the year of application fixed effect, ν_c is the competence fixed effect based on the reference matrix, κ_r is the region fixed effect, η_p is the booklet-specific position of the item fixed effect, and χ_q is the item difficulty quartile according to the proportion of individuals whose parents have a college degree who got the item correct. Standard errors are clustered at the individual level.

The parameter of interest is γ_1, which reflects the average performance gap induced by the marked items in the group of interest when the baseline is the disjoint group. The main assumption here is that the effects of the stressor and its group interactions are identified conditional on the difficulty, the target of the question, and the individual's ability. Note that the baseline is the item just prior to the one marked. Standard errors are clustered at the individual level.

[26]Some specifications also test the individual's conditional probability on this item, i.e., the interaction of the IRT structural parameters with the latent trait θ that is the basis of the official scores.

5.1 Sample restrictions

As I previously described, the analysis regarding the gaps has been done using all observation in each year meeting the inclusion criteria (i.e.: being high-stakes). But, since thousands of hundreds of test-takers take the ENEM each year, and since each of them would have 180 questions, the "curse of dimensionality" is binding, and I need to proceed with some sampling. Therefore, for each year, I draw a random sample of 10,000 per year form those that meet the inclusion criteria.[27] individuals who meet the conditions stated above. A sample big enough to have statistical power and being able to identify even small effect sizes. Table A1 provides the summary statistics of the individuals sampled.

5.2 How to flag items?

So far, I have developed testable hypotheses, but to test them, I need to identify the "treated" items. This requires a systematic approach to assigning items that potentially exhibit the properties linked to performance differences among groups, as suggested by the hypotheses.

A naive approach might involve assigning items based on the topic that represents them most strongly, but this method is neither clear nor intuitive. Each topic is a distribution of words, and each item is a mixture of topics. Consequently, assigning each item to a single representative topic risks losing valuable information. Instead, the approach I adopt involves flagging items based on their textual content using keyword-based classification with contextual filtering. For each performance gap, I will identify items that are flagged as either widening the gap or reducing the gap.

I proceeded by asking GPT-4 to generate a list of keywords for each of the hypotheses after providing it with the respective hypothesis.[28] I then used these lists of keywords to classify each item based on the occurrence of at least two keywords in the item. I do so to avoid false positives. Manual cross-checks suggest that the performance of the labeling is sufficiently accurate.

After completing the tagging process, a total of 1,098 items were categorized across the six defined categories. However, there is some overlap, with 40% of the items being labeled in two or more categories. Table A7 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 A8 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.

[27] The seed for each year I chose for these draws is based on www.random.org

[28] For example, for the hypothesis associated with widening the gender gap, the generated list includes: "wave", "height", "increase", "produce", "transportation", "technology", "energy", "force", "pressure", "speed", "frequency", "radiation", "volume", "mass", "density", "resistance", and "power," among others.

5.3 Cueing and performance

Table 2 presents the results of the estimation of equation 2. Each column displays the estimates for the factors influencing the probability of correctly answering a particular item (0/1), including interaction terms for each proposed hypotheses. The coefficients are rescaled to represent percentage points. The even-numbered columns also include Individuals per Area Fixed Effects. The sample in columns (5) and (6) is limited to Low SES individuals for consistency with the previous analysis.

Regarding the first two hypotheses related to SES gaps, the results suggest that these hypotheses hold when individual fixed effects are included. Specifically, the presence of an item flagged as containing financial terms widens the SES gap by 1.4 percentage points (a 22% increase from the average gap of 6.1 percentage points). Conversely, items that include more real-world situations, as described in H2, help reduce the gap by 0.8 percentage points.

For the hypotheses related to gender gaps, it appears that individual fixed effects might play a role in explaining the widening gender gap, as the sign of the coefficients flips when these effects are included. This change may be driven by a subset of women. H4, which is associated with items that reduce the gender gap, is supported by the data regardless of the specification, with an estimated reduction of 0.6 percentage points.

Regarding the ethnic gap, none of the hypotheses are supported by the data, with the estimates being highly imprecise. As a result, the two associated hypotheses are rejected and are not included in the subsequent analyses.

Finally, to account for the overlapping effects of items that have properties of more than one hypothesis, column (7) presents the estimates for all the tags together. These results are very stable.

5.4 Heterogeneous Effects

5.4.1 Role of Ability

In Section 4.2 of the descriptive analysis, I demonstrated that ability is a significant factor in explaining the observed average gaps. Therefore, a natural next step is to analyze how the effects vary across different levels of ability. This is crucial because the stakes are highest at the top end of the ability distribution, and if the effects are more pronounced among high-ability individuals, it could indicate that stereotype threat is most potent where expectations are greatest.

Table 3 presents the results of model (7), fully saturated as in the previous table, but with the sample split according to individuals' positions within each ability quartile. Note that this ability measure corresponds to the latent trait that is estimated for each of the test-takers based on the IRT models, therefore it is comparable across groups and years.

An interesting pattern emerges: regardless of the type of gap, the widening effect is most pronounced among individuals with higher ability. This suggests that the higher the ability, the greater the detrimental effect observed, with the most severe case being the gender gap, which

Table 2: Hypotheses and their impact in Performance (p.p)

	(1) Correct	(2) Correct	(3) Correct	(4) Correct	(5) Correct	(6) Correct	(7) Correct
Low SES	-6.141***						
	(0.060)						
SES Wd	0.748***	1.092***					1.191***
	(0.045)	(0.045)					(0.046)
SES Wd × Low SES	-1.551***	-1.424***					-1.420***
	(0.081)	(0.081)					(0.081)
SES Rd	-0.692***	-0.320***					-0.329***
	(0.031)	(0.031)					(0.032)
SES Rd × Low SES	1.128***	0.810***					0.806***
	(0.056)	(0.055)					(0.055)
Girl			-3.583***				
			(0.061)				
Gender Wd			0.277***	-0.078*			0.021
			(0.044)	(0.043)			(0.043)
Gender Wd × Girl			-0.486***	0.226***			0.206***
			(0.056)	(0.054)			(0.054)
Gender Rd			-2.423***	-0.967***			-0.963***
			(0.045)	(0.044)			(0.044)
Gender Rd × Girl			2.306***	0.644***			0.645***
			(0.059)	(0.056)			(0.056)
Non-White					-1.334***		
					(0.094)		
Ethnic Wd					-0.476***	-0.294***	-0.266***
					(0.107)	(0.104)	(0.042)
Ethnic Wd × Non-White					-0.102	0.138	-0.125**
					(0.122)	(0.118)	(0.055)
Ethnic Rd					-0.143	0.067	0.257***
					(0.221)	(0.218)	(0.084)
Ethnic Rd × Non-White					-0.411	-0.058	-0.104
					(0.256)	(0.251)	(0.114)
Const.	42.898***	41.722***	43.444***	41.909***	35.975***	35.547***	41.838***
	(0.064)	(0.054)	(0.072)	(0.054)	(0.135)	(0.104)	(0.054)
Item's Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Indiv X Area FE	No	Yes	No	Yes	No	Yes	Yes
Difficulty FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Obs.	22,006,964	22,006,964	22,006,964	22,006,964	5,914,067	5,914,067	22,006,964

Clustered Robust Standard Errors at Individual Level in parentheses.

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

The estimations follow equation 2.

Table 3: Hypotheses and their impact in Performance (p.p) by Ability Quartile

	(1) Correct Q1	(2) Correct Q2	(3) Correct Q3	(4) Correct Q4
SES Wd	-0.471***	-0.116	0.869***	2.809***
	(0.090)	(0.091)	(0.090)	(0.086)
SES Wd × Low Income	0.092	-0.162	-0.436**	-0.593**
	(0.133)	(0.149)	(0.175)	(0.251)
SES Rd	-0.033	0.111*	-0.049	-0.417***
	(0.064)	(0.065)	(0.063)	(0.059)
SES Rd × Low Income	0.187**	-0.011	0.353***	-0.016
	(0.093)	(0.105)	(0.121)	(0.168)
Gender Wd	0.933***	0.510***	0.256***	-0.151*
	(0.087)	(0.088)	(0.086)	(0.082)
Gender Wd × Girl	-0.027	-0.341***	-0.479***	-1.064***
	(0.102)	(0.106)	(0.108)	(0.115)
Gender Rd	-0.517***	-0.963***	-1.130***	-0.987***
	(0.083)	(0.092)	(0.090)	(0.082)
Gender Rd × Girl	0.486***	0.785***	0.668***	0.897***
	(0.104)	(0.114)	(0.115)	(0.111)
Ethnic Wd	-0.872***	-0.254***	-0.282***	-0.199***
	(0.087)	(0.087)	(0.084)	(0.074)
Ethnic Wd × Non-white	0.315***	0.319***	0.117	-0.200*
	(0.104)	(0.108)	(0.111)	(0.114)
Ethnic Rd	-0.094	0.861***	1.147***	0.258*
	(0.179)	(0.177)	(0.169)	(0.149)
Ethnic Rd × Non-White	0.065	-0.579**	0.144	0.082
	(0.219)	(0.229)	(0.234)	(0.240)
Const.	23.829***	40.193***	47.314***	57.295***
	(0.101)	(0.108)	(0.106)	(0.105)
Item's Controls	Yes	Yes	Yes	Yes
Indiv X Area FE	Yes	Yes	Yes	Yes
Difficulty FE	Yes	Yes	Yes	Yes
Obs.	5,508,994	5,501,288	5,497,670	5,499,012

Clustered Robust Standard Errors at Individual Level in parentheses.

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

Quartiles by Ability

shows a 1 percentage point difference from the first to the fourth quartile. However, this pattern is not mirrored in the reduction channel. For the SES gap and the ethnic gap, there appears to be no effect among the top performers, while the gender gap shows a consistent effect across all quartiles, with a noticeable jump in the last quartile.

The pattern based on ability for the SES widening channel is also helpful to reject one alternative explanation. It could be perfectly possible that these effects are explained by financial literacy problems, where the population coming from low-SES context is more deficient. However, if this were the case, one would expect to observe more pronounced effects at the lower end of the skill distribution.

5.4.2 Role of Stereotyped Items

In line with the evidence provided in Section 6, another interesting source of variation arises from observing how stereotyped and non-stereotyped items differentially affect the stated hypotheses. In this context, a stereotyped item is one whose competence is framed in a way that aligns with existing stereotypes about performance. Since these stereotypes are present across all competences related to SES and ethnic gaps, I focus my attention on gender gaps.

Based on the content of each of the competences described in the Reference Matrix, I asked ChatGPT to evaluate whether a specific description is likely to carry stereotypes against women's performance. Consequently, I conducted the heterogeneous analyses solely for the gender gap hypotheses. The results show that 31.3

As shown in Table A9, items belonging to more stereotyped categories tend to amplify the widening effect observed in items that exhibit the properties outlined in H3. This effect is particularly pronounced when the analysis is conducted on the entire sample, where the impact is largely concentrated in these items. However, when examining the within-quartile estimates, we observe that the effects of flagged items, whether stereotyped or not, are almost identical in size. This suggests that the effects on top-performing females differ, likely due to the additional pressure and expectations they face. In terms of the reduction channel, it can be seen that in stereotyped areas, the effect is nearly halved, with no effect observed at the top of the ability distribution.

5.4.3 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 A10 summarizes the results for the gender gap, while Table A11 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 exhibit H3 characteristics, thereby reducing gender gaps. However, this mirroring effect does not appear to be effective in addressing the widening effects associated with SES gaps.

5.5 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.[29]

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

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 A13 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.

[29] 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.

6 Discussion and Conclusion

This paper demonstrates that the contextual features of test questions significantly affect performance disparities in high-stakes standardized tests, based on 13 years of data from 3.8 million test-takers. The analysis proceeds in two stages.

First, individual performance data is matched to estimate item-by-item performance gaps for three key groups: gender, socioeconomic status (SES), and ethnicity. Leveraging text analysis techniques and the support of large-language models (LLMs) like ChatGPT, six testable hypotheses are proposed in a data-driven fashion—two for each gap, with hypotheses predicting both widening and reducing effects. These hypotheses are grounded in stereotype threat theory (Steele and Aronson, 1995), which suggests that the salience of group identity can trigger disparities.

In the second stage, these hypotheses are tested using the panel structure of the data to study performance within individuals over their test. The analysis finds supporting evidence for four out of the six hypotheses, with the ethnicity-related hypotheses not supported by the data. For the SES gap, the widening channel is driven by items involving abstract and financially related content, resulting in a gap increase of 1.4 percentage points (23% relative to the SES gap). The reducing channel is associated with items related to grounded, daily activities, with a 0.8 percentage point effect, particularly among lower-ability test-takers. For gender gaps, the widening effect is linked to items involving abstract scientific concepts and measurements, with the effect concentrated in the top half of the ability distribution. This effect is most severe for female test-takers in the fourth quartile, with an effect size of 1.06 percentage points (30% of the average gender gap). The reduction channel, on the other hand, is driven by items that involve practical problem-solving scenarios, creativity, emotions, and social interactions, with a consistent effect of 0.8 percentage points across quartiles.

Heterogeneity analyses reveal several important insights. First, the widening channels are concentrated among top performers, which aligns with stereotype threat theory, as these groups face the greatest pressures and expectations. In contrast, the reduction channels show relatively consistent effects across ability quartiles. Second, in the case of gender gaps, items associated with stereotyped competences exacerbate the widening effect, except among females in the top quartile of ability. There is no heterogeneous effect on the reduction channel based on whether an item is stereotyped. Third, for gender gaps, a positive mirroring effect is observed: when female test-takers encounter a female character in the item, the widening effect is completely offset (consistent with stereotype threat theory, the reduction channel is unaffected). However, for SES gaps, the mirroring effect actually amplifies the widening channel. This suggests that items featuring concepts linked to gender-related widening should be paired with female characters to mitigate the impact.

Stereotype threat theory is a widely accepted explanation for performance disparities across groups. It is appealing due to its simplicity and the promise that simple interventions can significantly reduce performance gaps, benefiting underprivileged groups. However, most of the existing evidence comes from underpowered samples or low-stakes settings (Priest et al., 2024; Shewach

et al., 2019; Flore et al., 2018). Advancing our understanding in this area is crucial for confirming or challenging the theory in real-world, high-stakes scenarios Levitt and List (2007).

In this paper, I advance the study of stereotype threat theory using one of the largest high-stakes testing scenarios in the world. The novel approach adopted here leverages the textual content of test items in a data-driven fashion and utilizes the capabilities of Artificial Intelligence platforms like ChatGPT. This paper is intended to help education researchers, organizations, and policymakers generate new insights into what drives performance gaps across groups in testing. Beyond the findings themselves, this approach offers a practical tool for researchers, organizations, and policymakers seeking to improve test design and understand its impact on different groups. This method could also help uncover other performance gaps that have not yet been explored in the literature.

For better or worse, standardized tests will continue to play a significant role, with consequences that provide valuable data for researchers and policymakers. This data is crucial for advancing our understanding of the mental processes that drive performance differences, which in turn serve as gateways to new and better development opportunities. Expanding the availability of test questionnaires, as Brazil has done, is essential for progress in this area.

References

Adukia, A., Eble, A., Harrison, E., Runesha, H. B., and Szasz, T. (2021). What we teach about race and gender: Representation in images and text of children's.

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.

Atkinson, R. C. and Geiser, S. (2009). Reflections on a century of college admissions tests. *Educational Researcher*, 38(9):665–676.

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.

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

Batista, R. M. and Ross, J. (2024). Words that work: Using language to generate hypotheses. Working paper. Available at https://drive.google.com/file/d/1VdD7u-2xG4zfyxc2JU4e-zqiJ1DsdDUV/view.

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.

Beilock, S. L., Rydell, R. J., and McConnell, A. R. (2007). Stereotype threat and working memory: mechanisms, alleviation, and spillover. *Journal of Experimental Psychology: General*, 136(2):256.

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

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

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

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. and Chen, Y. (2020). Social identity, group behavior, and teams. *Annual Review of Economics*, 12(1):691–713.

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

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.

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.

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

Galla, B. M., Shulman, E. P., Plummer, B. D., Gardner, M., Hutt, S. J., Goyer, J. P., D'Mello, S. K., Finn, A. S., and Duckworth, A. L. (2019). Why high school grades are better predictors of on-time college graduation than are admissions test scores: The roles of self-regulation and cognitive ability. *American Educational Research Journal*, 56(6):2077–2115.

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

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

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.

Iaria, A., Schwarz, C., and Waldinger, F. (2022). Gender gaps in academia: Global evidence over the twentieth century. *Available at SSRN 4150221*.

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.

Levitt, S. D. and List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic perspectives*, 21(2):153–174.

Levy, B. (1996). Improving memory in old age through implicit self-stereotyping. *Journal of personality and social psychology*, 71(6):1092.

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.

Mani, A., Mullainathan, S., Shafir, E., and Zhao, J. (2013). Poverty impedes cognitive function. *science*, 341(6149):976–980.

Manning, B. S., Zhu, K., and Horton, J. J. (2024). Automated social science: A structural causal model-based approach.

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*.

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

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.

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

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.

Schmader, T. and Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of personality and social psychology*, 85(3):440.

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.

Shih, M., Ambady, N., Richeson, J. A., Fujita, K., and Gray, H. M. (2002). Stereotype performance boosts: the impact of self-relevance and the manner of stereotype activation. *Journal of Personality and social psychology*, 83(3):638.

Shih, M., Pittinsky, T. L., and Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological science*, 10(1):80–83.

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.

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.

Walker, M. E. and Bridgeman, B. (2008). Stereotype threat spillover and sat® scores. *ETS Research Report Series*, 2008(1):i–10.

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.

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.

Zwick, R. (2019). Assessment in american higher education: The role of admissions tests. *The ANNALS of the American Academy of Political and Social Science*, 683(1):130–148.

A Appendix Tables

Appendix Table A1: Summary Statistics

	count	mean	sd	min	max
Low Income	2.27e+07	0.28	0.45	0.00	1.00
Girl	2.27e+07	0.57	0.49	0.00	1.00
Bianco	2.27e+07	0.43	0.50	0.00	1.00
Preta	2.27e+07	0.11	0.31	0.00	1.00
Parda	2.27e+07	0.43	0.49	0.00	1.00
Asiatic	2.27e+07	0.02	0.14	0.00	1.00
Indigenous	2.27e+07	0.01	0.08	0.00	1.00
Rural	2.27e+07	0.13	0.33	0.00	1.00
Federal School	2.27e+07	0.04	0.20	0.00	1.00
Private School	2.27e+07	0.23	0.42	0.00	1.00
Public School	2.27e+07	0.73	0.44	0.00	1.00
Father with Higher Education	2.07e+07	0.16	0.37	0.00	1.00
Father with Low Education	2.07e+07	0.24	0.43	0.00	1.00
With access to Internet	2.10e+07	0.75	0.43	0.00	1.00
Spanish Foreign Language	2.27e+07	0.50	0.50	0.00	1.00
Grade in Sciences (IRT)	2.27e+07	486.02	77.51	0.00	833.40
Grade in Social Sciences (IRT)	2.27e+07	528.57	84.78	0.00	873.20
Grade in Language (IRT)	2.27e+07	513.44	72.39	0.00	772.90
Grade in Math (IRT)	2.27e+07	515.97	112.56	0.00	990.70
Total Correct Science	2.27e+07	12.91	5.13	0.00	44.00
Total Correct Social Sciences	2.27e+07	17.47	6.79	0.00	45.00
Total Correct Language	2.27e+07	13.92	7.67	0.00	44.00
Total Correct Math	2.27e+07	12.89	6.00	0.00	45.00

Source: Author's own elaboration based on ENEM 2010-2022

B Appendix Figures

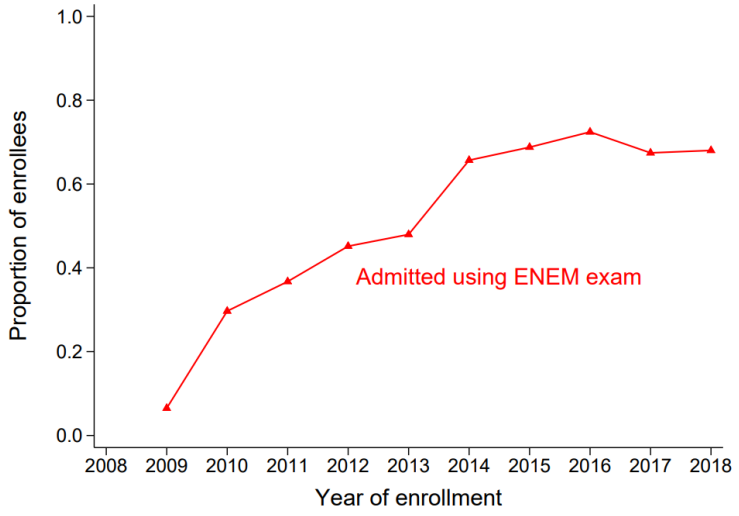
C How the way test are performed affect gaps

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. Moreover, reviewing and ranking the factors is challenging because these estimates were conducted independently, without accounting for potential interactions between factors. For instance, an individual may belong to multiple groups, each influenced by different mechanisms. If empirical strategies do not control for all these possible channels, the estimates are likely to be biased.

Appendix Table A2: Item Statistics

	(1)	(2)	(3)	(4)
	Social Sc	Natural Sc.	Language	Mathematics
Readability Score	-.2589312	-.3903698	.471756	.3869293
Sentiment Score	.1617963	.1344226	.1750167	.1047915
Lenght in Words (SD)	-.3734397	-.2649423	.6023386	-.0334954
Figures	.1597222	.2940141	.2174797	.4017699
Tables	.0034722	.0528169	0	.100885
Grounded	.0069444	.028169	.1504065	.2070796
Man Character	.1388889	.0985915	.2174797	.139823
Woman Character	.1753472	.0616197	.2378049	.0849558
Privileged/Underprivileged	.3055556	.096831	.2703252	.1610619
A parameter	2.159392	2.282638	2.062743	2.048555
B parameter	1.142611	1.394098	.8640873	1.906675
C parameter	.1650543	.1622701	.1458909	.1612244
% Correct with College Father	.5011967	.3766729	.3659024	.394432
Poor Stressor	.03125	.0052817	.0365854	.1221239
Girl Stressor	.03125	.0123239	.0487805	.0247788
Low SES Gap	-.037085	-.0216042	-.0202483	-.0284197
Gender Gap	-.0245527	-.0265529	.0049532	-.0442527
Observations	576	568	492	565



Appendix Figure A1: Proportion of test-takers that were admitted in a Federal University by ENEM. Source Reyes et al. (2023)

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

Appendix Table A3: Top 20 Words Influencing the Gender Gap (Effect in p.p.)

Effects of Gendered Language: Cohen et al. (2023), exploiting a policy change in the Israeli Admission test[30], 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

[30]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.

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

Appendix Table A4: Top 20 Words Influencing the SES Gap (Effect in p.p.)

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– on later on outcomes, and being an important generator

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	work (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

Appendix Table A5: Top 20 Words Influencing the Conditional Ethnic Gap (Effect in p.p.)

of disparities[31]. 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.

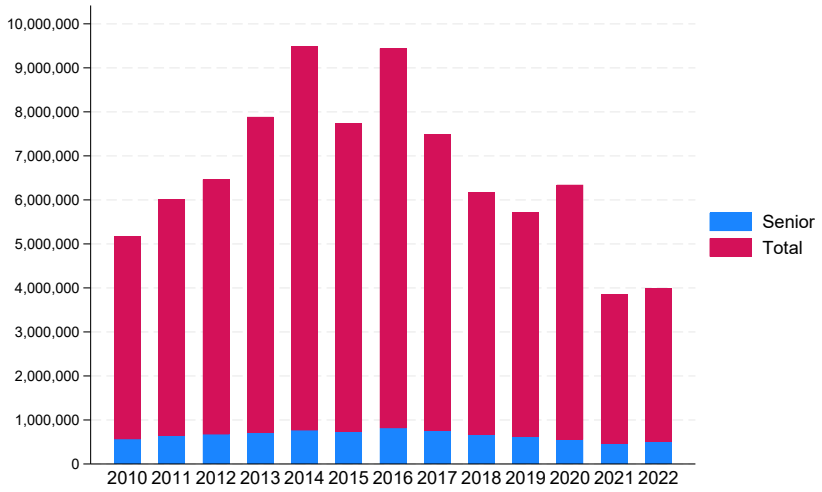
[31] Namely: *By reducing the contribution of endurance gaps to test-score gaps by half, the reform would: (i) Reduce the gender test-score gap by 0.85 percentage points (a 32% decrease from the pre-reform gap of 2.6 percentage points); (ii) Reduce the racial test-score gap by 0.08 percentage points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

Appendix Table A6: LASSO coefficients for each gap

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
woman	1.0106	topic_58	-1.9921
nu_param_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
nu_param_b	0.4269	topic_23	-1.0691
nu_param_b_squared	0.3553	NS-Science Development	-0.8247
topic_1	0.3356	topic_64	-0.5933
NS-Organism-Environment	0.2622	privileged	-0.2780
SC-Tech Transformations	0.1421	NS-Applied Technologies	-0.2409
man	0.1328	grounded1	-0.2401
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
nu_param_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
man	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
woman	0.3917	SC-Society-Environment	-0.5973
words	0.3661	SC-Tech Transformations	-0.4101
NS-Applied Technologies	0.3309	nu_param_a	-0.3239
topic_55	0.2219	privileged	-0.2550
NS-Scientific Methods	0.1010	underprivileged	-0.2356
readability	0.0429	words_squared	-0.1218
		readability_squared	-0.0492
Ethnicity (Color) Gap			
Features Reducing	**Coeff. (p.p.)**	**Features Widening**	**Coeff. (p.p.)**
nu_param_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
man	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	nu_param_b_squared	-0.1760
tables	0.0890	underprivileged	-0.0753
nu_param_a	0.0645	MT-Magnitude Measurement	-0.0627
words	0.0321	figures	-0.0298
readability	0.0071	words_squared	-0.0212
NS-Physics Knowledge	0.0124	grounded1	-0.0191
readability_squared	0.0036	readability_squared	-0.0036

	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%

Appendix Table A7: Distribution of Flagged Items by Subject Category



Appendix Figure A2: Proportion of test-takers that are high-school seniors each year relative to the total of test takers

D 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 competences described for each text. In turn, this competences can be grouped in 6 to 9 categories per subject. The following figures describe each of the specific competences by subject.

E Text-based features and its impact on performance

IN THIS SECTION I DESCRIBE THE EVIDENCE SHOWING WHY THE TEXT-BASED FEATURES I CHOSE ARE RELEVANT. THEY CAN ALSO SHED LIGHT ON THE MECHANISM

F 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

Appendix Table A8: Flagged Items and items characteristicas as predictors

	(1) SES Wd	(2) SES Rd	(3) Gender Wd	(4) Gender Rd	(5) Ethnic Wd	(6) Ethnic Rd
Figure	-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	0.033	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	0.046*	0.018	0.022	-0.010	-0.036	-0.006
	(0.026)	(0.029)	(0.026)	(0.023)	(0.023)	(0.010)
IRT 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)
IRT 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)
IRT 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

Robust Standard Errors in Parentheses

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

Appendix Table A9: Hypotheses and their impact in Performance (p.p) by Stereotyped Item (only Gender Gap)

	(1) Correct All	(2) Correct Q1	(3) Correct Q2	(4) Correct Q3	(5) Correct Q4
Gender Wd	-0.172***	1.361***	0.949***	0.240**	-1.016***
	(0.051)	(0.102)	(0.102)	(0.100)	(0.095)
Gender Wd × Girl	0.588***	0.064	-0.169	-0.205	-0.868***
	(0.064)	(0.121)	(0.124)	(0.126)	(0.135)
Stereotyped	-0.300***	0.417***	0.051	-0.269***	-0.810***
	(0.046)	(0.091)	(0.095)	(0.094)	(0.088)
Gender Wd × Stereotyped	0.288***	-1.583***	-1.468***	-0.162	2.166***
	(0.091)	(0.181)	(0.187)	(0.183)	(0.173)
Girl × Stereotyped	0.726***	0.476***	0.720***	0.769***	0.460***
	(0.054)	(0.102)	(0.109)	(0.111)	(0.111)
Gender Wd × Girl× Stereotyped	-1.182***	-0.291	-0.548**	-0.852***	-0.620**
	(0.115)	(0.218)	(0.229)	(0.234)	(0.243)
Gender Rd	-0.694***	-0.503***	-0.737***	-0.699***	-0.571***
	(0.051)	(0.098)	(0.107)	(0.106)	(0.095)
Gender Rd × Girl	0.561***	0.600***	0.691***	0.516***	0.760***
	(0.066)	(0.123)	(0.135)	(0.136)	(0.129)
Gender Rd × Stereotyped	-0.821***	-0.204	-0.562***	-1.041***	-1.224***
	(0.084)	(0.162)	(0.174)	(0.173)	(0.161)
Gender Rd × Girl × Stereotyped	0.245**	-0.334	0.292	0.457**	0.375*
	(0.109)	(0.204)	(0.219)	(0.225)	(0.223)
Const.	41.868***	23.437***	39.993***	47.371***	57.835***
	(0.056)	(0.103)	(0.111)	(0.109)	(0.107)
Item's Controls	Yes	Yes	Yes	Yes	Yes
Indiv X Area FE	Yes	Yes	Yes	Yes	Yes
Difficulty FE	Yes	Yes	Yes	Yes	Yes
Obs.	22,006,964	5,508,994	5,501,288	5,497,670	5,499,012

Clustered Robust Standard Errors at Individual Level in parentheses.

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

Appendix Table A10: Hypotheses and their impact in Performance (p.p) by Mirroring Effect on Gender Gaps

	(1) Correct All	(2) Correct Q1	(3) Correct Q2	(4) Correct Q3	(5) Correct Q4
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 Area FE	Yes	Yes	Yes	Yes	Yes
Difficulty FE	Yes	Yes	Yes	Yes	Yes
Obs.	22,006,964	5,508,994	5,501,288	5,497,670	5,499,012

Clustered Robust Standard Errors at Individual Level in parentheses.

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

Quartiles by Ability

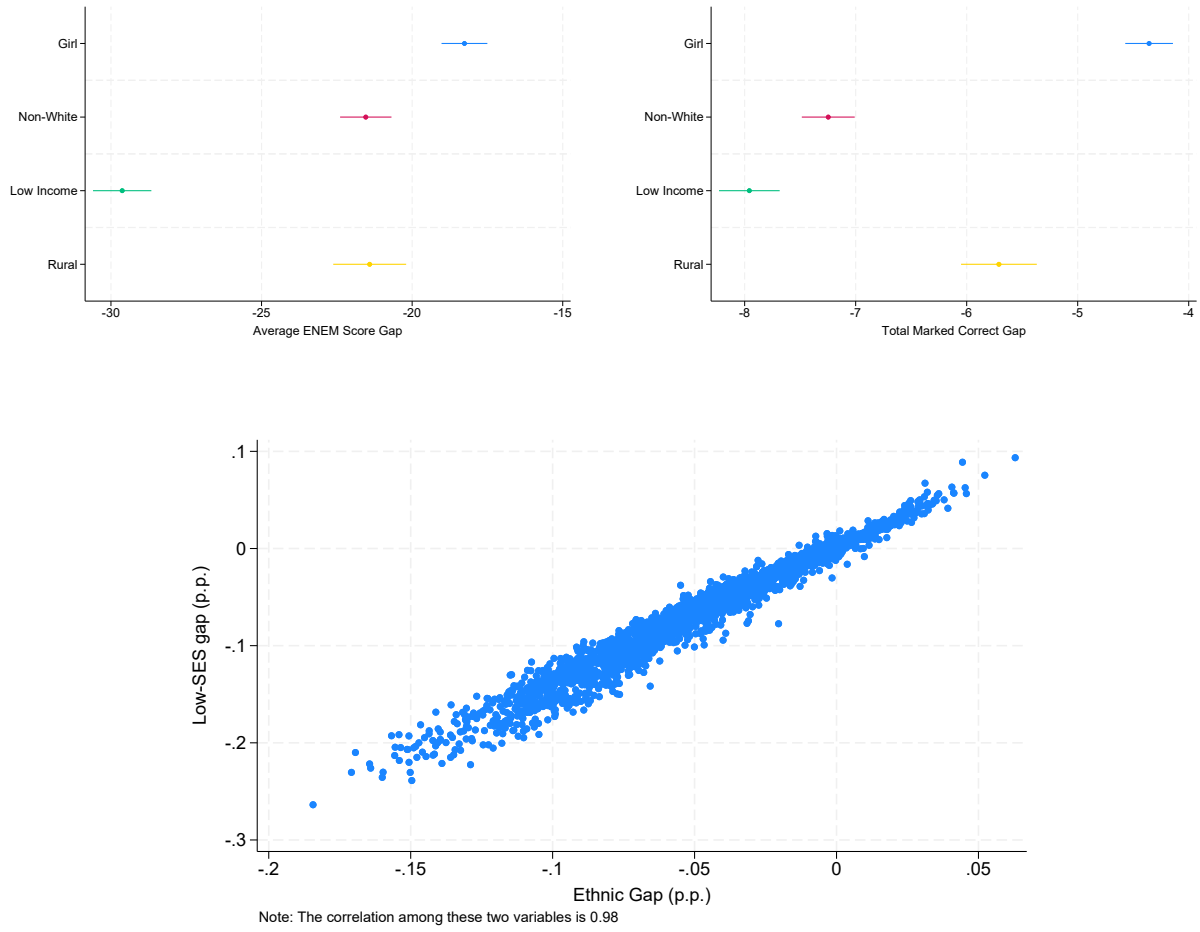
Appendix Table A11: Hypotheses and their impact in Performance (p.p) by Mirroring Effect on SES Gaps

	(1) Correct All	(2) Correct Q1	(3) Correct Q2	(4) Correct Q3	(5) Correct Q4
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 Area FE	Yes	Yes	Yes	Yes	Yes
Difficulty FE	Yes	Yes	Yes	Yes	Yes
Obs.	22,006,964	5,508,994	5,501,288	5,497,670	5,499,012

Clustered Robust Standard Errors at Individual Level in parentheses.

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

Quartiles by ability.



Appendix Figure A3: Correlation among Low-SES gap and Ethnic Gap

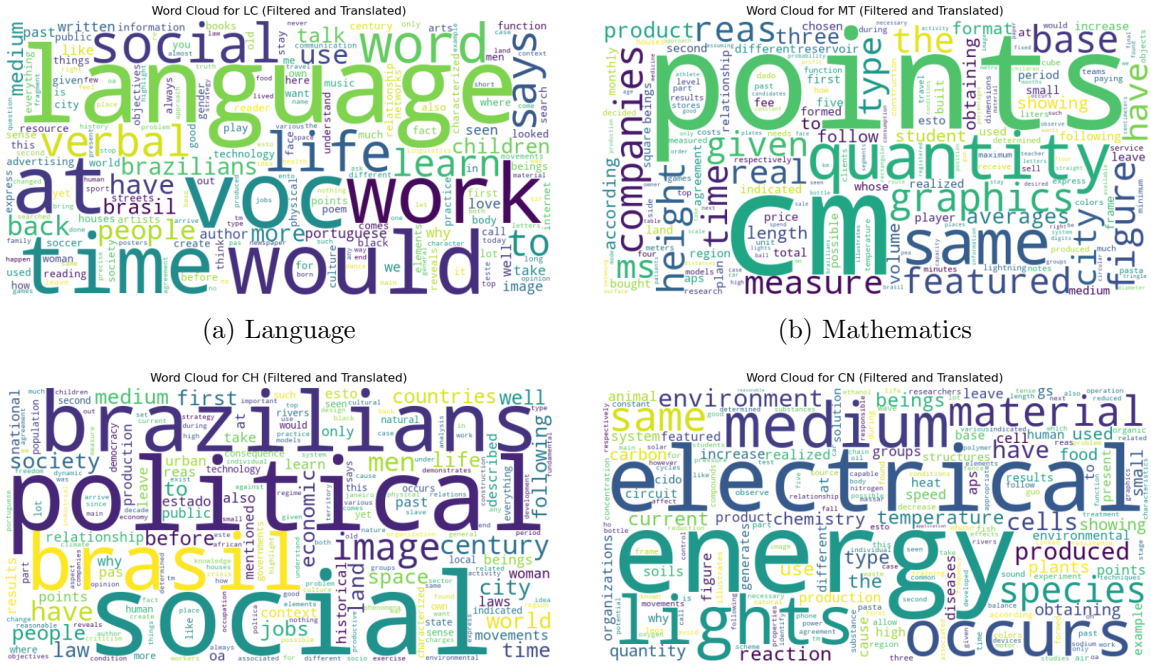
items and where their performance is assessed. Based on the performance of this training sample, the parameters of each items are calibrated. The model used to calibrate is the so called 3PL model.

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

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 A18 shows the so called characteristic curve for some general item.

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.

Appendix Figure A4: Words clouds by Subject



(a) Language (b) Mathematics

(c) Social Sciences (d) Natural Sciences

Note: These figures display the frequency of the most popular words used to contextualize items by subjects. The estimation excludes the 50 most frequent words and also includes a TF-IDF weighting. Original language is Portuguese. All words were translated using the API of DEEPL.

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 A19 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.

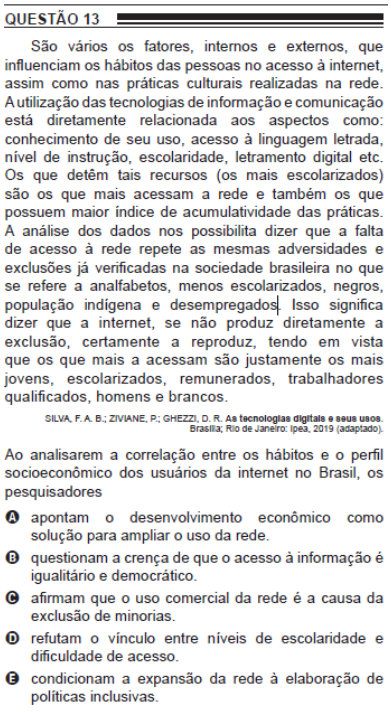
The information provided by the IRT scores might be helpful to control when estimating the gaps per question, as provide one way of include a proxy of ability. The latent ability estimation (parameter θ) is the result of the most consistent pattern across different set of question, calibrated in a sample 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 A19, 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.

Distortions and IRT interaction

One interesting thing that the data allows me to do is to check on what extent the gaps differ in

Appendix Figure A5: Some Items examples



QUESTÃO 13

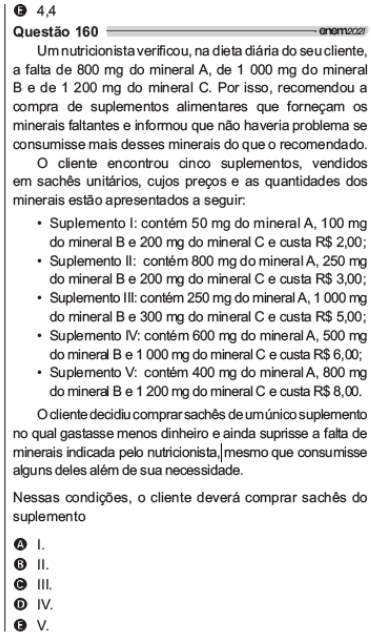
São vários os fatores, internos e externos, que influenciam os hábitos das pessoas no acesso à internet, assim como nas práticas culturais realizadas na rede. A utilização das tecnologias de informação e comunicação está diretamente relacionada aos aspectos como: conhecimento de seu uso, acesso à linguagem 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



D) 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:

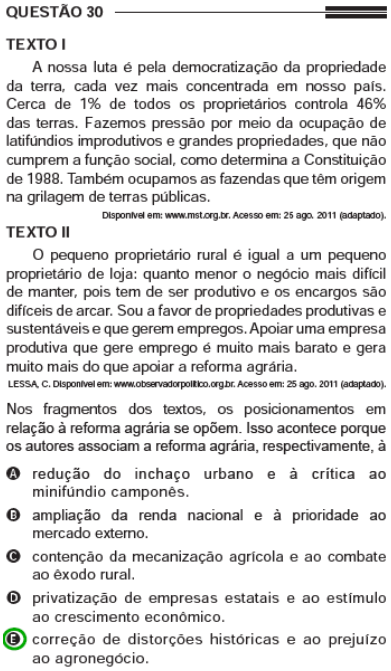
- Suplemento I: contém 50 mg do mineral A, 100 mg do mineral B e 200 mg do mineral C e custa R$ 2,00;
- Suplemento II: contém 800 mg do mineral A, 250 mg do mineral B e 200 mg do mineral C e custa R$ 3,00;
- Suplemento III: contém 250 mg do mineral A, 1 000 mg do mineral B e 300 mg do mineral C e custa 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

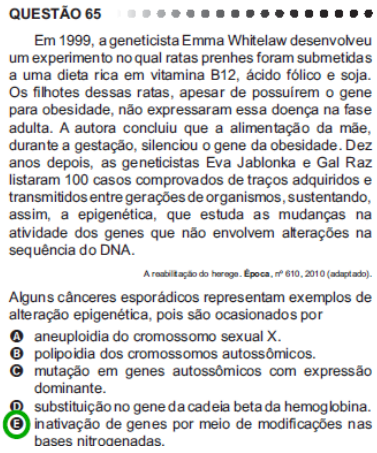
O pequeno proprietário rural é igual a um pequeno proprietário de loja: quanto menor o negócio mais difícil de manter, pois tem de ser produtivo e os encargos são difíceis de arcar. Sou a favor de propriedades produtivas e sustentáveis e que gerem empregos. Apoiar uma empresa produtiva que gere emprego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

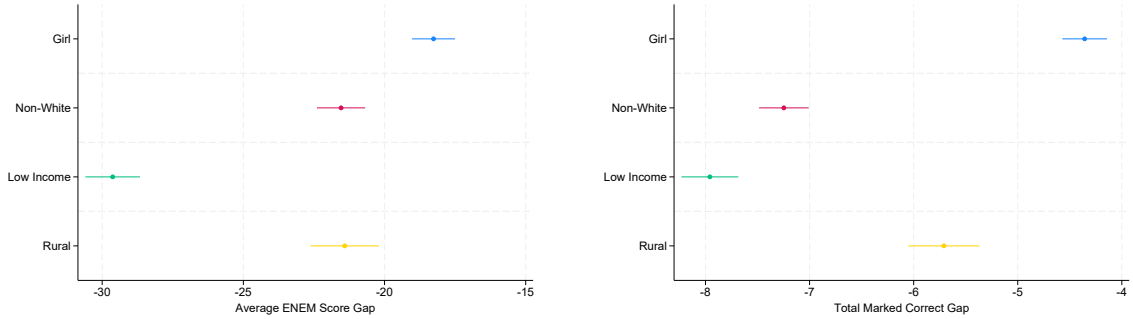
Alguns cânceres esporádicos representam exemplos de alteração epigenética, pois são ocasionados por

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.

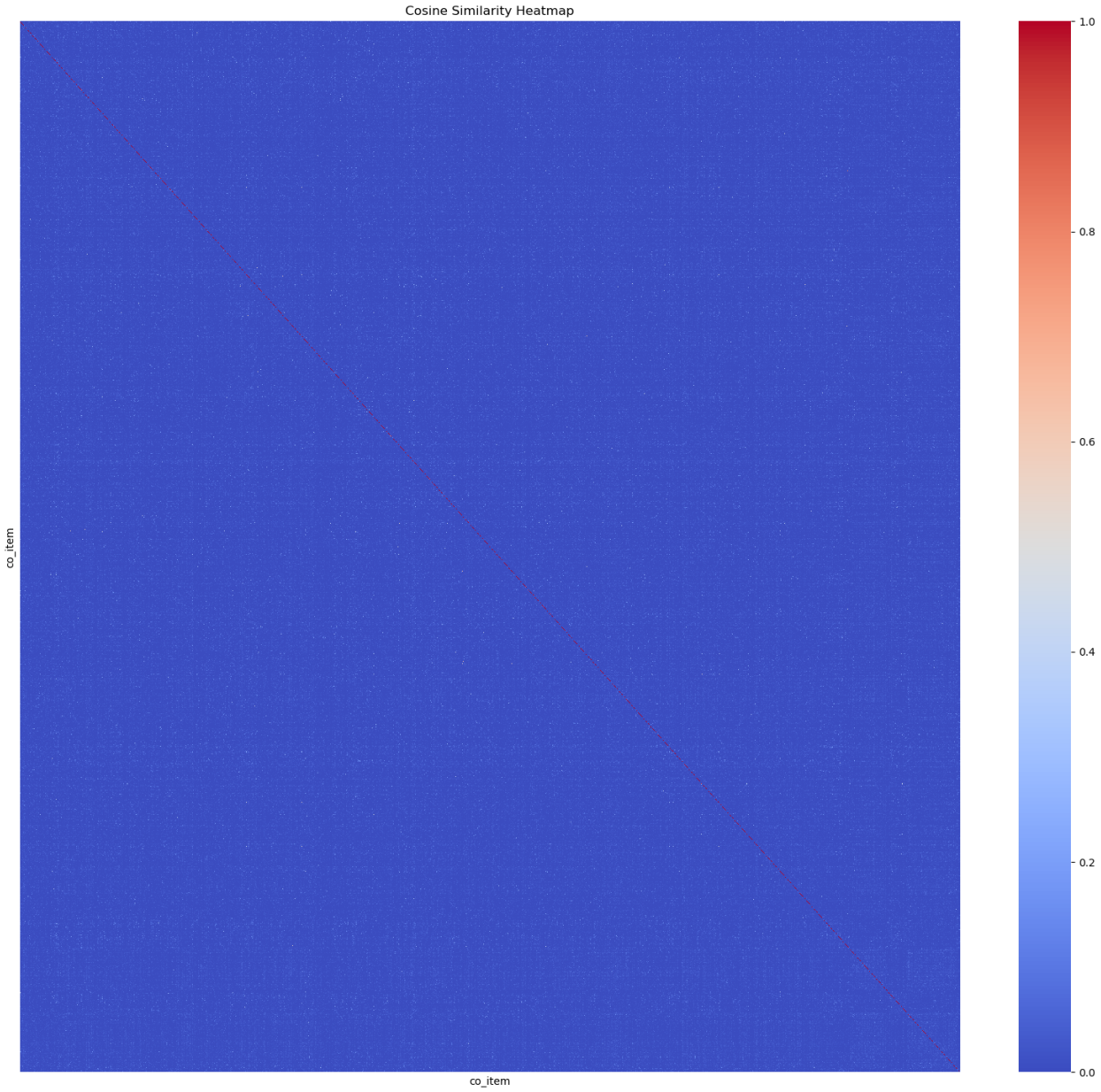
Appendix Figure A6: Words with the highest predicting power by type of performance gap



(a) Score Gap by Group

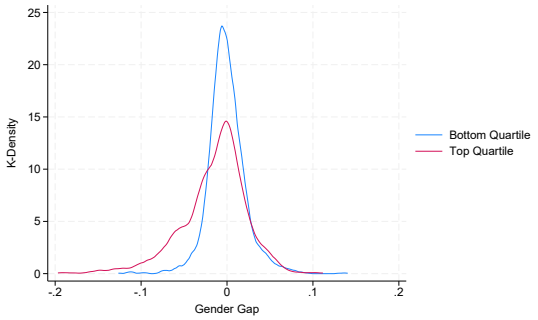
(b) Number of Correct Marking by Group

Note: These plots show the gaps observed in terms of total score on the ENEM by relevant group. The left panel shows the gap in terms of the score estimated using the IRT model. The right panel shows the gap in terms of the total number of items that were correctly identified by group.

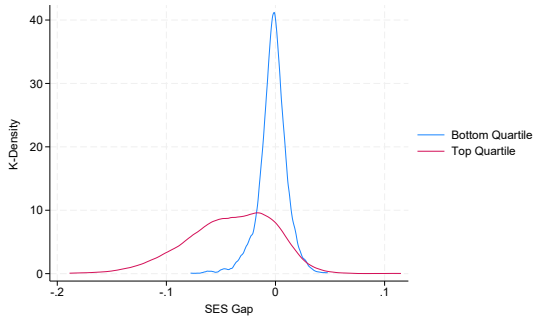


Appendix Figure A7: Cosine Similarity across items

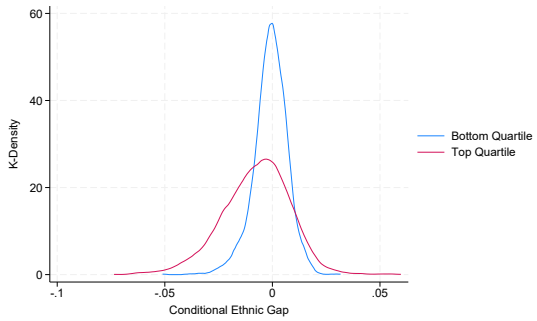
Appendix Figure A8: performance Gaps density by different tagging



(a) Gender Performance Gaps (p.p.) by ability quartile



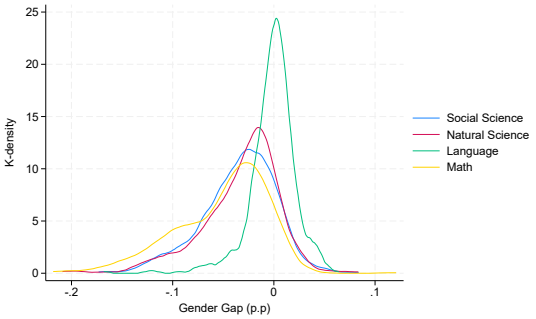
(b) SES Performance Gaps (p.p.) by ability quartile



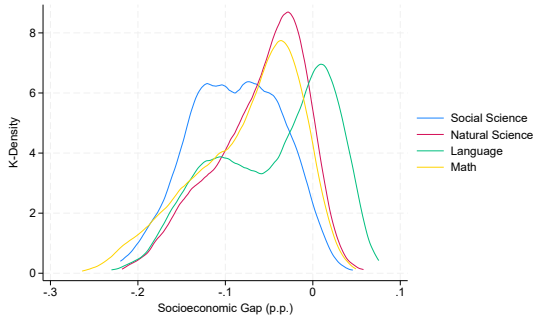
(c) K-Density of Ethnic Performance Gap conditional on LowSES (p.p) by ability quartile

Note: These plots show the k-density estimations for the distribution of gaps according to each category by ability quartile (first v/s fourth).

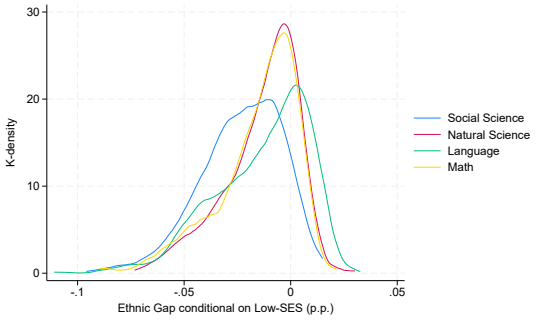
Appendix Figure A9: performance Gaps density by different tagging



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



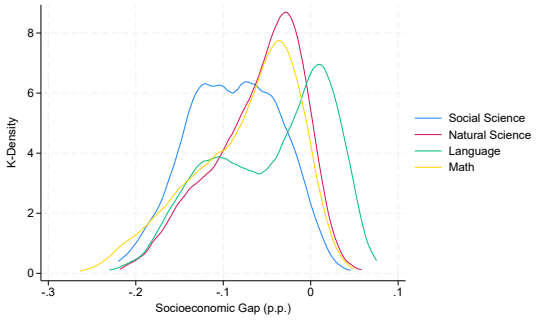
(b) K-Density of Socioeconomic Performance Gaps (p.p.)



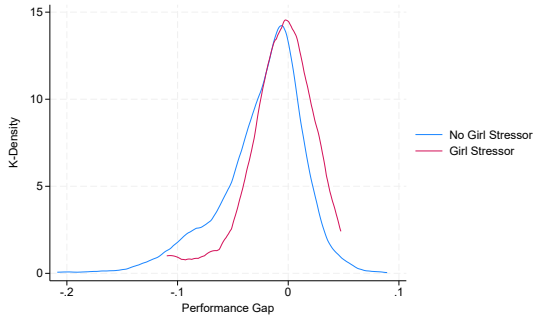
(c) K-Density of Ethnic Performance Gap conditional on LowSES (p.p)

Note: These plots show the k-density estimations for the distribution of gaps according to each category.

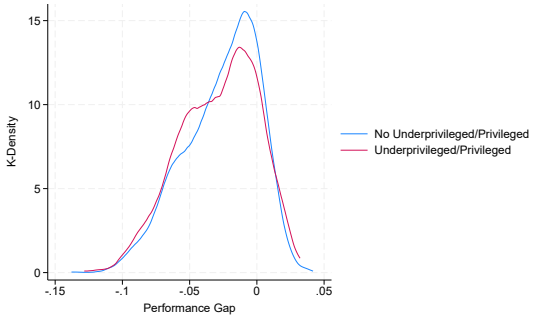
Appendix Figure A10: performance Gaps density by different tagging



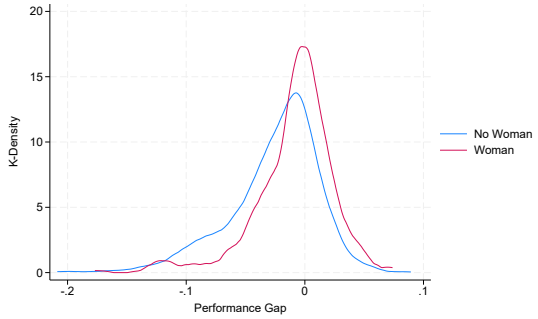
(a) K-Density by Poor Stressor



(b) K-Density by Girl Stressor



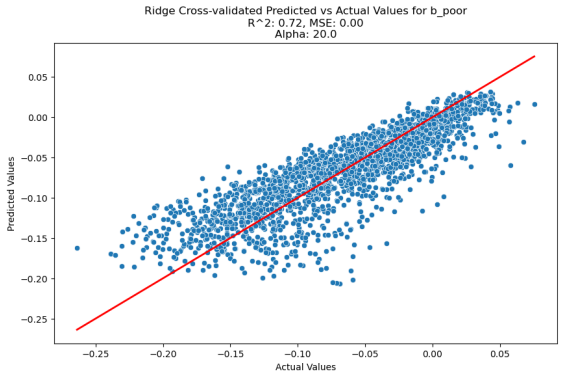
(c) K-Density by Girl Stressor



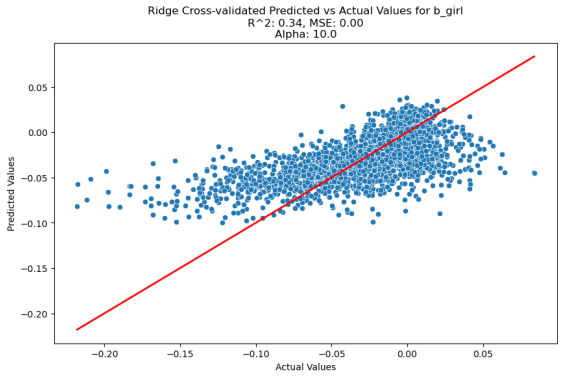
(d) K-Density by Girl Stressor

Note: These images show the words in the text that have the highest predictive power in explaining the observed gaps. Those in blue are words that influence the gaps by reducing them, while those in red widen the gap. The words were estimated using a Ridge model with an alpha value of 10, chosen by the method of grid search.Figure A11 shows the fit of the models.

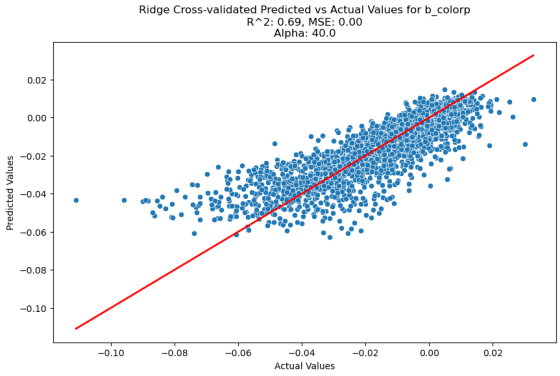
Appendix Figure A11: Words with the highest predicting power by type of performance gap



(a) SES gap

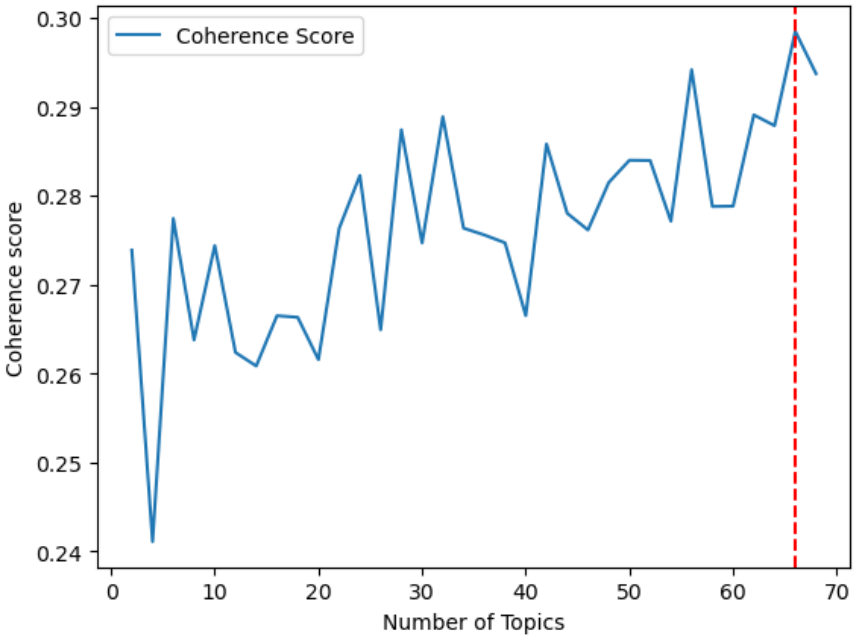


(b) Gender Gap



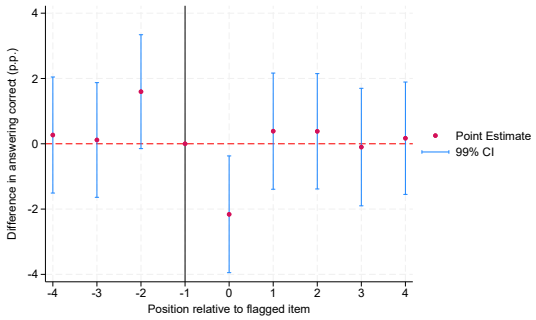
(c) Cond. Ethnic Gap

Note: These images show the words in the text that have the highest predictive power in explaining the observed gaps. Those in blue are words that influence the gaps by reducing them, while those in red widen the gap. The words were estimated using a Ridge model with an alpha defined by grid search and including the item features and competences fixed effects as controls. Figure A11 shows the fit of the models.

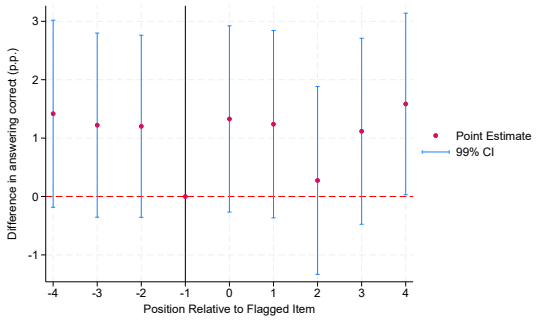


Appendix Figure A12

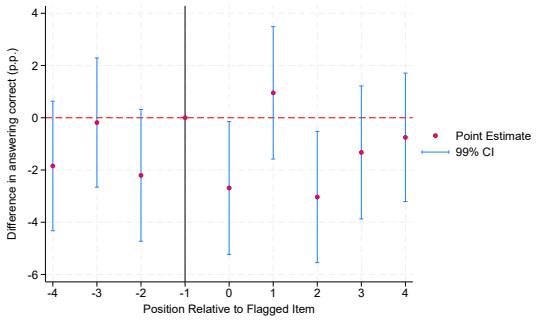
Appendix Figure A13: Performance Gaps dynamics by relative position to flagged item



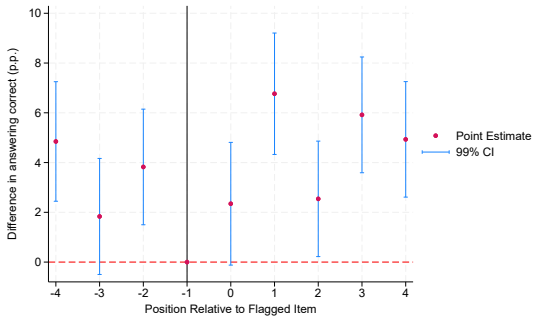
(a) SES relative gap in the presence of a flagged item with a widening effect



(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 when the equation 3 is executed. The time window considered is -4 items to +4 items, the one in the middle being the marked one. Note that the baseline is the intimidate item before.

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

Appendix Figure A14: Competences 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

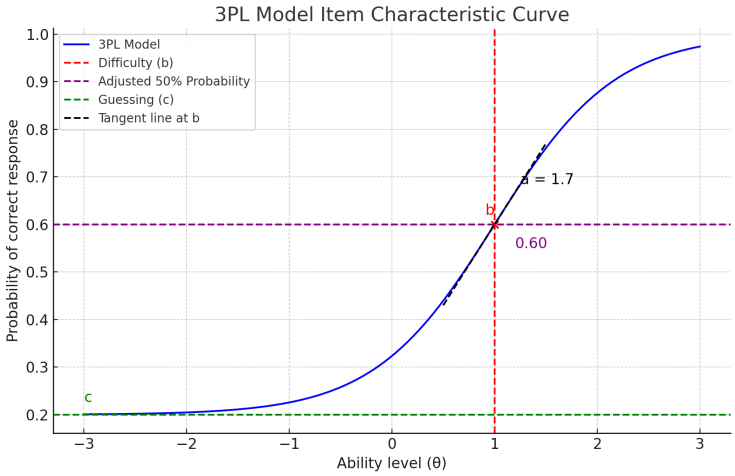
Appendix Figure A15: Competences descriptions for Math

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

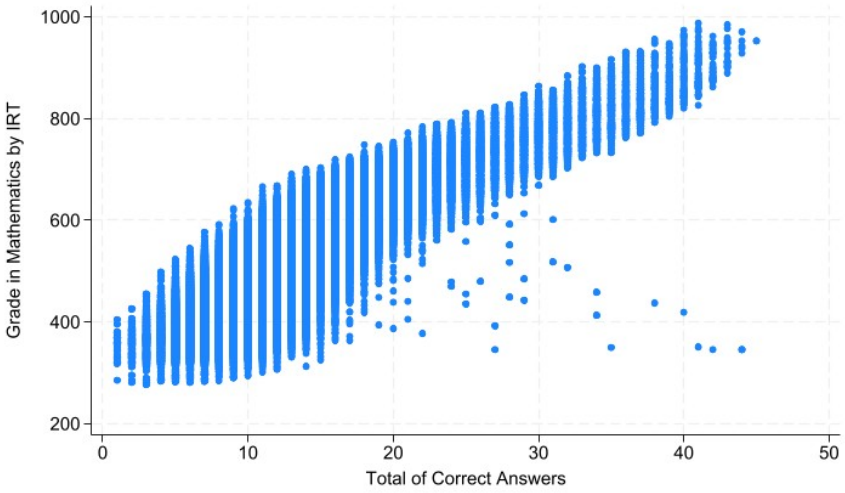
Appendix Figure A16: Competences descriptions for Science

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

Appendix Figure A17: Competences descriptions for Social Sciences



Appendix Figure A18: 3LMP Characteristic Curve- Mathematics Test



Appendix Figure A19: IRT v.s Actual Number of Corrects