Analyzing the Equity of the Brazilian National High School Exam by Validating the Item Response Theory's Invariance

Vitoria Guardieiro[1], Marcos M. Raimundo[1,2], Jorge Poco[1]

[1]Getulio Vargas Foundation [2]Universidade Federal do Rio de Janeiro

The Brazilian National High School Exam (ENEM)

- One of the most extensive entrance exams globally - over 5 million participants in 2019.
- It has several functions:
  - **Individual-level:** admission test to access the federal universities (through the Unified Selection System) and access to the federal scholarship programs (University for All Program).
  - **Collective level:** comparison between schools and municipalities, and it also serves as an indicator for public educational policies.
- Composed of four objective tests, each containing 45 multiple choice questions and an essay:
  - The tests evaluate the knowledge areas: Mathematics (MT), Languages and Codes (LC), Human Sciences (HS), and Natural Sciences (NS).
  - The LC test consists of five foreign language questions (English or Spanish), and the remaining 40 questions are in Portuguese.
  - Each participant must take all tests and choose one of the two foreign languages.

ENEM Score Estimation

The methodology of score estimation is from Item Response Theory (IRT), which models the probability of a participant responding correctly to an *item* (or *question*) as a function of its parameters and the participant's *ability* (or *proficiency*).

Three-parameter Logistic Function

ENEM uses the three-parameter logistic function, that is the probability of a correct answer by participant j to item i (event $U_{ij} = 1$) given the proficiency parameter θ_j and item parameters a_i, b_i, and c_i:

$$P(U_{ij} = 1|\theta_j, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{1}{1 + e^{-a_i(\theta_j - b_i)}}$$

Item Characteristic Curve

The relationship between $P(U_{ij} = 1|\theta_j)$ and the parameters a, b, and c is called the *Item Characteristic Curve* (ICC):

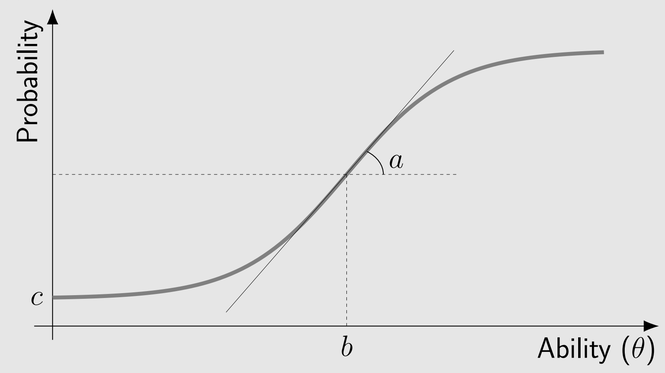

Figure 1. Example of an Item Characteristic Curve (ICC).

Invariance

Item Response Theory starts from the premise that, for a given question, a single function maps an examinee's ability to his probability of answering it correctly. If the model is well specified, all populations' parameters are the same.

This premise implies the property called *invariance of item and ability parameters*, which is the primary distinction between IRT and classical test theory. Therefore, we would expect that given two populations of participants, their estimated ICC should be similar, even if their ability distributions are different.

Research Goal

Analyze if the ENEM Score Estimation methodology is egalitarian for students of different population groups based on sensitive characteristics. To do so, we evaluate whether a critical premise of this methodology, namely **invariance**, holds for the subpopulations considered. If such a premise does not hold, we can conclude that the score estimation of the subpopulations is not egalitarian, even if their ability distributions are different.

Methodology

Data

We used ENEM's most recent microdata from 2019. We analyzed three group characteristics reported by the participant: Gender (Man or Woman), Income (Low, Medium, and High income), and Race (White, Black, and Brown).

AUICC Inspection

Under invariance, we expect the ICC of an item to be similar to all groups. Therefore, the area under ICC (AUCICC) should be near equal to all groups. We calculated how different the observed AUICC is for each group of a specific characteristic with the following steps:

1. **Observed ICC:** For each question and subpopulation; the observed ICC was the proportion of participants who answered correctly given their score range.
2. **Item AUICC:** Area under the item characteristic curve for each subpopulation and the total population.
3. **AUICC discrepancy:** Standard deviation of AUICC for the groups and normalized this value by dividing by the AUICC found for the total population. This normalized value indicates whether inequality seems to hold for this item (small values) or not (bigger values).

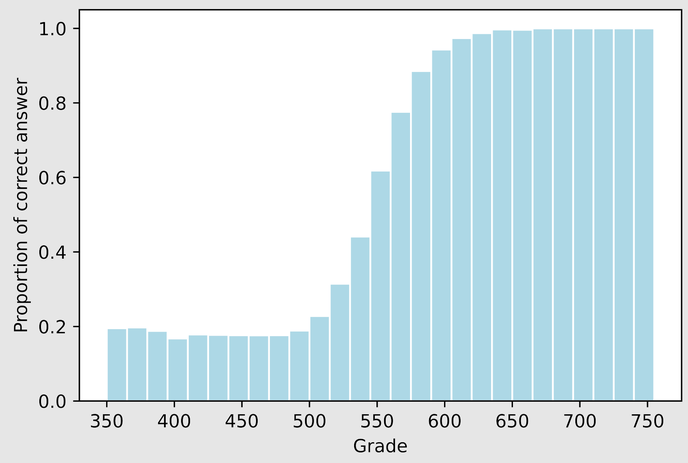

Figure 2. Illustration of the Observed ICC for a given question. The light blue bars represent the proportion of correct answers the participants gave in a grade range. The AUICC is the sum of the area of the bars.

Invariance Checking

The item-by-item AUICC comparison indicates potentially troubling questions. To show if there is a consistent difference among the social/gender groups, we performed a non-parametric Friedman ranking test.

We determine the one-by-one comparisons through a Finner posthoc test for the combination of tests and features whose null hypothesis was rejected by the Friedman test.

Results

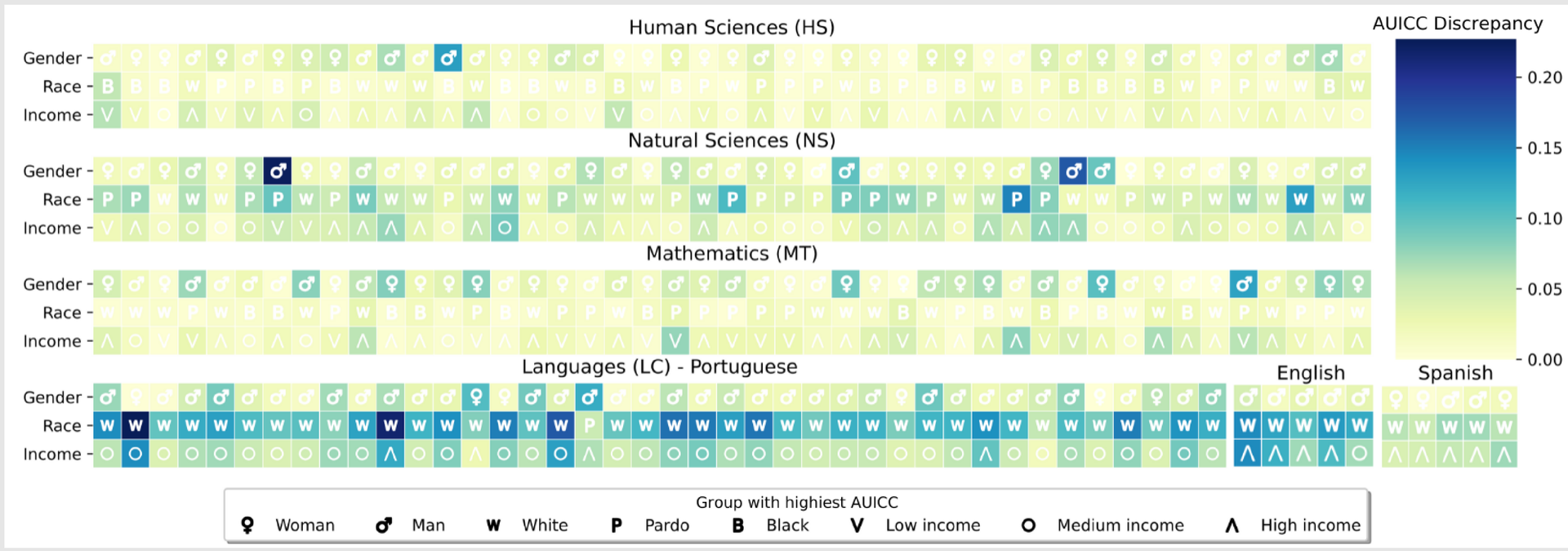

Figure 3. Heatmap of the Area discrepancy based on the Observed ICC. The color of each square indicates the value of Area discrepancy and the symbol indicates the group with the highest area. As the importance of the indicated grows with the discrepancy, the symbols' visibility also grows with the color.

AUICC Inspection

Visually in Figure 3, this result indicates that:

- The first three tests are egalitarian for gender, race, and income. Although a few questions have high difficulty discrepancy, the group with the highest area in such questions varies significantly.
- The LC test, mainly the foreign language questions, is not egalitarian for race and income, with white having higher AUICC in all of the LC questions and the high-income group dominating the foreign language questions, mainly in English.
- Men have greater AUICC than women in almost all of the LC questions, but the discrepancy is low.

Invariance Checking

- The HS and MT tests did not show a consistent favoring or disfavoring for any group.
- For Gender, the LC questions in Portuguese and English showed a relevant difference (*women* having lower AUICC rank).
- For Race, the NS test (*black* having lower rank than *white* and *pardo*), LC Portuguese questions (*black* and *pardo* having lower rank than *white*), and both foreign languages questions (*black* having lower rank than *white*) were not egalitarian.
- For Income, the *low-income* group ranked lower than the *high-income* for NS and also LC in all languages.