FACULTAD DE INGENIERÍA

REVISTA DE LA FACULTAD DE INGENIERIA

www.revistaingenieria.uda.cl

Fraud detection in exams with polytomous response items

Oilson Alberto Gonzatto Junior[1], Marcos Jardel Henriques[2], Claudia Evelyn Escobar Montecino[2], Josimara Tatiane da Silva[2], Vanderly Janeiro[3], and Terezinha Aparecida Guedes[3]

[1]Institute of Mathematical and Computer Sciences, University of São Paulo (ICMC-USP), São Carlos, São Paulo, Brazil.
[2]Interinstitutional Postgraduate Program in Statistics (PIPGEs) UFSCar-USP (Federal University of São Carlos (DES-UFSCar) and University of São Paulo (ICMC-USP)). São Carlos, São Paulo, Brazil. Email: jardel@usp.br
[3]Postgraduate Program in Biostatistics, Department of Statistics, State University of Maringá (UEM). Maringá, Paraná, Brazil

The database addressed in this work results from the administration of a test to over 11,000 individuals. The test comprised 50 multiple-choice questions, each with five mutually exclusive options. Among the evaluated individuals, four small groups became suspects of cheating on the exam. As exploratory analysis tools, among other methods, the characteristic curves graph, obtained with a Nominal Response Model, and the dendrogram, which was useful for determining degrees of similarity between the answer sheets, were used. The probabilistic analysis included the approach via Classical Test Theory, using the indices: K, K_1, K_2, S_1, and S_2, and also a modern approach based on Item Response Theory, in this case, using the indices: GBT and ω. From the probabilistic analysis of the Copy Indices, it can be concluded that the occurrence of fraud among the individuals in the suspect groups is highly probable, justifying a possible detailed investigation of their members.

Keywords: Copy Indices. Item Response Theory (IRT). Nominal Response Model.

La base de datos abordada en este trabajo es el resultado de la aplicación de una prueba a más de 11 mil individuos. La prueba estuvo compuesta por 50 preguntas de opción múltiple, cada una con cinco opciones mutuamente exclusivas. Entre los individuos evaluados, 4 pequeños grupos se convirtieron en sospechosos de haber cometido fraude en el examen. Como herramientas de análisis exploratorio, entre otros métodos, se utilizaron el gráfico de las curvas características, obtenido con un Modelo de Respuesta Nominal, y el dendrograma, que fue útil para determinar grados de similitud entre las hojas de respuestas. El análisis probabilístico contó con el enfoque de la Teoría Clásica de Tests, utilizando los índices: K, K_1, K_2, S_1 y S_2, y también con un enfoque moderno basado en la Teoría de Respuesta al Ítem, en este caso, utilizando los índices: GBT y ω. Del análisis probabilístico de los Índices de Copia se puede concluir que la ocurrencia de fraude entre los individuos de los grupos sospechosos es muy probable, lo que justifica una posible investigación detallada de sus miembros.

Palabras claves: Índices de Copia. Teoría de Respuesta al Ítem (TRI). Modelo de Respuesta Nominal.

1 Introduction

Item Response Theory - IRT can assist one in these decision-making processes by considering categorical responses as latent, that is, calculating how likely each category is to be true. It is through methodologies like these that the average test score is reached for all possible results as a probabilistic score. This allows one to unravel patterns such as measuring how likely that response is to be true.

According to Wollack (1997) [11], there are two basic statistical approaches to detecting copying among individuals taking a multiple-choice test. One approach involves comparing the similarity between the responses of two individuals, considering an expected amount of similarity; the other is based on Item Response Theory, whose updated description can be found in the texts by [7, 8, 9].

The first approach, as outlined by [13], uses evidence found in the unusual clustering of responses on two answer sheets. The characterization of ”evidence” occurs in two ways: the exclusive use of incorrectly identical responses and the exclusive use of both correctly and incorrectly identical responses. However, according to [13], these interpretations have been legally criticized, as identical responses may be considered evidence of copying, the opposite should attest to ”non-copying”. A more in-depth discussion of this can be found in the work of [2].

Other perspectives take into account the information contained in all the items answered by all the examined subjects. Among these methods are those based on Item Response Theory, more specifically, the Nominal Response Models proposed by [1]. This approach considers that the probability of an examined individual correctly answering a given item (considering an estimate of their latent ability) is independent of any other examined subject. In this context, the response pattern of an examined individual is compared with what is expected to be observed from other individuals with similar latent abilities.

The context discussed here involves the administration of a test to 11,234 individuals. The test consisted of 50 multiple-choice items, each with five response options. In this sense, the general considerations for the analysis assume that:

- a candidate only gets a question wrong when they truly do not know how to solve it;
- if there is no fraud, a candidate who does not know how to solve a question randomly chooses one of the five response options;
- if there is fraud, a candidate who does not know how to solve a question chooses one of the five response options non-randomly.

Among the classified individuals, there are four main small groups suspected of violating the aforementioned randomness assumptions. Considering possible fraudulent practices, the collection of evidence aims to argue against or in favor of this suspicion.

2 The methods

The analysis of the overall response pattern was conducted considering only individuals who filled out the entire exam answer sheet without any erasures; in this case, $10,769$ individuals remained whose observed responses were included in the analyses. Although not detailed here, appeals made to the assessment organization were also considered to relate and identify inconsistencies between empirical observations of the descriptions made and the resources submitted for each item.

Among the exploratory analyses used, the following stands out: the characteristic curve plot 2, obtained with a Nominal Response Model, which helps to perceive response patterns according to the degree of latent ability of the examined subjects (as well as bar graphs 3); and the dendrogram 1, useful for determining degrees of similarity between answer sheets, this method was important for identifying groups that, due to excessive proximity, could be classified as suspicious groups.

To group evidence of “copy” or “non-copy” among the $10,769$ individuals' responses to the 50 questions, the indices GBT, ω, K, K_1, K_2, S_1, and S_2 were used. For their respective calculations, all codes are implemented in the `CopyDetect` package, created by [12], and available in the `R` software [4].

- **Generalized Binomial Test – GBT:**

  The GBT index was developed by [10] based on the derivation of the exact probability distribution of the number of identical responses between two answer sheets (considering a known model for polytomous responses). The individual probabilities, P_{jic} and P_{jis}, of choosing alternative j on item i, for the copier (c) and copied (s), respectively, are determined based on the Nominal Response Model. Considering the probability of observing exactly m coincidences out of n items, one can determine the probability of observing at least O_{cs} identical responses and then compare it to a critical value of interest.

- **Index ω:**

  Proposed by [11], it combines Classical Test Theory (used in indices of the first approach) with Item Response Theory and was specifically designed to detect copies. The probabilities associated with each possible response are also obtained through the Nominal Response Model. Its test statistic is a normal approximation of the exact probability distribution for the number of identical responses between two response vectors.

- **Index K and its variants:**

  Indices K, K_1, and K_2 are based on the Binomial distribution to determine the probability of observing at least W_{cs} identically incorrect responses between two sets of responses. The calculation is based on a random variable M representing the number of items equally incorrect as those of the copied individual for each group of examined subjects sharing the same number of incorrect items. Since this random variable can be approximated by a Binomial distribution, it is necessary to determine the parameter representing the proportion of individuals that determine it; in this case, the indices K, K_1, and K_2 were defined using distinct approaches:

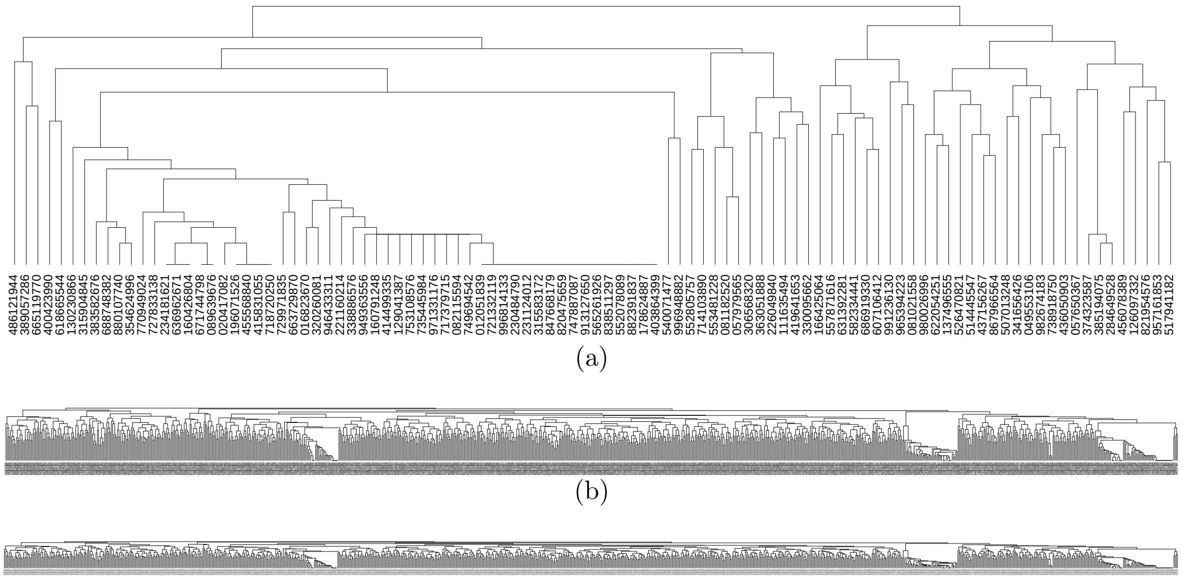
  - **Index K:** The Binomial proportion is empirically determined, the first approximation, proposed by [3], is based on the observed average value of M and then determines the proportion of incorrect responses in the population.
  - **Index K_1:** The Binomial proportion is determined through a linear approximation, using the proportion of incorrect responses of each examinee in each subgroup sharing the same number of incorrect responses.
  - **Index K_2:** The Binomial proportion is determined through a quadratic approximation, using the proportion of incorrect responses of each examinee in each subgroup sharing the same number of incorrect responses.

  In summary, index K corresponds to the proportion of examined individuals who have the same number of incorrect responses as the copier and whose number of identically incorrect items coinciding with the copied is at least as large as W_{cs}. Thus, when K is a very small quantity, there is statistical evidence that the copier indeed copied the copied individual's exam. Further considerations regarding index K are available in the works of [3, 5, 6].

  Indices S_1 and S_2, proposed by [6], originate from the same idea that characterizes index K, i.e., considering only errors as evidence, however, they use the Poisson distribution to understand the quantity W_{cs} and employ a log-linear model to estimate M. Index S_2 was proposed to overcome a limitation of index K (and its variants), for this purpose, its construction also incorporates information from identically correct responses.

The determination of copy indices was performed after identifying suspicious groups (GS) through descriptive and visual processes. Once individuals from each group were identified, all possible pairs (copier and copied) were considered and their copy indices calculated. The process was carried out in this manner because there is a distinction between the indices obtained when investigating pairs (A;B) and (B;A), and it is not possible to identify a "suspect individual" for copying or being copied.

As a control measure, two control groups (GC) were selected, which corresponded to random samples of $2,200$ individuals among those who took the test. In their case, it is impractical to identify all possible pairs, so $10,000$ pairs were randomly selected for determining copy indices.



(a)

(b)

(c)

Figure 1: Dendrograms for the top 100, 500, and 1000 individuals based on their scores.

Source: the authors

3 Results

The degree of similarity between sets of individual responses was organized in the form of a dendrogram for a more coherent visualization of proximity patterns among responses and identification of groups of suspicious individuals. For illustration, observe in Figure 1 the patterns exposed in the dendrograms for the top 100, 500, and 1000 candidates based on their scores. Under normal circumstances in this context, what is expected in a figure like this is the pattern perceived on the right side of Figure 1a, or the left side of Figure 1b, and for the most part of Figure 1c.

It is visually impractical to reproduce a dendrogram in this text that represents all candidates; however, it was constructed, and, given a similarity threshold, the six groups with the highest frequency ($GS1 - -GS6$) were extracted. From the sixth group onwards, the frequency of individuals with very similar response sets is four or fewer. Once identified, the most frequent groups that share a high enough degree of similarity to be considered suspicious, the analysis proceeds probabilistically. Similarly, one can also understand and conclude about such questions, through the observable frequencies and probabilities presented by Figures 2 and 3.

In the graphs of these two figures, one can try to notice, for example, candidates who made good guesses when answering the alternatives of each question. Or those who visibly and/or possibly cheated on the exam. For example, in the National High School Exam - ENEM (a higher education admission test conducted by the National Institute of Educational Studies and Research Anísio Teixeira, an autonomous agency linked to the Ministry of Education of Brazil), one can see how such characteristic curves and even IRT. It is through methodologies allows attribute the differentiation between students who guessed and those who genuinely got questions wrong due to lack of knowledge is made. In other words, two students taking the same test and getting the same number of questions right (half). Assuming the first student got half of the difficult questions right, and the second student got half of the easy questions right. In this case, the second student will end up with a much higher score than the first. This is because, initially, it is concluded that: how does a student get all the difficult questions right and all the easy ones wrong? Interpretations like this, among many others, make us at least want to look more closely at what happened with these candidates. Graphs like those in Figures 2 and 3 allow us to see strong indications on this subject.

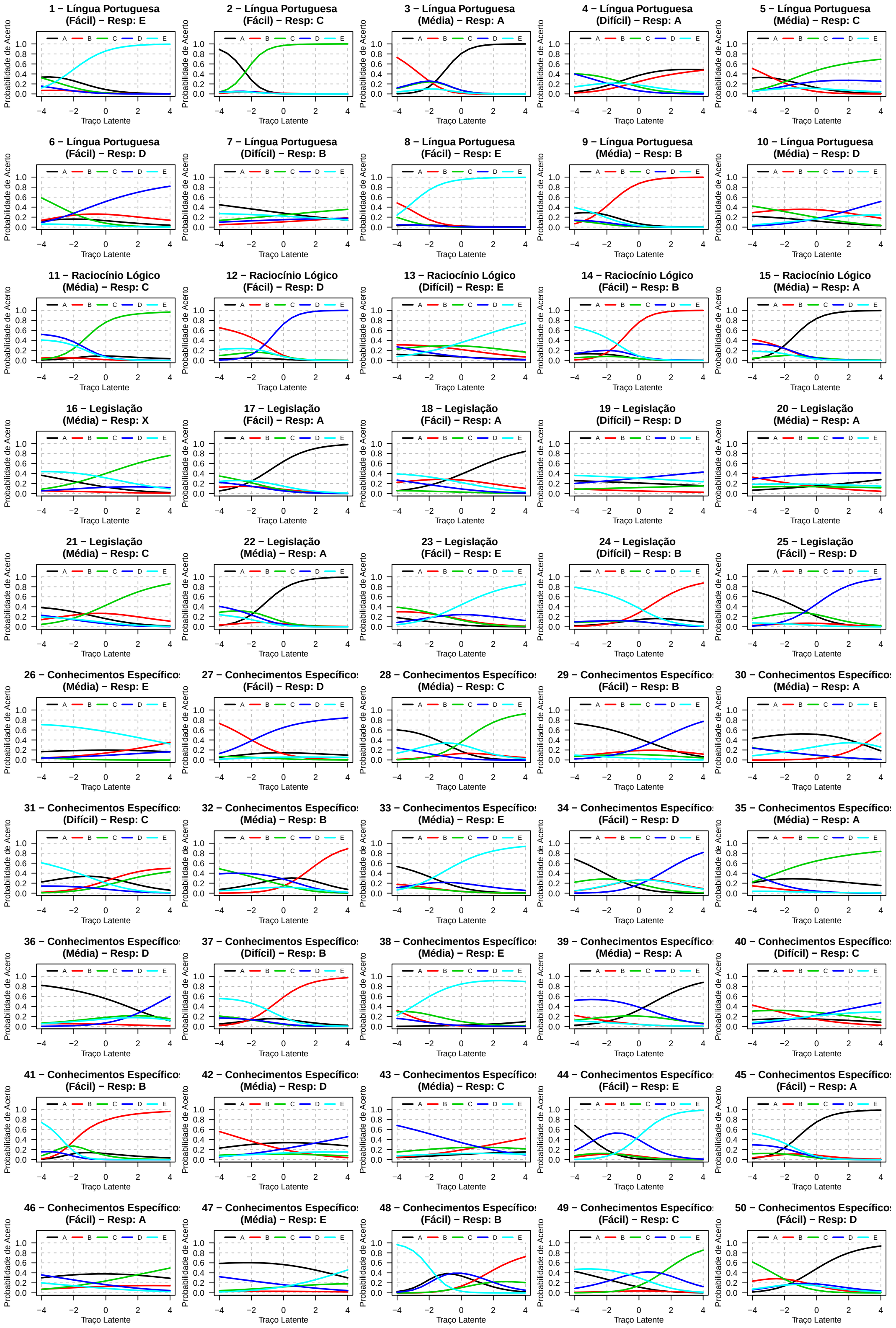


Figure 2: Characteristic curves - of the candidates' responses

Source: the authors

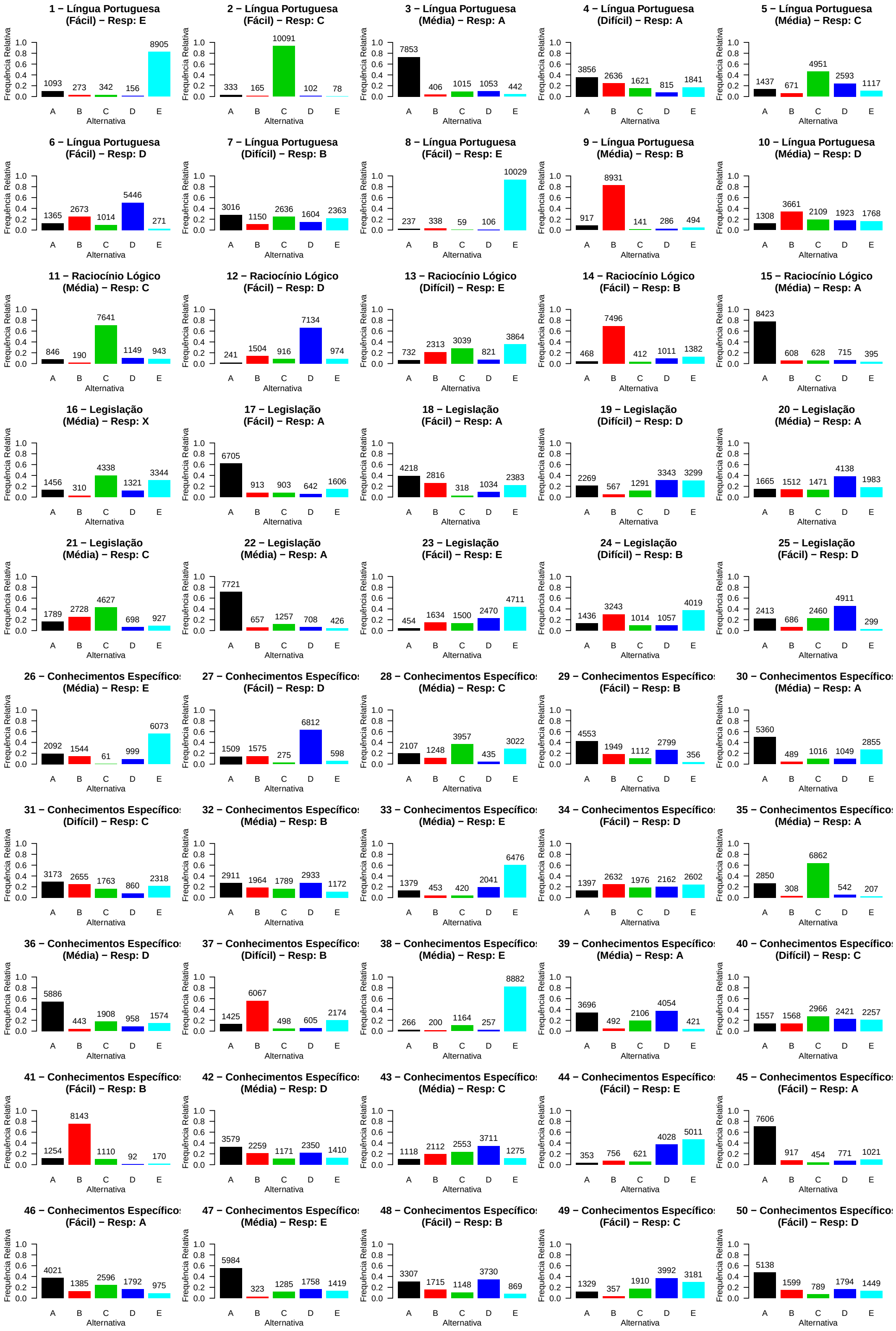


Figure 3: Bar graphs - of the candidates' responses

Source: the authors

Table 1: Observed relative frequencies (in percentage) of detection of copying per examined group.

Group Analyzed	Copy Detection Indices						
	S_2	S_1	K_2	K_1	K	GBT	ω
GS1	0,00	0,02	0,12	0,43	0,79	98,85	99,88
GS2	84,35	89,89	97,62	97,44	45,48	99,77	100,00
GS3	99,50	99,66	99,83	99,83	93,20	95,38	99,92
GS4	95,67	95,46	98,69	99,29	89,72	93,15	98,08
GS5	0,00	0,00	0,00	2,38	0,00	23,81	28,57
GS6	0,00	4,76	4,76	9,52	4,76	19,05	28,57
GC1	0,16	0,27	0,29	0,78	0,25	0,85	1,23
GC2	0,12	0,24	0,20	0,59	0,20	0,47	1,12

Source: the authors

The results presented here correspond to the observed relative frequency of evidence of copying for each of the suspicious groups and the two control groups. Establishing a 1% level of false detections, in random and independent situations (such as with the two control groups), one would expect to observe an approximate 1% frequency of evidence of copying. A significantly higher frequency across all indices would indicate evidence that individuals in that group cheated on the exam.

According to the results shown in Table 1, it is clear that there is sufficient evidence to further investigate individuals in suspicious groups 2, 3, and 4, as they show evidence of copying across all indices. Suspicious group 1 was not identified by the first four indices, likely due to low frequency (its individuals are part of the highest scoring group in the exam).

4 Discussion

From the probabilistic analysis of the Copy Indices, it can be concluded that there is a high likelihood of fraud among the individuals in each of the first four suspicious groups, justifying a potential detailed investigation of their members. In addition to the information presented here, all other analyzed data indicate an unexpected clustering in the proximity between answer sheets in groups 1 to 4.

Furthermore, although not graphically presented here, the characteristic curves indicated that the latent ability, especially of $GS1$, justifies its error patterns; that is, the errors made by these individuals correspond to the errors that individuals with higher latent traits would make. This supports the subjects in the sense that there would be no abnormality in the errors, as they are individuals "better prepared". On the other hand, this may also indicate that the responses came from a common source.

References

[1] R. Darrell Bock. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 03 1972.

[2] W. G. Buss and M. R. Novick. The detection of cheating on standardized tests: Statistical and legal analysis. *Journal of Law and Education*, 9:1–64, 1980.

[3] Paul W. Holland. *Assessing Unusual Agreement Between the Incorrect Answers of Two Examinees Using the K-Index: Statistical Theory and Empirical Support.* Educational Testing Service, Princeton, New Jersey, 1996.

[4] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2016.

[5] Leonardo S. Sotaridona and Rob R. Meijer. Statistical properties of the k-index for detecting answer copying. *Journal of Educational Measurement*, pages 115–132, 2002.

[6] Leonardo S. Sotaridona and Rob R. Meijer. Two new statistics to detect answer copying. *Journal of Educational Measurement*, 40, 2003.

[7] Wim J. van der Linden. *Handbook of Item Response Theory, Volume One.* Taylor & Francis Ltd, New York, EUA, 2016.

[8] Wim J. van der Linden. *Handbook of Item Response Theory, Volume Two.* Taylor & Francis Inc, New York, EUA, 2016.

[9] Wim J. van der Linden. *Handbook of Item Response Theory, Volume Three.* Taylor & Francis Inc, New York, EUA, 2018.

[10] Wim J. van der Linden and L. Sotaridona. Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics*, 31, 01 2006.

[11] J. A. Wollack. A nominal response model approach for detecting answer copying. *Applied Psychological Measurement*, 21, 12 1997.

[12] Cengiz Zopluoglu. *CopyDetect: Computing Statistical Indices to Detect Answer Copying on Multiple-Choice Tests*, 2016. R package version 1.2.

[13] E. C. Zopluoglu, C.; Davenport. The empirical power and type i error rates of the GBT and ω indices in detecting answer copying on multiple-choice tests. *Educational and Psychological Measurement*, 72, 12 2012.