RIDU Revista Digital de Investigación en Docencia Universitaria
E-ISSN: 2223-2516

V. 14, n°2, jul-dec | PERÚ | 2020
https://doi.org/10.19083/ridu.2020.1240

THEORETICAL ARTICLE

How to assure valid assessments and detect cheating in synchronous remote tests

Pilar Rodríguez Morales
Universidad de la República, Maldonado, Uruguay.
https://orcid.org/0000-0003-1929-4961

Mario Luzardo Verde
Universidad de la República, Montevideo, Uruguay.
https://orcid.org/0000-0002-9360-2806

Received: 28/07/20 *Revised: 15/10/20* *Accepted: 27/10/20* *Published: 12-11-2020*

Abstract

The passage from face to face teaching to distance learning, as a measure to face the Covid-19, brought about the need for the validation of the tests' results taken in electronic format. It is considered that students are more likely to commit fraud in tests conducted remotely. The objective of the article is to present the study of the academic cheating as a report to the analysis of the psychometric validity of the tests. Through a bibliographic review, the concept of academic cheating and its types are analyzed. The main methods to detect it are presented, which can be used to ensure the validity of the results in synchronous tests, multiple choice type. The uses, potentialities and limitations of the methods used are described. Finally, the main challenges to overcome to validate the synchronous test' results carried out remotely are presented.

Keywords: Student cheating; student evaluation; testing; methodology.

Cómo asegurar evaluaciones válidas y detectar falseamiento en pruebas a distancia síncronas

Resumen

El pasaje de la enseñanza presencial a la modalidad a distancia, como medida para enfrentar el Covid-19, trajo como consecuencia la necesidad de la validación de los resultados de las pruebas tomadas en formato electrónico. Se considera que los estudiantes tienen mayor facilidad para cometer fraudes en pruebas realizadas a distancia. El objetivo del artículo es presentar el estudio del falseamiento como un aporte al análisis de la validez psicométrica de las pruebas. A través de una revisión bibliográfica se analiza el concepto de falseamiento y sus tipos. Se presentan los principales métodos para detectarlo, que pueden ser utilizados para asegurar la validez de los resultados en las pruebas síncronas, tipo opción múltiple. Se describen los usos, potencialidades y limitaciones de los métodos presentados. Por último, se plantean los principales desafíos por superar para la validación de los resultados de pruebas síncronas realizadas a distancia.

Palabras clave: Fraude académico; evaluación del estudiante; prueba; metodología.

Autor corresponsal: *Pilar Rodríguez Morales*
Mail: *pilarod@gmail.com*

Como garantir avaliações válidas e detectar falsidade em testes remotos síncronos

Resumo

A passagem do ensino para o a distância, como medida de enfrentamento ao Covid-19, trouxe como conseqüência a necessidade de validar os resultados dos testes realizados em formato eletrônico. Os alunos são considerados mais propensos a cometer fraudes em testes realizados remotamente. O objetivo do artigo é apresentar o estudo da falsificação como um relatório para a análise da validade psicométrica dos testes. Por meio de uma revisão bibliográfica, analisa-se o conceito de falsificação e seus tipos. São apresentados os principais métodos de detecção, que podem ser utilizados para garantir a validade dos resultados em testes síncronos, tipo múltipla escolha. São descritos os usos, potencialidades e limitações dos métodos utilizados. Por fim, são apresentados os principais desafios a serem superados para a validação dos resultados dos testes síncronos realizados remotamente.

Palavras-chave: Fraude acadêmica; avaliação do aluno; teste; metodologia.

How to cite this article:

Rodríguez, P. & Luzardo, M. (2020). How to assure valid assessments and detect cheating in synchronous remote tests. *Revista Digital de Investigación en Docencia Universitaria, 14*(2), e1240. https://doi.org/10.19083/ridu.2020.1240

Introduction

Universities faced multiple challenges to continue providing distance education after the suspension of face-to-face classes due to Covid-19. One of the biggest challenges has been the evaluation of learning in the context of virtuality. In Europe, universities were the first to take measures on remote evaluation. Then, other institutions did the same and provided guidance. In Spain, the Conference of Spanish University Rectors offers guidelines on how to carry them out (Conferencia de Rectores de Universidades Españolas [CRUE], 2020). In the United Kingdom, the Quality Assurance Agency, in its document called *Assessing with Integrity in Digital Delivery*, urges universities to work to prevent cheating or falsehood, plagiarism, or other improper behavior (Quality Assurance Agency for Higher Education [QAA], 2020).

In Latin America, most universities have implemented distance learning through virtual platforms. The rectors of Latin American leading universities have identified the lack of instruments for evaluation or accreditation of knowledge in virtual education as a weakness (Inter-American Development Bank [IDB], 2020).

In Uruguay, the declaration of a health emergency due to the detection of the first Covid-19 cases lead Universidad de la República (Udelar), the country's main university, to immediately suspend face-to-face classes and adopt the distance learning method. The rector's office provided general guidelines and, through the main university bodies, provided support and technical guidance. The issue of evaluation for the certification of student learning was initially raised, and an adequate design and programming of the tests on the platform was recommended to reduce copying problems (Universidad de la República/ Comisión Sectorial de Enseñanza [Udelar/CSE] (2020a). Later, due to the teachers' concerns about remote evaluation tools and methods, concrete guidelines were provided to ensure the quality of the instruments and control mechanisms that can be implemented (Udelar/CSE, 2020b).

Besides, before the pandemic, Udelar was taking its first steps in distance education as a tool to address large-enrollment courses, which are characteristic of this open-admission university, and also allow access to higher education throughout the country. Although

various technological tools were being used for virtual teaching, no progress had been made in remote evaluation for knowledge accreditation. The context of the pandemic accelerates the implementation of distance learning and the maintenance of health measures requires the use of remote assessments in multiple-choice format, especially in the schools with the highest enrollment per cohort (Udelar/ CSE, 2020b).

For these reasons, this article addresses the latest theoretical and methodological advances to ensure the validity of the results in synchronous remote tests.

The concept of fraud or falsehood is explored through a literature review, to then focus on methods to detect falsehood in multiple-choice tests.

First, the theoretical background on the subject and the psychometric approach by which fraud or falsehood should be studied are presented. Then, the main methods used to detect it will be described, including the latest developments. Finally, the challenges that, in the authors' opinion, higher education institutions face for the validation of the results of synchronous tests carried out remotely will be described.

Background

In this section, the concepts of fraud or falsehood, its various types, and the problem of validating test results will be discussed.

Test or exam fraud, or what is commonly referred to as "cheating," is an inherent problem with assessments. It has been extensively studied and, unfortunately, it has been proven that the scope of this practice has had a certain magnitude of consideration (Whitley, 1998; Arthur, Glaze, Villado, & Taylor, 2010).

While the analysis of data on test fraud began in the 1990s, there has been an interest in detecting it for almost a hundred years (Bird, 1927, 1929). Toward the end of the twentieth century, Whitley (1998) found, in a review of studies, that 43% of university students admitted to cheating on tests. In a study carried out in Spain, half of the surveyed students declared to have cheated at least once during an exam (Sureda, Comas, & Gili, 2009). A third of the fraud acts found by Friedman, Blau, & Eshet-Alkalai (2016) was carried out using technology. Although students seem to be more attracted to committing fraud on unmonitored online tests, the results are not as encouraging for those who commit fraud because their performance correlates negatively with the other courses in which their assessments are monitored (Arnold, 2016).

Technically, we will use the term falsehood for different types of fraud in tests. Cizek (2012) defines falsehood as any action taken before, during, or after the administration of a test or task, with the intent to take unfair advantage or produce inconsistent results. Distance education has focused on this issue since both teachers and students consider that falsehood in a test taken at distance is much easier than in a face-to-face paper-based test (Arnold, 2016; Chirumamilla, Sindre, & Nguyen-Duc, 2020). Some authors, such as Brimble (2016) and Sutherland-Smith (2016) go further and suggest that digital environments seem to promote fraud because of the facilities they offer for obtaining information, cutting and pasting, or accessing external help.

Types of Falsehood

Different types of falsehood can be distinguished. Chirumamilla et al. (2020) synthesize what has been studied in the literature in recent years and differentiate the following types of falsehood:

a) Substitution: Someone else performs the test.
b) Prohibited Aids: Using documents or tools not allowed during the test.
c) Copying: Copying the responses of other students. It can be done with the consent of the student who is being copied or without his or her consent.
d) Cooperation between students: Students cooperate with each other to answer the test.
e) Outside Assistance: Getting illegitimate (qualified) help from an outside party during the test.
f) Collusion: Illegitimate assistance of a teacher, official, or another student during the test or to gain access to the test in advance.

Most of these types of falsehood can be controlled or limited by a continuous identity authentication monitoring system. However,

the application of synchronous tests in multiple-choice format entails the need to analyze if some type of falsehood has occurred.

The question as to how to validate the results of tests taken remotely then arises. The concept of validity is central because, in addition to properly designed, reliable tests, the scores obtained need to be valid (Abad, Olea, Ponsoda, & García, 2011). It is assumed that teachers prepare assessments that meet the following criteria:

a) It is consistent with the learning goals and objectives.
b) It is correctly designed.
c) The items or tasks correctly represent the construct to be evaluated (Rodríguez Morales, 2017).
d) The items or tasks were properly constructed (Unidad de Apoyo a la Enseñanza [UAE], 2020).
e) The items are conditionally independent of each other and the distractors are appropriate, that is, the quality of the items has been studied (Rodríguez & Luzar, 2014).

These criteria provide evidence of the validity of the test. However, it is necessary to be able to provide evidence of the validity of remote assessment results.

Validation of Distance Learning Through Synchronous Tests

The need for the falsehood study as part of the evidence to be collected to validate the interpretation of the results is the approach that will be taken in this article. Considering the effects of fraud as reprehensible, the focus will be on the evidence of validity provided by falsehood studies. *The Standards for Educational and Psychological Testing* point out that validity is the aspect that should be considered as fundamental in the development and evaluation of tests (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education [AERA, APA, & NCME], 2014). Validity refers to how test scores are interpreted to determine whether the intended use is correct or not. Test scores are considered valid when the interpretations or inferences based on those scores are correct or accurate (Cizek & Wollack, 2017). Falsehood operates directly on test scores, making them less accurate, so it is necessary to analyze and detect this situation. In this way, the study of the falsehood contributes to the validity of the interpretation of the scores.

There are three moments in which it is necessary to collect evidence on the validity of the test results, that is, tasks to be performed before, during, and after its application (Rodríguez Morales, 2020), as presented in Figure 1.

Sindre and Vegendla (2015) demonstrated that remote testing is no less valid than paper-based testing and that the safety of online testing depends on the measures taken during these three moments.

The first thing to consider is the design and programming of the test. Preparing multiple-choice tests is not easy, so it is necessary to take

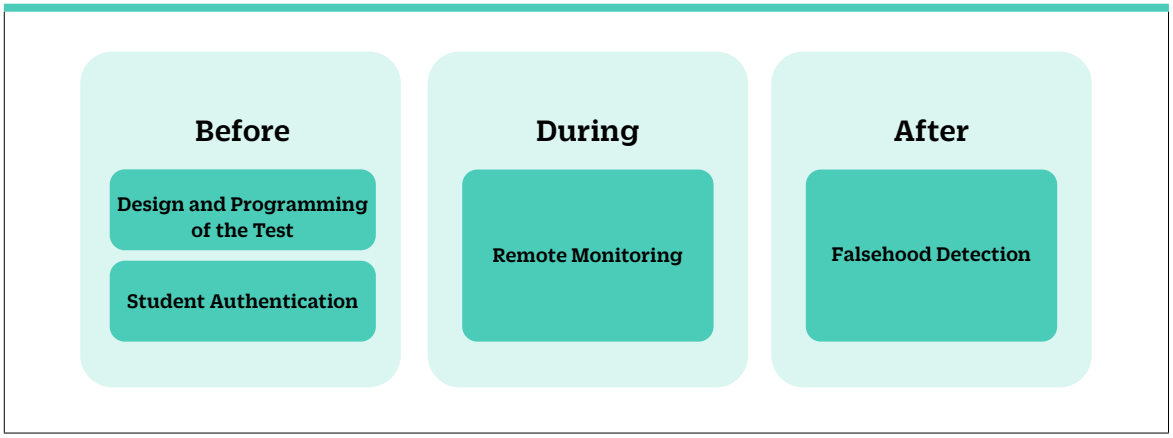

Figure 1. Moments to Gather Evidence on the Validity of Test Results

certain precautions when designing them. The following suggestions are provided:

- **a)** Elaborate a wide group of items, making sure to write the same amount of easy, medium, and difficult items. Muñiz and Fonseca-Pedrero (2019) recommend doubling the number of necessary items.
- **b)** Randomly arrange the questionnaire items. This prevents students from identifying the item with a position on the test.
- **c)** If fixed tests are designed, make equivalent booklets in terms of content coverage, item difficulty, number, and type.
- **d)** Randomly set the response options within each item (Udelar/CSE, 2020b).
- **e)** Avoid true/false items because they increase the probability of random responses.
- **f)** Choose between three to five response options for each item. Abad, Olea, and Ponsoda (2001) recommend items with three alternatives as they are less susceptible to the presence of partial knowledge. Ortega Torres and Chávez Álvarez (2020) also propose that three is the optimal number of alternatives for response options since only two plausible distractors need to be constructed.
- **g)** Avoid incorrect (distracting) response choices that are too obvious.
- **h)** Establish a reasonable time for the duration of the test in a synchronous way for the whole cohort. As Prieto and Delgado (1996) point out, it is necessary to achieve a balance between a number of items that ensure adequate reliability and the appropriate duration for the study level.
- **i)** Establish deferred feedback so that it is conducted once the test is finished and synchronously for all students (Udelar/CSE, 2020b).

Before the test, it is also necessary to authenticate the students and verify them on the platform used for the evaluation.

When synchronous tests are applied to students ranging from 150 to 2,500, as is the case in Udelar, it is necessary to implement monitoring systems. Virtual learning platforms have a continuous student authentication system. They provide constant monitoring through a set of software tools that use biometric facial recognition techniques, authorize the recording or photography of the students while they are taking the test, access to microphones for the recording of ambient audio, and control the opening of tabs on the student's computer. These systems are often invasive and controversial in terms of privacy. Students may perceive the invasion of personal space and feel watched (Noguera, Guerrero-Roldán, & Rodríguez, 2017).

Other types of remote test monitoring programs have been created. Through the TeSLA project, a verification system has been developed for student authentication and authorship of their work. Identity verification is carried out with biometric data obtained through the virtual platform, taking special care of the students' privacy, and the European educational, technological, and ethical standards. It has been used to evaluate constructed-response tests, e-portfolios, and peer-reviewed collaborative learning (Baró-Solé et al., 2018). In addition, it is especially useful for tests of a formative or continuous nature (Amigud, Arnedo-Moreno, Daradoumis, & Guerrero-Roldán, 2017).

However, in Latin American universities, these identity verification and monitoring systems are less common. At Udelar, teachers do not have the monitoring application in the Virtual Learning Environment platform. Therefore, the task of corroborating the identity of the person who is taking and monitoring its execution cannot be performed. For this reason, we have to make special use of the falsehood detection techniques that can be used after the application of multiple-choice tests (Yates, Godbey, & Fendler, 2017). The analysis to detect falsehood is very important in this context. So, it is necessary to know the different falsehood detection methods, their characteristics, potentialities, and limitations.

Falsehood Detection Methods

In this section, a literary review of the main falsehood detection methods is carried out, pointing out their uses, potentialities, and limitations. Those indexes that refer to falsehood detection in multiple-choice items are taken into account. Those that detect falsehood in essays or constructed-response tests were not considered

since the focus is on those that apply to the type of test used at Udelar. Research works that present or apply new methods, and those that compare the performance of existing methods in different scenarios, were reviewed.

All of these methods cannot prove that a fraud has taken place; they can only assign a probability that the cheating has occurred (Zopluoglu, 2017).

There are great detection methods for some types of falsehood. Thus, copying and cooperation among students have been addressed in the indexes that look for evidence of copying, that is, indexes of similar responses.

There are two main methods for detecting the different types of falsehood: response *similarity* indexes and person-fit indexes. The first ones analyze the degree of agreement between two response vectors. The second ones examine whether a single response vector is aligned with a certain response model. The response similarity indexes can be classified based on two characteristics: a) the reference statistical distribution and b) the similarity degree between the likelihoods of two response vectors (Zopluoglu, 2017). At the end of this section, some more recent approaches with alternative methods will also be presented.

Similarity Indexes

The first group of *similarity* indexes, which work with the number of identical incorrect responses, is constituted by the K (Saretsky, 1984), ESA (Bellezza & Bellezza, 1989), K1, and K2 (Sotaridona & Meijer, 2002) indexes, which use the binomial distribution. While an S1 (Sotaridona & Meijer, 2003) is based on the Poisson distribution. Within the group of indexes based on empirical distribution, the most recently developed index is VM (Belov, 2011).

The second group of indexes works on the number of identical correct and incorrect responses. S2 stands out in this group (Sotaridona & Meijer, 2003).

The third group of indexes considers all the items. In this group, ω stands out (Wollack, 1997) and has been widely used. Zopluoglu (2017) found it to be the best fit for similar responses.

Maynes (2017) studied the potential individual collusion with a *similarity* analysis. The types of falsehood where examinees receive help from an external source, communicate, or work together to obtain responses (cooperation), or where there is collusion, can be addressed with *similarity* statistics. Therefore, they are often very useful for the study of falsehood in electronic remote tests since students can easily communicate with external sources or other students. The bivariate statistic M4 (Maynes, 2017) is a recent option for the detection of this type of falsehood.

One approach to the study of collusion, as prior knowledge of the test items, is presented by Eckerly (2017), in which he used the Deterministic Gated Item Response Theory Model (DGM) (Shu, Leucht, & Henson, 2013).

Person-Fit Indexes

Psychometric analyses using person-fit statistics are important for detecting aberrant response patterns that produce inaccurate test scores. Aberrant response patterns include falsehood, but also other behaviors that lead to inaccurate measurements such as careless, creative, or random responses.

Person-fit indexes can be classified into parametric and non-parametric. The latter are not based on the parameters estimated by the Item Response Theory, but they are rather calculated from the data set of scores obtained in a test. Parametric indexes measure the distance between the test data set and the estimated response predictions derived from the parameter estimates of an Item Response Theory model.

There are a large number of person-fit indexes. However, there is little research on which ones are more useful. The last comparative study was that of Karabastos (2003), where 36 person-fit indexes (25 parametric and 11 non-parametric) were compared under different conditions to obtain a better consensus regarding their performance.

Besides, person-fit indexes have limitations to detect copying, as demonstrated by Zopluoglu (2017). However, HT (Sijtsma & Meijer, 1992) and D (Trabin & Weiss, 1983) are the best, as pointed out by Karabastos (2003).

Alternative Methods

In addition to the *similarity* and person-fit indexes, other procedures can be used depending on the available test information, as competitive or

complementary methods for falsehood detection. Remote tests in electronic multiple choice format allow obtaining relevant information such as response times. Several works incorporate this parameter to the Item Response Theory models, for example, van der Linden et al. (2006 & 2010). Recently, this type of modeling has been used as an alternative option for falsehood detection (Qian, Staniewska, Reckase, & Woo, 2016; Sinharay & Johnson, 2019; and Kasli Zopluoglu, & Toton, 2020). Another tool is to start from the assumption that students who falsify test results have a similar response pattern. Therefore, various automatic classification methods can provide information about these groups (Zopluoglu, 2019a; & Man, Harring, & Sinharay, 2019). Different algorithms can be implemented, e.g. k-means, SVM, random forests, neural networks, and their results can be compared with those obtained through classic indexes (Zopluoglu, 2019b).

Some Methodological Challenges

Similarity and person-fit indexes work under certain conditions and were developed for some specific test analyses. It is necessary to perform a comparative study of all the indexes, including the most recent ones, and contrast their performance in different situations. One of the latest index comparison studies was conducted almost 20 years ago. Karabastos (2003) studied 36 person-fit indexes and found that the HT indexes of Sijstma & Meijer (1992) and the D index of Trabin & Weiss (1983) have an acceptable performance. Among the similarity indexes, ω and GBT, which use Item Response Theory, the K index, and its counterpart, which does not use Item Response Theory, and the VM index were found to work well in terms of power and Type I error rates. In Uruguay, there are very few assessments that use Item Response Theory for item calibration. Within the university environment, it has only been used for diagnostic evaluation tests developed at Centro Universitario Regional del Este (Rodríguez Morales, 2017). Thus, it is desirable to study more deeply those that do not require the Item Response Theory. Doyoung, Woo, & Dickinson (2017) found that the U3 statistic, which does not use the Item Response Theory, is very easy to calculate, is promising, and its performance is similar to parametric indexes.

Table 1 shows a list of the main response-similarity and people-fit indexes, which present better fits under certain conditions. The indexes recommended by Karabastos (2003), Haney and Clarke (2007), de la Torre and Deng (2008), Guo and Drasgow (2010), Belov (2011, 2015), Eckerly, Babcock, & Wollack (2015), Doyoung et al. (2017), Maynes (2017), Wollack & Cizek (2017), Zopluoglu (2016, 2017, 2019a, 2019b), and Sanzvelasco, Luzardo, García, and Abad (2020), which were obtained through simulation studies, were selected.

As for the methodology to apply these indexes, Belov & Armstrong (2010) suggest a two-stage analysis: first, performing a screening using

Table 1

Similarity and Person-Fit Indexes

Response-Similarity Indexes	Person-Fit Indexes
ω (Wollack, 1997)	Z (Guo & Drasgow, 2010)
GBT (van der Linden & Sotaridona, 2006)	AMC, LRT, MSRLT (Sanzvelasco et al., 2020)
K (Kling apud Saretsky, 1984)	H^T (Sijtsma & Meijer, 1992)
K_1 y K2 (Sotaridona & Meijer, 2002)	D (Trabin y Weiss, 1983)
ESA (Bellezza &Bellezza, 1989),	U3 (van der Flier, 1980)
DGM (Shu et al., 2013) y *scale-purified* DGM (Eckerly et al., 2015)	*lz* (Drasgow, Levine & Williams, 1985 y de la Torre & Deng, 2008)
M4 (Maynes, 2017)	MCI (Harnisch & Linn, 1981)
VM (Belov, 2011)	

person-fit indexes to identify potential falsifiers and then applying *similarity* indexes between those potential falsifiers. This is an appropriate strategy for the detection needs in learning tests in the context of Udelar; however, more evidence of its scope and power is needed because some of its properties have not yet been sufficiently studied. New approaches should also be explored, such as the modeling of response times for falsehood detection proposed by Qian et al. (2016) and Sinharay & Johnson (2019) or the automatic classification methods presented by Zopluoglu (2019b) and Man, Harring, & Sinharay (2019).

As explained by Wollack and Cizek (2017), many of the approaches developed so far to detect falsehood identify unusual behavior without a solid understanding of the nature of that fraud. Also, the properties of the detection methods are not clearly identified. In other words, these approaches represent a great risk because they can detect falsehood where it did not occur or attribute some form of fraud to harmless atypical behavior. Wollack and Fremer (2013) suggest that the best way to transcend these risky approaches is to study these methodologies through research that combines studies through simulation and application to real data that help to better understand the properties of these methods for a diverse range of situations. Therefore, it is necessary to conduct research on falsehood by comparing these methods, identifying their properties, potentialities, and weaknesses through simulations at the beginning, and then applying data extracted from real tests.

Falsehood detection is a topic that includes test design, item analysis and calibration, and index application and interpretation. Some of these topics are not accessible to all university teachers. For this reason, it is necessary to create an application that can easily provide this information to teachers or those responsible for the assessments, so that they can make decisions based on the type of evidence found. These decisions can range from improving the design and validation of instruments to observing possible falsehood behavior that leads to applying complementary assessments or making decisions regarding main policies.

Discussion

The health crisis the world is going through has posed great challenges in the educational field, especially in terms of assessment of learning. The main challenge is how to ensure valid assessments and detect falsehood in synchronous remote tests. There are several aspects in which universities should advance to achieve these objectives. First, it is necessary to establish standards for the design and application of remote tests, just as they are developed for face-to-face tests (Rodríguez Morales, 2017). Establishing protocols that guide teachers in the design of valid and reliable evaluation instruments is a fundamental aspect of measurement. All phases of its development must be taken care of, from the framework, the definition of variables, the construction of items, the editing, the piloting, and the application of the test, as stated by Muñiz & Fonseca-Pedrero (2019). The work on the design and validation of tests should be developed through the creation of item banks, their analysis and calibration, and booklet matching. This may imply a limitation since it requires prior work on technical features, which not all academic units are in a position to cope with.

At the same time, it is necessary to install monitoring systems for remote testing through virtual teaching platforms or create other platforms exclusively for the application of tests. The properties of authentication and monitoring systems based on biometric techniques are described in Baró-Solé et al. (2018) and Noguera et al. (2017) for constructed-response tests. Hernández-Ortega, Daza, Morales, Fiérrez, Ortega-García (2019) also tested this type of system for different types of assessments, including multiple-choice tests. A limitation of the approach of this article is that it only addresses the falsehood detection indexes for this type of test, not the indexes for constructed-response tests or essays.

Additionally, in order to apply falsehood detection methods to multiple-choice tests with greater ease, it would be very useful to develop a web application, which with a set of test data, would provide information to professors about the analysis of the items, their calibration, and

different tests for the detection of falsehood. Research in falsehood detection methods, with the aim of finding those that work better in different contexts, is central to making these contributions more concrete. In this sense, it is necessary to carry out comparative studies of the performance and benefits of the different indexes presented through simulation and real data studies.

Electronic remote evaluation may be an option that can be offered in the future for students to take their tests—beyond the health emergency posed by Covid-19—as the correlate of a teaching style that will surely take hybrid formats where face-to-face education is combined with distance education. This way, it will be possible to combat overcrowding and democratize access for those who are geographically distant or have family or employment commitments, which are the postulates of this university (Udelar, 2020). In addition, it will be possible to promote flexibility, mobility, and accessibility, principles to which European universities are aligned (Noguera et al., 2017). Not all students will choose this option because they believe it has advantages; there is already evidence that many students are aware of the difficulties involved in remote assessments and do not choose it, as James (2016) found in his research.

Therefore, it is necessary to create academic units with technical skills to carry out these evaluation and research activities and to train more human resources qualified to perform these tasks.

References

Abad, F. J., Olea, J. & Ponsoda, V. (2001). Analysis of the optimum number alternatives from the Item Response Theory. *Psicothema*, 13 (1), pp. 152-158.

Abad, F. J., Olea, J., Ponsoda, V., & García, C. (2011). *Medición en ciencias del comportamiento y de la salud*. Madrid: Editorial Síntesis.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME]. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Amigud, A., Arnedo-Moreno, J., Daradoumis, T. & Guerrero-Roldán, A. (2017). Open Proctor: An Academic Integrity Tool for the Open Learning Environment. En: Barolli L., Woungang I., Hussain O. (Eds.) Advances in Intelligent Networking & Collaborative Systems. Lecture Notes on Data Engineering and Communications Technologies, vol 8. Springer, Cham. https://doi.org/10.1007/978-3-319-65636-6_23

Arnold, I. J. (2016). Cheating at online formative tests: Does it pay off? *Internet and Higher Education*, 29, 98–106.

Arthur, W., Glaze, R. M., Villado, A. J. & Taylor, J. E. (2010). The Magnitude and Extent of Cheating and Response Distortion Effects on Unproctored Internet-Based Tests of Cognitive Ability and Personality. *International Journal of Selection and Assessment*, 18 (1), 1-16.

Baró-Solé, X., Guerrero-Roldan, A.E., Prieto-Blázquez, J., Rozeva, A. Marinov, O., Kiennert, Ch., Rocher, P.O., Garcia-Alfaro, J. (2018). Integration of an adaptive trust-based e-assessment system into virtual learning environments—The TeSLA project experience. *Internet Technology Letters,* 1:e56. https://doi.org/10.1002/itl2.56

Banco Interamericano de Desarrollo [BID] (2020). La educación superior en tiempos de Covid-19. Aportes de la segunda reunión del Diálogo Virtual con Rectores de Universidades Líderes de América Latina. BID. https://publications.iadb.org/publications/spanish/document/La-educacion-superior-en-tiempos-de-COVID-19-Aportes-de-la-Segunda-Reunion-del-Di%C3%A1logo-Virtual-con-Rectores-de-Universidades-Lideres-de-America-Latina.pdf

Bird, C. (1927). The detection of cheating in objective examinations. *School and Society, 25*(635), 261–262.

Bird, C. (1929). An improved method of detecting cheating in objective examinations. *The Journal of Educational Research, 19*(5), 341–348.

Bellezza, F. S., & Bellezza, S. F. (1989). Detection of cheating on multiple-choice tests by using error similarity analysis. *Teaching of Psychology*, 16(3), 151–155.

Belov, D. I. & Armstrong, R. D. (2010). Automatic Detection of Answer Copying via Kullback-Leibler Divergence and K-Index. *Applied Psychological Measurement*, 34(6) 379–392. https://doi.org/10.1177/0146621610370453.

Belov, D. I. (2011). Detection of answer copying based

on the structure of a high-stakes test. *Applied Psychological Measurement*, 35(7), 495–517.

Belov, D. I. (2015). Comparing the performance of eight item preknowledge detection statistics. *Applied Psychological Measurement*, *40*(2), 83-97. https://doi.org/10.1177/0146621615603327

Brimble, M. (2016). Why students cheat: An exploration of the motivators of student academic dishonesty in Higher Education. En T. Bretag (Ed.), *Handbook of academic integrity* (pp. 365-382). Springer-Nature: Springer Science-Business Media Singapore.

Cizek, G. J. (2012). *Ensuring the integrity of test scores: Shared responsibilities*. Annual Meeting of the American Educational Research Association, Vancouver, British Columbia.

Cizek, G. J. & Wollack, J. A. (2017). Exploring cheating on tests. En G. J. Cizek y J. A. Wollack (Eds). *Handbook of quantitative methods for detecting cheating on tests* (pp. 3-19). New York: Routledge.

Conferencia de Rectores de Universidades Españolas [CRUE] (2020). *Informe sobre el impacto normativo de los procedimientos de evaluación online: protección de datos y garantía de los derechos de los y las estudiantes*. https://www.usal.es/files/informe_procedimientos_evaluacion_no_presencial_crue_16-04-2020.pdf.pdf

Chirumamilla, A., Sindre, G. & Nguyen-Duc, A. (2020): Cheating in e-exams and paper exams: the perceptions of engineering students and teachers in Norway. *Assessment & Evaluation in Higher Education*. https://doi.org/10.1080/02602938.2020.1719975.

de la Torre, J., & Deng, W. (2008). Improving person-fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement*, 45(2), 159–177.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38(1), 67–86.

Doyoung, K., Woo, A. & Dickison, P. (2017). Identifying and investigating aberrant responses using psychometrics-based and machine learning based approaches. En G. J. Cizek y J. A. Wollack. *Handbook of quantitative methods for detecting cheating on tests* (pp. 70-98). New York: Routledge.

Eckerly, C. A. (2017) Detecting preknowledge and item compromise. En G. J. Cizek & J. A. Wollack (Eds.) *Handbook of quantitative methods for detecting cheating on tests* (pp. 214-231). New York: Routledge.

Eckerly, C. A., Babcock, B., & Wollack, J. A. (2015) *Preknowledge detection using a scale-purified deterministic gated IRT model*. Annual meeting of the National Conference on Measurement in Education, Chicago, IL.

Guo, J., & Drasgow, F. (2010). Identifying cheating on Unproctored Internet Tests: the Z-test and the likelihood ratio test. *International Journal of Selection and Assessment*, 18(4), 351-364.

James, R. (2016). Tertiary student attitudes to invigilated, online summative examinations. *International Journal of Educational Technology in Higher Education*, 13-19. https://doi.org/10.1186/s41239-016-0015-0.

Friedman, A., Blau, I., & Eshet-Alkalai, Y. (2016). Cheating and feeling honest: Committing and punishing analog versus digital academic dishonesty behaviors in higher education. *Interdisciplinary Journal of e-Skills and Life Long Learning*, 12, 193-205. http://www.informingscience.org/Publications/3629

Hernandez-Ortega, J., Daza, R., Morales, A., Fierrez, J., & Ortega-Garcia, J. (2019). edBB: Biometrics and behavior for assessing remote education. *arXiv preprint arXiv:1912.04786*.

Karabastos, G. (2003). Comparing the Aberrant Response Detection Performance of Thirty-Six Person-Fit Statistics. *Applied Measurement in Education*, 16 (4), 277-298. https://doi.org/0.1207/S15324818AME1604_2

Kasli, M., Zopluoglu, C. & Toton, S. (2020). A deterministic gated lognormal response time model to identify examinees with item preknowledge. *PsyArXiv*, 9. https://doi.org/10.31234/osf.io/bqa3t

Haney, W. M., & Clarke, M. J. (2007). Cheating on tests: Prevalence, detection, and implications for online testing. In *Psychology of academic cheating* (pp. 255-287). Academic Press. https://doi.org/10.1016/B978-012372541-7/50015-2

Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*, 18, 133–146.

Man, K., Harring, J. R. & Sinharay, S. (2019). Use of data mining methods to detect test fraud. *Journal of Educational Measurement*, 56 (2), 251-279.

Maynes, D. D. (2014). Detection of non-independent test

taking by similarity analysis. In N. M. Kingston and A. K. Clark (Eds.) *Test fraud: Statistical detection and methodology*. Routledge: New York, NY, pp. 53–82.

Maynes, D.D. (2017). Detecting potential collusion among individual examinees using similarity analysis. En G. A. Cizek y J. A. Wollack (Eds.) *Handbook of quantitative methods for detecting cheating on tests* (pp. 47-69). New York: Routledge.

Muñiz, J. & Fonseca-Pedrero, E. (2019). Diez pasos para la construcción de un test. *Psicothema*, 31 (1), 7-16. https://doi.org/10.7334/psicothema2018.291

Noguera I., Guerrero-Roldán A.E., Rodríguez M.E. (2017) Assuring authorship and authentication across the e-assessment process. En: D. Joosten-ten Brinke, M. Laanpere (Eds.) *Technology Enhanced Assessment. TEA 2016. Communications in Computer and Information Science* (pp.86-92), vol 653. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-57744-9_8

Ortega Torres, L. & Chávez Álvarez, C. A. (2020). Eliminación del tercer distractor de ítems de opción múltiple en exámenes a gran escala. *Revista de Educación*, 388, 133-165. https://doi.org/10.4438/1988-592X-RE-2020-388-450

Prieto, G., & Delgado, A. R. (1996). Construcción de ítems. En J. Muñiz (Coord.). *Psicometría* (pp.105-138). Madrid: Universitas.

Qian, H., Staniewska, D., Reckase, M. & Woo, A. (2016). Using response time to detect item preknowledge in computerbased licensure examinations. *Educational Measurement: Issues and Practice*, 35 (1), 38-47.

Quality Assurance Agency for Higher Education (QAA) (2020). *Assessing with Integrity in Digital Delivery*. Covid-19 supporting resources. https://www.qaa.ac.uk/docs/qaa/guidance/assessing-with-integrity-in-digital-delivery.pdf

Rodríguez, P. & Luzardo, M. (2014). Study the quality of items using isotone nonparametric regression in a mathematics test. *International Meeting of Psychometric Society*. Madison, Wisconsin, Estados Unidos.

Rodríguez Morales, P. (2017). Creación, desarrollo y resultados de la aplicación de pruebas de evaluación basadas en estándares para diagnosticar competencias en Matemática y Lectura al ingreso a la Universidad. *Revista Iberoamericana de Evaluación Educativa*, 10 (1), 89 – 107. https://doi.org/10.15366/riee2017.10.1.005

Rodríguez Morales, P. (2020). Evaluación de Aprendizajes a Distancia. Desafíos y dificultades. *Seminario de Desafíos de la evaluación de los procesos de aprendizaje y proyección de los nuevos escenarios de enseñanza en la Universidad*. https://www.cse.udelar.edu.uy/blog/2020/05/21/seminario-virtual-sobre-los-desafios-de-la-evaluacion-y-los-nuevos-escenarios-de-la-ensenanza/

Sanzvelasco, S., Luzardo, M., García, C. & Abad. F., (2020). Comparing statistics to detect cheating on recruitment contexts: an application for small items' banks. Psicothema, 32 (4), 549-558. https://doi.org/10.7334/psicothema2020.86.

Saretsky, G. D. (1984). *The treatment of scores of questionable validity: The origins and development of the ETS Board of Review (ETS Occasional Paper)*. Princeton, NJ: Educational Testing Service. http://files.eric.ed.gov/fulltext/ED254538.pdf.

Sinharay, S. & Johnson, M. S. (2019). The use of item scores and response times to detect examinees who may have beneted from item preknowledge. *British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1111/bmsp.12187

Shu, Z., Leucht, R., & Henson, R. (2013). Using deterministic, gated item response theory model to detect test cheating due to item compromise. *Psychometrika*, 78, 481–497.

Sijtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's non-parametric IRT model. *Applied Psychological Measurement*, 16(2), 149–157.

Sindre, G., & A. Vegendla. 2015. E-Exams versus Paper Exams: A Comparative Analysis of Cheating-Related Security Threats and Countermeasures. *Norwegian Information Security Conference* (NISK) 8 (1): 34 - 45.

Sotaridona, L. S., & Meijer, R. R. (2002). Statistical properties of the K-index for detecting answer copying. *Journal of Educational Measurement*, 39(2), 115–132.

Sotaridona, L. S., & Meijer, R. R. (2003). Two new statistics to detect answer copying. *Journal of Educational Measurement*, 40(1), 53–69.

Sureda Negre, J., Comas Forgas, R. y Gili Planas, M. (2009). Prácticas académicas deshonestas en el desarrollo de exámenes entre el alumnado universitario español. *Estudios sobre Educación*, 17, 103-122.

Sutherland-Smith, W. (2016). Authorship, ownership, and

plagiarism in the Digital Age. In T. Bretag (Ed.), *Handbook of academic integrity* (pp. 575-589). SpringerNature: Springer Science-Business Media Singapore.

Trabin, T. E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item response theory models. En D. J. Weiss (Ed.) *New horizons in testing* (pp. 83–108). New York, NY: Academic Press.

Unidad de Apoyo a la Enseñanza (UAE) (2020). *Pautas para la evaluación a distancia*. Maldonado: CURE. https://www.cse.udelar.edu.uy/recursos/wp-content/uploads/sites/16/2020/05/EVALUACION-EN-LINEA-final_UAECURE.pdf

Universidad de la República/Comisión Sectorial de Enseñanza [Udelar/CSE] (2020a). *Udelar en línea. Orientaciones básicas para el desarrollo de la enseñanza y la evaluación*. Montevideo: Comisión Sectorial de Enseñanza. https://www.cse.udelar.edu.uy/wp-content/uploads/2020/04/UdelarEnLinea-OrientacionesBasicas.pdf

Universidad de la República/Comisión Sectorial de Enseñanza [Udelar/CSE] (2020b). *Enseñanza en línea. Orientaciones para la aplicación de pruebas objetivas masivas en línea*. Montevideo: Comisión Sectorial de Enseñanza. https://www.cse.udelar.edu.uy/wp-content/uploads/2020/07/PautasEvaluacionEnLinea-v2.pdf

Universidad de la República [Udelar] (2020). *Propuesta al país 2020-2024. Plan estratégico de desarrollo de la Universidad de la República*. https://udelar.edu.uy/portal/wp-content/uploads/sites/48/2020/09/Presupuesto_2020-2024.pdf

van der Flier, H. (1980). *Vergelijkbaarheid van individuele testprestaties* [Comparability of individual test performance]. Lisse: Swets & Zeitlinger.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, 31(2), 181-204.

van der Linden, W. J., & Sotaridona L. S. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics*, 31, 283–304.

van der Linden, W. J., Klein Entink, R. H. & Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, 34(5): 327-347.

Yates, M. C., Godbey, J. & Fendler, R. (2017). Observed Cheating and the Effects of Random Seat Assignment. *SoTL Commons Conference*, GA, USA.

Whitley, B. E. (1998). Factors associated with cheating among college students: A review. *Research in Higher Education*, 39(3), 235–274.

Wollack, J. A. (1997). A nominal response model approach for detecting answer copying. *Applied Psychological Measurement*, 21(4), 307–320.

Wollack, J. A. & Cizek, G. J. (2017). The future of quantitative methods for detecting cheating. En G. A. Cizek y J. A. Wollack (Eds.) *Handbook of quantitative methods for detecting cheating on tests* (pp. 390-399). New York: Routledge.

Wollack, J. A., & Fremer, J. J. (2013). *Handbook of test security*. New York: Routledge.

Zopluoglu, C. (2016). Classification performance of answer-copying indices under different types of irt models. *Applied psychological measurement*, *40*(8), 592-607. https://doi.org/10.1177/0146621616664724

Zopluoglu, C. (2017) Similarity, answer copying and aberrance. En G. A. Cizek y J. A. Wollack (Eds.) *Handbook of quantitative methods for detecting cheating on tests* (pp. 25-46). New York: Routledge.

Zopluoglu, C. (2019a). Computation of the Response Similarity Index M4 in R under the Dichotomous and Nominal Item Response Models. *International Journal of Assessment Tools in Education*, 6 (5), 1–19. https://doi.org/10.21449/ijate.527299

Zopluoglu, C. (2019b). Detecting examinees with item preknowledge in large-scale testing using extreme gradient boosting (xgboost). *Educational and Psychological Measurement*, 79(5):931-961.

RIDU / Revista Digital de Investigación en Docencia Universitaria / e-ISSN: 2223-2516