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

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

ARTÍCULO TEÓRICO

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

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

Recibido: 28/07/20 *Revisado: 15/10/20* *Aceptado: 27/10/20* *Publicado: 12-11-2020*

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.

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

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.

Autor corresponsal: *Pilar Rodríguez Morales*
Correo: *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.

Citar como:

Rodríguez, P. & Luzardo, M. (2020). Cómo asegurar evaluaciones válidas y detectar falseamiento en pruebas a distancia síncronas. *Revista Digital de Investigación en Docencia Universitaria, 14*(2), e1240. https://doi.org/10.19083/ridu.2020.1240

Introducción

Las universidades afrontaron múltiples desafíos para continuar impartiendo la enseñanza a distancia después de la suspensión de clases presenciales a causa del Covid-19. Uno de los mayores retos ha sido la evaluación de aprendizajes en el contexto de la virtualidad. En Europa son las universidades, en primer lugar, las que toman medidas sobre la evaluación a distancia. Luego, otras instituciones también lo hacen y brindan orientaciones. En España, la Conferencia de Rectores de Universidades Españolas ofrece pautas sobre cómo llevarlas a cabo (Conferencia de Rectores de Universidades Españolas [CRUE], 2020). En el Reino Unido, la Agencia de Aseguramiento de la Calidad, en su documento titulado *Assessing with Integrity in Digital Delivery,* insta a que las universidades trabajen para evitar el engaño o falseamiento, el plagio u otros comportamientos impropios (Quality Assurance Agency for Higher Education [QAA], 2020).

En América Latina, la mayor parte de las universidades han implementado clases a distancia a través de plataformas virtuales. Se ha identificado como una debilidad por parte de los Rectores de Universidades líderes de Latinoamérica la carencia de instrumentos de evaluación o acreditación de saberes en la enseñanza virtual (Banco Interamericano de Desarrollo [BID], 2020).

En Uruguay, la declaración de emergencia sanitaria a causa de la detección de los primeros casos de Covid-19 hizo que la Universidad de la República (Udelar), la principal del país, suspendiera inmediatamente las clases presenciales y adoptara la modalidad a distancia. El rectorado impartió los lineamientos generales y mediante los órganos centrales de la universidad brindó apoyo y orientaciones técnicas. Se planteó inicialmente el tema de la evaluación para la certificación de los aprendizajes de los estudiantes, recomendándose un adecuado diseño y programación de las pruebas en la plataforma para reducir los problemas de copia (Universidad de la República/Comisión Sectorial de Enseñanza [Udelar/CSE] (2020a). Más adelante, en virtud de las inquietudes de los docentes sobre las herramientas y métodos de evaluación a distancia, se brindaron pautas concretas para cuidar la calidad de los instrumentos y los mecanismos de control que se pueden poner en práctica (Udelar/CSE, 2020b).

Por otro lado, antes de la pandemia, la Udelar estaba dando sus primeros pasos en la educación a distancia como una herramienta para afrontar las clases con matrícula numerosa, características de nuestra universidad que tiene libre ingreso, y también permitir la accesibilidad a la educación superior en todo el territorio nacional. Si bien diversas herramientas tecnológicas estaban siendo usadas para la enseñanza en forma virtual, no se había avanzado en la evaluación a distancia para la acreditación de saberes. El contexto de la pandemia acelera la implementación de la enseñanza a distancia y el mantenimiento de las medidas sanitarias obliga a utilizar evaluaciones a distancia en formato múltiple opción, especialmente en las facultades con mayor matrícula por cohorte (Udelar/ CSE, 2020b).

Por estos motivos, se aborda en este artículo los últimos avances teóricos y metodológicos para asegurar la validez de los resultados en las pruebas a distancia en forma síncrona.

A través de una revisión bibliográfica se indaga en el concepto de fraude o falseamiento, para luego centrarse en los métodos para detectar falseamiento en pruebas de opción múltiple.

En primer lugar, se presentan los antecedentes teóricos sobre la temática y el enfoque psicométrico por el que debe ser estudiado el fraude o falseamiento. En un segundo apartado se describirán los principales métodos que son utilizados para detectarlo, incluyendo los últimos desarrollos. Por último, se describirán los desafíos que, a juicio de los autores, enfrentan las instituciones de educación superior para la validación de los resultados de pruebas síncronas realizadas a distancia.

Antecedentes

En este apartado se referirán los conceptos de fraude o falseamiento, sus diversos tipos y por último, se planteará el problema de la validación de los resultados de las pruebas.

El fraude en una prueba o examen, o lo que comúnmente se denomina "copia", es un problema inherente a las evaluaciones. Ha sido ampliamente estudiado y, lamentablemente, se ha comprobado que el alcance de esta práctica ha tenido cierta magnitud de consideración (Whitley, 1998; Arthur, Glaze, Villado y Taylor, 2010).

Si bien el análisis de datos sobre fraude en pruebas comenzó en la década de los noventa del siglo pasado, ya hace casi cien años que se da cuenta del interés de detectarlo (Bird, 1927, 1929). Hacia fines del siglo XX, Whitley (1998) encontró, en una revisión de estudios, que el 43% de los estudiantes universitarios admitieron haber copiado en pruebas. En un estudio realizado en España, la mitad de los estudiantes encuestados declararon haber copiado al menos una vez durante un examen (Sureda, Comas & Gili, 2009). Un tercio de los actos de fraude encontrados por Friedman, Blau y Eshet-Alkalai (2016) fueron realizados utilizando tecnología. Si bien los estudiantes parecen ser más atraídos a cometer fraude en pruebas en línea sin monitoreo, los resultados no son tan alentadores para los que cometen fraude porque su desempeño correlaciona negativamente con los demás cursos en los que sus evaluaciones son monitoreadas (Arnold, 2016).

Técnicamente, denominaremos falseamiento a distintos tipos de fraude en pruebas. Cizek (2012) define al falseamiento como cualquier acción tomada antes, durante o después de la administración de una prueba o tarea, con la intención de tomar ventaja injusta o producir resultados incorrectos. La educación a distancia se ha enfocado en este tema, ya que, tanto docentes como estudiantes consideran que es mucho más sencillo el falseamiento en una prueba tomada a distancia que en una presencial en soporte papel (Arnold, 2016; Chirumamilla, Sindre & Nguyen-Duc, 2020). Algunos autores, como Brimble (2016) y Sutherland-Smith (2016) van más allá y sugieren que los entornos digitales parecen promover el fraude por las facilidades que ofrecen para obtener información, cortar y pegar o acceder a ayuda externa.

Tipos de Falseamiento

Se pueden distinguir diferentes tipos de falseamiento. Chirumamilla et al. (2020) sintetizan lo estudiado por la literatura en los últimos años y diferencian los siguientes tipos de falseamiento:

a) Sustitución: alguien más realiza la prueba.
b) Ayudas prohibidas: usar documentos o herramientas no permitidas durante la prueba.
c) Copia: copiar las respuestas de otros estudiantes. Puede darse con el consentimiento del estudiante al que se le copia o sin su consentimiento.

d) Cooperación entre estudiantes: los estudiantes cooperan entre sí para responder la prueba.
e) Asistencia externa: conseguir ayuda ilegítima (calificada) de alguien externo durante la prueba.
f) Colusión: ayuda ilegítima de un docente o funcionario u otro estudiante durante la prueba o para tener acceso previo a la prueba.

La mayoría de estos tipos de falseamiento se puede controlar o limitar mediante un sistema de monitoreo continuo de autenticación de la identidad. Sin embargo, la aplicación de pruebas síncronas en formato opción múltiple conlleva la necesidad de analizar si se ha producido algún tipo de falseamiento.

Entonces, surge la interrogante sobre cómo validar los resultados de las pruebas tomadas a distancia. El concepto de validez aparece como central, porque además de pruebas adecuadamente diseñadas, fiables, necesitamos que las puntuaciones obtenidas sean válidas (Abad, Olea, Ponsoda & García, 2011). Se parte del supuesto de que el docente planteó una evaluación que reúne las siguientes condiciones:

a) Es acorde a los propósitos y objetivos de aprendizaje.
b) Está correctamente diseñada.
c) Los ítems o tareas representan correctamente el constructo a evaluar (Rodríguez Morales, 2017).
d) Los ítems o tareas fueron construidos adecuadamente (Unidad de Apoyo a la Enseñanza [UAE], 2020).
e) Los ítems son condicionalmente independientes unos de otros y los distractores son apropiados, es decir, que se haya estudiado la calidad de los ítems (Rodríguez & Luzardo, 2014).

Estas condiciones proporcionan evidencias de la validez de la prueba. Sin embargo, es necesario poder brindar evidencias sobre la validez de los resultados de las evaluaciones a distancia.

Validación de los Aprendizajes a Distancia a través de Pruebas Síncronas

La necesidad del estudio de falseamiento como parte de las evidencias por recoger para la validez de la interpretación de los resultados es el enfoque que tomaremos en este artículo. Considerando como reprobables los efectos del fraude, el foco estará en las evidencias sobre la validez que aportan los estudios de falseamiento. Los *Standards for Educational and Psychological Testing* señalan que la validez es el aspecto que se debe considerar como fundamental en el desarrollo y evaluación de las pruebas (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education [AERA, APA, & NCME], 2014). La validez refiere a la forma en que se interpretan las puntuaciones en las pruebas para determinar si el uso que se pretende es correcto o no. Las puntuaciones en las pruebas son consideradas válidas cuando las interpretaciones o inferencias basadas en esos puntajes son correctas o precisas (Cizek & Wollack, 2017). El falseamiento opera directamente sobre las puntuaciones de las pruebas, haciéndolas menos precisas, por eso es necesario que se analice y detecte esta situación. De esta forma, el estudio del falseamiento aporta a la validez de la interpretación de las puntuaciones.

Se pueden identificar tres momentos en los que es necesario recoger evidencias sobre la validez de los resultados de la prueba, esto es, tareas para realizar antes de su aplicación, durante y después de ella (Rodríguez Morales, 2020), como se presentan en la figura 1.

Sindre y Vegendla (2015) demostraron que las pruebas a distancia no son menos válidas que las realizadas en papel y que la seguridad de las pruebas en línea depende de las medidas tomadas durante estos tres momentos.

En primer lugar, se debe considerar el diseño y programación de la prueba. Construir pruebas de opción múltiple no es sencillo, por eso es necesario tomar ciertas precauciones a la hora de diseñarlas. Recogemos las siguientes sugerencias:

a) Elaborar un banco de ítems amplio, cuidando de redactar la misma cantidad de ítems fáciles, medios y difíciles. Muñiz y Fonseca-Pedrero (2019) recomiendan que se elabore el doble de ítems de los necesarios.
b) Disponer en forma aleatoria los ítems del cuestionario. De esta forma se evita que los

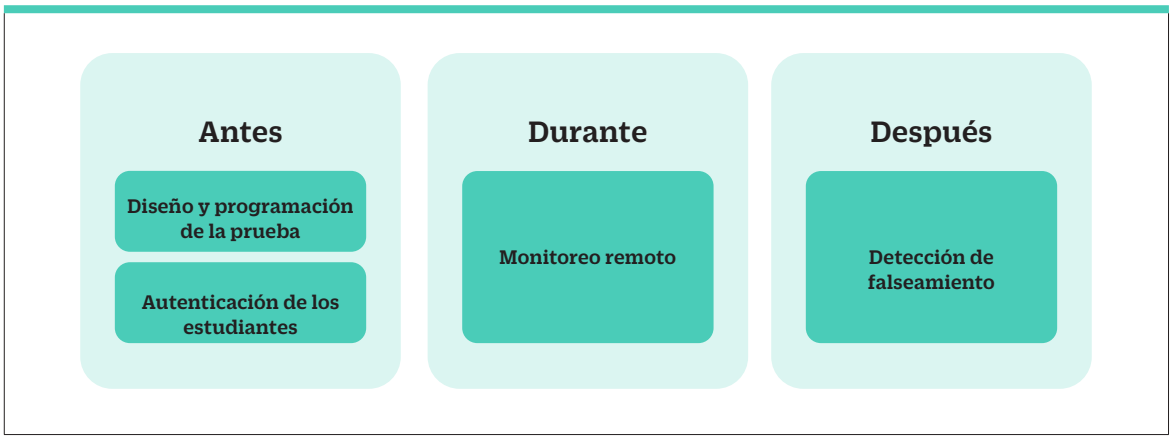

Figura 1. Momentos en Que Se Recogen Evidencias Sobre la Validez de los Resultados de las Pruebas

estudiantes identifiquen el ítem con una posición en la prueba.

c) Si se diseñan pruebas fijas, construir cuadernillos equivalentes en cuanto a la cobertura de contenidos, dificultad de los ítems, cantidad y tipo de ítems.

d) Establecer en forma aleatoria las opciones de respuesta dentro de cada ítem (Udelar/CSE, 2020b).

e) Evitar los ítems *verdadero/falso* porque aumenta la probabilidad de respuestas al azar.

f) Elegir entre tres y cinco opciones de respuestas para cada ítem. Abad, Olea y Ponsoda (2001) recomiendan los ítems de tres alternativas por ser menos susceptibles a la presencia de conocimiento parcial. Ortega Torres y Chávez Álvarez (2020) también proponen que tres opciones de respuesta es el número de alternativas óptimo, ya que solamente hay que construir dos distractores plausibles.

g) Evitar las opciones de respuestas incorrectas (distractores) que resulten demasiado obvias.

h) Establecer un tiempo razonable para la duración de la prueba en forma sincrónica para toda la cohorte. Como señalan Prieto y Delgado (1996) es necesario lograr un equilibrio entre un número de ítems que asegure una fiabilidad adecuada y la duración apropiada al nivel de estudios.

i) Establecer la retroalimentación diferida para que sea una vez finalizada la prueba y en forma sincrónica para todos los estudiantes (Udelar/CSE, 2020b).

Antes de la prueba también es necesario la autenticación de los estudiantes y su verificación en la plataforma utilizada para la evaluación.

Cuando se aplican pruebas síncronas a estudiantes que oscilan entre 150 y 2.500, como sucede en la Udelar, es necesario implementar sistemas de monitoreo. Las plataformas virtuales de aprendizaje poseen un sistema continuo de autenticación de los estudiantes. Proporcionan monitoreo constante a través de un conjunto de herramientas de software que utilizan técnicas biométricas de reconocimiento facial, autorizan la grabación o fotografía de los estudiantes mientras están realizando la prueba, el acceso a micrófonos para la grabación de audio ambiente y controlan la apertura de pestañas en la computadora del estudiante. Estos sistemas suelen ser invasivos y surgen controversias en cuanto a la privacidad. Los estudiantes pueden percibir invasión del espacio personal y sentirse vigilados (Noguera, Guerrero-Roldán & Rodríguez, 2017).

Se ha creado otro tipo de programas para el monitoreo de las pruebas a distancia. A través del proyecto TeSLA se ha desarrollado un sistema de verificación de la autenticación de los estudiantes y la autoría de sus trabajos. La verificación de la identidad se realiza con datos biométricos obtenidos a través de la plataforma virtual cuidando especialmente la privacidad de los estudiantes, las normas educativas, tecnológicas y éticas europeas. Se ha utilizado para evaluar pruebas de respuesta construida, e-portafolios y aprendizaje colaborativo revisado por pares (Baró-Solé et al., 2018). También

resulta especialmente útil para pruebas de carácter formativo o continuo (Amigud, Arnedo-Moreno, Daradoumis & Guerrero-Roldán, 2017).

Sin embargo, en las universidades latinoamericanas estos sistemas de verificación de la identidad y monitoreo son menos comunes. En la Udelar no tenemos disponible la aplicación para el monitoreo en la plataforma del Entorno Virtual de Aprendizaje, por lo tanto, no se puede realizar la tarea de corroboración de la identidad de quien realiza la prueba y su monitoreo durante su ejecución. Por eso, tenemos que valernos especialmente de las técnicas de detección de falseamiento que se pueden utilizar después de la aplicación de pruebas de opción múltiple (Yates, Godbey & Fendler, 2017). Los análisis para detectar falseamiento cobran vital importancia en nuestro contexto. Para ello es necesario conocer los distintos métodos de detección de falseamiento, cuáles son sus características, potencialidades y limitaciones.

Métodos de Detección de Falseamiento

En esta sección se realiza una revisión bibliográfica de los principales métodos de detección de falseamiento, señalando sus usos, potencialidades y limitaciones. Se toman en cuenta aquellos índices que se refieren a detección de falseamiento para ítems de opción múltiple. No se consideraron los que detectan falseamiento en ensayos o pruebas de respuesta construida, ya que nos concentramos en los que se aplican al tipo de prueba utilizada en nuestra universidad. Se revisaron tanto aquellas investigaciones donde se presentan y aplican nuevos métodos como las que comparan el desempeño de los existentes en distintos escenarios.

Todos estos métodos no pueden probar que ha tenido lugar un fraude, únicamente pueden asignar una probabilidad de que ese engaño haya ocurrido (Zopluoglu, 2017).

Algunos de los tipos de falseamiento han tenido un gran desarrollo de métodos para su detección. Así, la copia y la cooperación entre estudiantes ha sido abordada en los índices que buscan evidencias sobre copiado de respuestas, es decir, los índices de respuestas similares.

Se distinguen, principalmente, dos métodos para la detección de los distintos tipos de falseamiento: índices de *similaridad* de respuestas y los índices de ajuste de personas. Los primeros analizan el grado de acuerdo entre dos vectores de respuesta y los segundos examinan si un único vector de respuesta está alineado con cierto modelo de respuesta. Los índices de *similaridad* de respuesta pueden ser clasificados basándose en dos atributos: a) la distribución estadística de referencia y b) el grado de similitud entre las verosimilitudes de dos vectores de respuesta (Zopluoglu, 2017). También presentaremos, para finalizar este apartado, algunos enfoques más recientes con métodos alternativos.

Índices de Similaridad

Un primer grupo de índices de *similaridad*, que trabajan con el número de respuestas incorrectas idénticas, está constituido por los índices K (Saretsky, 1984), ESA (Bellezza & Bellezza, 1989), K_1 y K2 (Sotaridona & Meijer, 2002), que utilizan la distribución binomial. En tanto, el S_1 (Sotaridona & Meijer, 2003) está basado en la distribución de Poisson. Dentro del grupo de índices basados en la distribución empírica, el índice desarrollado más recientemente es el VM (Belov, 2011).

Un segundo grupo de índices trabaja sobre el número de respuestas idénticas correctas e incorrectas. En este grupo se destaca el S_2 (Sotaridona & Meijer, 2003).

Un tercer grupo de índices toma todos los ítems. En este grupo se destaca el ω (Wollack, 1997), que ha sido ampliamente utilizado. Zopluoglu (2017) encontró que era el que presentaba mejor ajuste en respuestas similares.

Maynes (2017) estudia la potencial colusión individual con análisis de *similaridad*. Los tipos de falseamiento donde los examinados reciben ayuda de una fuente externa, se comunican o trabajan en conjunto para obtener las respuestas (cooperación) o donde existe colusión, pueden ser abordados con estadísticos de *similaridad*. Por lo tanto, suelen ser muy útiles para el estudio del falseamiento en las pruebas a distancia en formato electrónico, ya que los estudiantes pueden tener facilidades en el acceso a la comunicación con fuentes externas u otros estudiantes. El estadístico bivariado M4 (Maynes, 2017) es una opción reciente para la detección de este tipo de falseamiento.

Un abordaje del estudio de la colusión, en tanto preconocimiento de los ítems de la prueba, es

presentado por Eckerly (2017), donde utiliza el *Deterministic Gated Item Response Theory Model* (DGM) (Shu, Leucht & Henson, 2013).

Índices de Ajuste de Personas

Los análisis psicométricos realizados mediante estadísticos de ajuste de personas son importantes para detectar patrones de respuesta aberrantes que producen puntuaciones inexactas en las pruebas. Los patrones de respuestas aberrantes incluyen el falseamiento, pero también otras conductas que llevan a mediciones poco precisas como las respuestas descuidadas, creativas o al azar.

Los índices de ajuste de personas se pueden clasificar en paramétricos y no paramétricos. Estos últimos no se basan en los parámetros estimados por la Teoría de Respuesta al Ítem, se calculan a partir del conjunto de datos de las puntuaciones obtenidas en una prueba. Los índices paramétricos miden la distancia entre el conjunto de datos de la prueba y las predicciones de respuestas estimadas derivadas del parámetro de estimaciones de un modelo de Teoría de Respuesta al Ítem.

Existen una gran cantidad de índices de ajuste de personas; sin embargo, hay pocas investigaciones sobre cuáles son más útiles. El último estudio comparativo fue el de Karabastos (2003) donde compara 36 índices de ajuste de persona (25 paramétricos y 11 no paramétricos), en diferentes condiciones, para obtener un mejor consenso en cuanto a sus desempeños.

Por otra parte, los índices de ajuste de personas presentan limitaciones para detectar copia, como demostró Zopluoglu (2017). Sin embargo, el H^T (Sijtsma & Meijer, 1992), el D (Trabin & Weiss, 1983) son los mejores, como señaló Karabastos (2003).

Métodos Alternativos

Además de los índices de *similaridad* y de ajustes de personas, otros procedimientos pueden ser utilizados, según la información disponible de la prueba, como métodos competitivos o complementarios para la detección del falseamiento. Las pruebas a distancia en formato electrónico tipo opción múltiple permiten obtener información relevante como los tiempos de respuesta. Varios trabajos incorporan este parámetro a los modelos de Teoría de Respuesta al Ítem, por ejemplo, van der Linden et al. (2006 y 2010). Recientemente, este tipo de modelización ha sido utilizado como un camino alternativo para la detección de falseamiento (Qian, Staniewska, Reckase & Woo, 2016, Sinharay & Johnson, 2019 y Kasli Zopluoglu & Toton, 2020). Otra herramienta por considerar es partir del supuesto de que los estudiantes que falsean los resultados en las pruebas tengan un patrón de respuestas similar, por lo tanto, diversos métodos de clasificación automática pueden arrojar información acerca de estos grupos (Zopluoglu, 2019a y Man, Harring & Sinharay, 2019). Pueden ser implementados diferentes algoritmos, por ejemplo k-medias, SVM, bosques aleatorios, redes neuronales y pueden compararse sus resultados con los obtenidos a través de los índices clásicos (Zopluoglu, 2019b).

Algunos Desafíos Metodológicos

Los índices de *similaridad* y de ajuste de personas funcionan en determinadas condiciones y fueron desarrollados para algunos análisis específicos de pruebas. Es necesario realizar un estudio comparativo de todos los índices, incluyendo los más recientes y contrastar su desempeño en diferentes situaciones. Uno de los últimos estudios de comparación de índices fue realizado hace casi 20 años. Karabastos (2003) estudió 36 índices de ajuste de personas y encuentra que los índices H^T de Sijtsma & Meijer (1992) y el D de Trabin & Weiss (1983) tienen un desempeño aceptable. Dentro de los índices de *similaridad* se encuentra que el ω y el GBT que utilizan Teoría de Respuesta al Ítem, el índice K y su contraparte que no usa Teoría de respuesta al Ítem y el índice VM funcionan bien en términos de potencia y de tasas de error tipo I. En nuestro país son muy escasas las evaluaciones que utilizan Teoría de Respuesta al ítem para la calibración de los ítems. Dentro del ámbito universitario solamente se ha utilizado para las pruebas de evaluación diagnóstica desarrolladas en el Centro Universitario Regional del Este (Rodríguez Morales, 2017). Entonces, es deseable estudiar más profundamente aquellos que no requieren Teoría de Respuesta al Ítem. Doyoung, Woo y Dickinson (2017) encuentran que el estadístico U3, que no utiliza Teoría de Respuesta al Ítem, es muy fácil de calcular, es prometedor y su desempeño es similar a los índices paramétricos.

En la Tabla 1, se presenta una lista de los principales índices de similaridad de respuesta y de ajuste de personas, que presentan mejores ajustes en determinadas condiciones. Se seleccionaron los índices recomendados por Karabastos (2003), Haney y Clarke, 2007, de la Torre y Deng, 2008, Guo y Drasgow (2010), Belov (2011, 2015), Eckerly, Babcock, & Wollack, 2015, Doyoung et al. (2017), Maynes (2017), Wollack & Cizek (2017), Zopluoglu (2016, 2017, 2019a, 2019b) y Sanzvelasco, Luzardo, García y Abad, 2020), que se obtuvieron a través de estudios de simulación.

En cuanto a la metodología para aplicar estos índices, Below & Amstromg (2010) sugieren un análisis en dos etapas: primero, hacer un *screening* usando índice de ajuste de personas para identificar potenciales falseadores y luego aplicar índices de *similaridad* entre esos potenciales falseadores. Esta es una estrategia adecuada para las necesidades de detección en pruebas de aprendizaje en el contexto de nuestra universidad, pero se necesitan más evidencias de su alcance y potencia porque algunas de sus propiedades aún no han sido suficientemente estudiadas. También se deben explorar nuevos enfoques, como la modelización sobre los tiempos de respuesta para la detección de falseamiento, que proponen Qian et al. (2016) y Sinharay & Johnson (2019) o los métodos de clasificación automática presentados por Zopluoglu (2019b) y Man, Harring & Sinharay (2019).

Como explican Wollack y Cizek (2017) muchos de los enfoques desarrollados hasta el momento para detectar falseamiento identifican comportamientos inusuales, sin una comprensión sólida de la naturaleza de ese fraude y no se identifican claramente las propiedades de los métodos de detección. Es decir, estos enfoques presentan un gran riesgo porque pueden detectar falseamiento donde no ocurrió o atribuir a comportamientos atípicos inocuos alguna forma de fraude. Wollack y Fremer (2013) sugieren que la mejor forma para trascender estos enfoques riesgosos es estudiar estas metodologías a través de investigaciones que combinen estudios mediante simulaciones y aplicaciones a datos reales que ayuden a comprender mejor las propiedades de estos métodos para una diversa variedad de situaciones. Por eso, es necesario investigar sobre el falseamiento, comparando estos métodos, identificando sus propiedades, potencialidades y debilidades a través de simulaciones en primer lugar y, luego, realizar aplicaciones a datos extraídos de pruebas reales.

La detección del falseamiento es un tema que incluye aspectos de diseño de pruebas, análisis y calibración de ítems, aplicación e interpretación de índices. Algunos de estos temas no son accesibles a todos los docentes universitarios. Por este motivo, es necesario crear una aplicación que pueda, en forma sencilla, aportar esta información a los docentes o responsables de las evaluaciones, de manera que se puedan tomar decisiones en función del tipo de evidencia encontrada. Estas

Tabla 1

Índices de Similaridad y de Ajuste de Personas

Índices de similaridad de respuesta	Índices de ajuste de personas
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)	

decisiones pueden ir desde la mejora en los diseños y validación de instrumentos hasta la observación de posibles conductas de falseamiento que lleven a aplicar evaluaciones complementarias o generar decisiones de políticas de centrales.

Discusión

Esta crisis sanitaria que transitamos ha planteado grandes desafíos en el ámbito educativo, especialmente en la evaluación de aprendizajes. El principal reto lo constituye cómo asegurar evaluaciones válidas y detectar falseamiento en las pruebas a distancia síncronas. Varios son los aspectos en que las universidades deberían avanzar para lograr estos objetivos. En primer lugar, se necesita establecer estándares para el diseño y aplicación de pruebas a distancia, de la misma forma en que se desarrollan para las pruebas presenciales (Rodríguez Morales, 2017). Acordar protocolos que guíen a los docentes en el diseño de instrumentos de evaluación válidos y fiables es un aspecto fundamental para la medición. Se deben cuidar todas las fases de su desarrollo, desde el marco, la definición de las variables, la construcción de ítems, la edición, el pilotaje y la aplicación de la prueba, como señalan Muñiz y Fonseca-Pedrero (2019). El trabajo en el diseño y validación de pruebas debe ser desarrollado a través de la creación de bancos de ítems, su análisis y calibración, y la equiparación de cuadernillos. Esto puede implicar una limitación, ya que requiere de un trabajo previo de características técnicas, que no todas las unidades académicas están en condiciones de afrontar.

A su vez, se deben instalar sistemas de monitoreo de pruebas a distancia a través de las plataformas virtuales de enseñanza o crear otras plataformas exclusivamente para la aplicación de pruebas. Las propiedades de los sistemas de autenticación y monitoreo basados en técnicas biométricas son descritas en Baró-Solé et al., (2018) y Noguera et al. (2017) para pruebas de respuesta construida. Hernández-Ortega, Daza, Morales, Fiérrez, Ortega-García (2019) también comprobaron este tipo de sistemas para distintos tipos de evaluaciones, incluidas las pruebas de opción múltiple. Una limitación del enfoque de este artículo es que solamente se abordan los índices de detección de falseamiento para este tipo de pruebas, no considerándose los índices para pruebas de respuesta construida o ensayos.

Asimismo, para aplicar los métodos de detección de falseamiento a las pruebas de opción múltiple con mayor facilidad sería muy útil desarrollar una aplicación en la web, donde dado un conjunto de datos de pruebas, proporcione información a los docentes acerca del análisis de los ítems, su calibración y distintas pruebas para la detección del falseamiento. La investigación en métodos de detección de falseamiento, con el objetivo de encontrar los que se desempeñan mejor en distintos contextos, es central para concretar estos aportes. En este sentido, es necesario realizar estudios comparativos del desempeño y bondades de los distintos índices presentados mediante estudios de simulación y con datos reales.

La evaluación a distancia en formato electrónico tal vez sea una opción que pueda ofrecerse en el futuro a los estudiantes para tomar sus pruebas, más allá del contexto de la emergencia sanitaria planteada por el Covid-19, como el correlato de una enseñanza que seguramente tomará formatos híbridos donde se combine la educación presencial con la a distancia. De esta manera se podrá combatir la masificación, democratizar el acceso de los más alejados geográficamente o comprometidos familiar o laboralmente, que son los postulados de nuestra universidad (Udelar, 2020). Además, se podrá promover los principios de flexibilidad, movilidad y accesibilidad a los que se alinean las universidades europeas (Noguera et al., 2017). No todos los estudiantes elegirán esta opción por creer que pueden encontrar ventajas, ya existen evidencias de que muchos estudiantes son conscientes de las dificultades que implica una evaluación a distancia y no la eligen, como encontró James (2016) en su investigación.

Por lo tanto, se necesita crear unidades académicas con capacidades técnicas para realizar estas actividades de evaluación e investigación y formar más recursos humanos que estén capacitados para realizar estas tareas.

Referencias

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