XVI Congreso Nacional de Investigación Educativa
CNIE-2021

Análisis de sospecha de copia en aplicaciones desde casa para los exámenes del Ceneval

Laura Ortega Torres
Dirección de Investigación, Calidad Técnica e Innovación Académica
Centro Nacional de Evaluación para la Educación Superior (Ceneval)

César Antonio Chávez Álvarez
Dirección de Investigación, Calidad Técnica e Innovación Académica
Centro Nacional de Evaluación para la Educación Superior (Ceneval)

Área temática 12. Evaluación educativa.
Línea temática: Evaluación a gran escala. Aportes metodológicos a la evaluación educativa.

Resumen

A partir de las restricciones para realizar actividades presenciales impuestas por la pandemia del COVID-19, el Centro Nacional de Evaluación para la Educación Superior (Ceneval) implementó la modalidad de aplicación *examen desde casa*, en la que los sustentantes resuelven las pruebas en su hogar. Esta obligó a incorporar distintas medidas de seguridad que se llevan a cabo antes, durante y después de una aplicación. El análisis forense de datos examina información después de que ocurrió la aplicación para detectar rápidamente posibles conductas indebidas como la copia, que sucede cuando un sustentante obtiene de otro las respuestas en el momento de la aplicación. Para contribuir a mantener la seguridad en las aplicaciones y verificar la equidad de las evaluaciones, como parte de estos análisis se han desarrollado diversos algoritmos matemáticos que detectan comportamientos sospechosos de copia. Para corroborar su eficacia, se simularon aplicaciones con un porcentaje de sustentantes con respuestas idénticas, como si estos evaluados se hubiera copiado. Estos datos se analizaron con algunos algoritmos de copia y se concluyó que los más efectivos para analizar las aplicaciones de los exámenes del Centro son el índice K y su variante K2. Ambos índices fueron implementados para verificar la seguridad en aplicaciones en la modalidad *examen desde casa* de procesos de admisión tanto para la educación media superior como superior del 2020 y confirmar la validez de las evaluaciones. Los resultados respaldaron la calidad del proceso de la evaluación, aportando información única sobre la seguridad de las aplicaciones.

Palabras clave: *Exámenes a gran escala, aplicaciones remotas, análisis forense de datos, índices de copia.*

La pandemia del COVID-19 ha impedido reunir a las personas que presentan algún examen a gran escala en un centro de evaluación, como se ha hecho de manera tradicional. El Centro Nacional de Evaluación para la Educación Superior (Ceneval), consciente de las necesidades de las instituciones educativas y de las demandas generadas por la nueva situación, puso a disposición la modalidad de aplicación *examen desde casa*, en la que los sustentantes lo resuelven, precisamente, desde su hogar, sin la presencia física de un aplicador que los monitoree. Esta modalidad demanda contar con nuevos mecanismos de seguridad, diferentes a la vigilancia que en situaciones normales llevan a cabo los aplicadores de forma presencial. Al respecto, el Ceneval ha incorporado distintas medidas, que se llevan a cabo antes, durante y después de la aplicación.

Las medidas incorporadas a partir del análisis de información después de que ocurrió una aplicación se denominan análisis forense de datos, que se refiere a la investigación de eventos para clasificar, identificar o interpretar incidentes raros o inusuales (Cizek & Wollack, 2016). Esta información y el análisis de las condiciones de la aplicación que pudieran promover o dejar que aparezcan estos incidentes permiten identificar aquellas que requieran modificarse y actualizar las políticas y regulaciones de las evaluaciones (International Test Commission, 2014).

La seguridad en las aplicaciones es uno de los motivos principales del análisis forense de datos en evaluación, el cual hace referencia a los métodos, análisis y procesos que pueden detectar diferentes comportamientos indebidos de los sustentantes y sus efectos negativos en el proceso de evaluación. Su objetivo es identificar estos comportamientos tan pronto como sea posible para tomar acciones que aseguren que los puntajes de estos sustentantes sean invalidados y que en el futuro este comportamiento sea eliminado o al menos desmotivado para que se disminuya su frecuencia.

Como se mencionó, en las aplicaciones presenciales en lugares controlados, como las instalaciones de una institución o las del Ceneval, existe la presencia del aplicador, quien tiene entre sus funciones la de monitorear a los sustentantes para disminuir la posibilidad de comportamientos indebidos. Si bien, para la modalidad del *examen desde casa*, en lo que se refiere a la seguridad, el Centro ha incorporado distintas medidas, como la corroboración de la identidad del sustentante por medio de la fotografía de su identificación junto con su rostro, el bloqueo de las funcionalidades de su equipo de cómputo durante la aplicación del examen para que sólo pueda acceder a éste sin posibilidad de abrir alguna otra aplicación o página web al mismo tiempo, así como la grabación del sustentante y de la pantalla de su computadora durante toda la aplicación del examen, se requieren medidas forenses que permitan detectar comportamientos indebidos rápidamente.

En la literatura sobre evaluación, un comportamiento indebido se denomina *fraudulento* y se ha definido como cualquier acción realizada antes, durante o después de la aplicación de una prueba con la intención de obtener una ventaja o alterar sus resultados. Una situación específica es el trabajo *coludido*, que es cuando dos o más personas trabajan juntas para obtener una ganancia para al menos uno de ellos (Cizek & Wollack, 2016). Esto incluye cuando uno de los sustentantes comparte sus respuestas de manera deliberada, hay comunicación

indebida durante el examen o se reciben las respuestas por parte de los profesores. La copia de respuestas es un caso especial de este trabajo coludido, donde un sustentante obtiene las respuestas de otro en el momento de la aplicación, con o sin la participación de éste.

Tomando en cuenta la importancia de evitar la ocurrencia de este tipo de comportamientos indebidos para la validez y equidad de la evaluación, se han desarrollado y se continúan desarrollando algoritmos matemáticos para la detección de copia (Ordoñez Camacho, 2011). De manera general, estos algoritmos comparan las respuestas entre cada pareja de sustentantes que participaron en un mismo proceso de evaluación. Se asume que si dos sustentantes contestaron exactamente lo mismo, se equivocaron en las mismas preguntas o en estos casos incluso eligieron la misma opción incorrecta, entonces hubo algún tipo de conducta indebida, como copia física, comunicación durante la resolución del examen o acceso previo a las preguntas. Estos algoritmos producen indicadores estadísticos de la ocurrencia de eventos muy poco usuales, por lo que por sí mismos en realidad son evidencia de una sospecha de copia.

Las primeras propuestas sencillamente comparaban el número de respuestas iguales entre un par de sustentantes, la cantidad de respuestas incorrectas idénticas (sin importar la opción elegida) o el número de errores en los que respondieron exactamente con la misma opción de respuesta. Sin embargo, estos valores no se estandarizaban ni se consideraba el nivel de habilidad de los sustentantes quienes, si tienen un alto conocimiento del tema, responderán correctamente a la mayoría de las preguntas, por lo que sus respuestas correctas en común serán muy similares a las de otros sustentantes con un conocimiento alto del tema e, incluso, es probable que sus respuestas incorrectas también sean similares. En estas situaciones los sustentantes eran marcados con sospecha de copia por estos índices. Posteriormente, se propuso estandarizar estos conteos, aunque sin mucho éxito para pruebas cortas y con un gran número de sustentantes, como el índice de Bellezza y Bellezza (1989).

Otras propuestas, como los índices de error o de error estandarizado de Guttman, evalúan los errores de los sustentantes individualmente por medio del cálculo de valores de acuerdo con lo que se espera del patrón de respuesta de los sustentantes dada su calificación global en la prueba (Meijer, 1994). Por ejemplo, si un sustentante contestó correctamente la mayoría de las preguntas se esperaría que tuviera errores en aquellas que son más difíciles, y que aquellos con calificaciones bajas hayan contestado correctamente las preguntas más fáciles, pero no las más difíciles. Cuando un sustentante se sale de estos patrones esperados, se detecta una sospecha de conducta indebida. Estos algoritmos tampoco fueron exitosos, por lo que se continuó desarrollando métodos de análisis de las respuestas o errores comunes entre parejas de sustentantes, pero estandarizados de acuerdo con la información obtenida de los sustentantes analizados.

El índice K y todas sus variantes estiman la probabilidad de obtener un número de respuestas incorrectas idénticas entre un par de sustentantes, referidos como *fuente* y *copiador* (Holland, 1996; Lewis & Thayer, 1998; Sotaridona & Meijer, 2002; 2003). La idea general es comparar sus respuestas elegidas cuando ambos

contestaron erróneamente. Se calcula una probabilidad de ocurrencia del número compartido de respuestas incorrectas y cuando ésta es muy baja se marca al sustentante *copiador* con sospecha de copia. Por ejemplo, cuando un par de sustentantes cometen 20 errores en una prueba es poco probable que dentro de estas respuestas incorrectas, 18 sean idénticas, mientras que es más probable que en esta cantidad de errores, tres respuestas sean iguales. Con los índices K, en el primer caso se obtendría una probabilidad menor de 0.05 (un evento muy poco usual), mientras que en el segundo caso el valor sería mayor (un evento común). En el primer ejemplo se marcaría al sustentante *copiador* lo que indicaría una sospecha de conducta de copia entre éste y el referido como *fuente*. La diferencia entre el índice K y sus variantes radica en cómo con cada uno de ellos se calcula la probabilidad de ocurrencia de las respuestas incorrectas idénticas.

Los algoritmos más complejos se basan en todas las respuestas (correctas, incorrectas e incluso las omitidas) entre cada pareja de sustentantes y usan información obtenida a partir de la teoría clásica de los test como los índices g2 (Frary, Tideman & Watts, 1977) y de Wesolowsky (2000); o de la teoría de respuesta al ítem como los índices ω (Wollack, 1997) y Zk, (Sotaridona, van der Linden, & Meijer, 2006), es decir, los modelos estadísticos más usados para calificar a los sustentantes, así como para analizar la calidad de los reactivos y las pruebas. Al realizar la calificación con estas teorías se obtiene una medida de habilidad de cada sustentante, mientras que al analizar la calidad de los reactivos se pueden conseguir diferentes parámetros del comportamiento psicométrico de los reactivos, como su dificultad, discriminación o incluso la probabilidad de elección de cada una de las opciones de respuesta (correcta e incorrectas). Dicha información es usada en estos índices de sospecha de copia para estimar la probabilidad de observar las mismas respuestas entre cada pareja de sustentantes y que dicha probabilidad no se deba al azar. Actualmente, existen variantes de algunos de estos índices que exploran diferentes condiciones matemáticas en sus cálculos, tales como distribuciones binomiales o de Poisson o regresiones lineales o cuadráticas.

En virtud de que los exámenes del Ceneval tienen características particulares, especialmente el gran número de sustentantes que presentan un examen en una misma fecha de aplicación, condición que ha mostrado afectar la efectividad de los índices de detección de copia, se llevó a cabo una investigación en la que se comparan algunos de los métodos antes mencionados. Se simularon datos de sustentantes que aplicaron los exámenes nacionales de ingreso del Ceneval para un mismo proceso de admisión a una institución de nivel medio superior o superior (licenciatura), los Exámenes Nacionales de Ingreso a la Educación Media Superior (EXANI-I) y Superior (EXANI-II), respectivamente. Ambos exámenes contienen reactivos de opción múltiple; en el EXANI-I Admisión se responden 92 preguntas y 112 en el EXANI-II Admisión. Las dos pruebas son de alto impacto en el futuro de los estudiantes en México porque sus resultados determinan el ingreso a una institución educativa.

Se simularon datos de aplicaciones donde se presentaron entre 50 y 9,600 sustentantes por aplicación. Se generaron datos similares a los obtenidos en estas aplicaciones, suponiendo que no hubo copia entre los sustentantes; posteriormente, se duplicaron todas las respuestas de un sustentante y con ellas se sustituyeron

todas las respuestas de otros sustentantes, como si le hubieran copiado al primero. Se manipularon los datos de cada aplicación simulada para que cada una tuviera 5% de sustentantes con respuestas idénticas, como si hubiera existido una situación de copia entre ese porcentaje de aplicados. Sabiendo que existía este porcentaje de respuestas idénticas en estos datos, se analizó si los valores de detección de sospecha de copia obtenidos con los índices de Belleza y Belleza, el índice K y sus variantes: K1, K2, S1 y S2, el índice g2 y el de Wesolowsky indicaban que, efectivamente, hubo "copia" entre los sustentantes cuyas respuestas fueron idénticas. Los resultados mostraron que los índices de Belleza y Belleza no detectaron este nivel de copia, que el g2 y el de Wesolowsky lo hicieron únicamente con grandes cantidades de sustentantes; mientras que los K1, S1 y S2 lo detectaron únicamente con cantidades pequeñas y que los índices K y K2 son los que consumen más tiempo en hacer los cálculos, pero detectaron la copia de forma más estable en todas las aplicaciones estudiadas. Se concluyó que los índices más efectivos para los exámenes del Centro son el K y su variante K2. Posteriormente, ambos fueron implementados para verificar la seguridad en algunas aplicaciones de ambos EXANI en la modalidad de *examen desde casa* del verano del 2020, donde aun cuando se cuente con mecanismos de monitoreo remoto continuo por medio de la cámara web y del micrófono del equipo del sustentante, dado el impacto del resultado de éstos exámenes, los aspirantes pudieran ser más propensos a comunicarse con otros por medio de sus dispositivos electrónicos para obtener las respuestas, lo que alteraría los resultados de los procesos de evaluación.

Como se indicó, estos análisis se introdujeron para garantizar la seguridad de las aplicaciones en la modalidad *examen desde casa*; sin embargo, también se llevaron a cabo en aplicaciones presenciales realizadas con el apoyo de un aplicador en 2019, para tener un referente de modalidades monitoreadas. Se compararon los resultados obtenidos en ambas modalidades de una misma institución; los resultados de dos de ellas se reportan a continuación.

Generalmente, las instituciones estudiadas tienen varios procesos de admisión en un mismo año; sin embargo, dadas las restricciones de la pandemia, en 2020 sólo se llevó a cabo uno, con la mayoría de los aspirantes en la modalidad *examen desde casa*. Para evitar que el número distinto de examinados generara diferencias en las estimaciones de los índices, se utilizaron los datos de la aplicación presencial del verano de 2019, con un número similar de sustentantes de la de 2020. De igual manera, se comparó la distribución de calificaciones en cada aplicación para asegurar que las poblaciones presentadas en ambos años fueran similares. Con esta selección se buscó aislar el efecto de la modalidad de aplicación para que la posible diferencia en el porcentaje de aspirantes detectados con sospecha de copia se pudiera atribuir a los actos indebidos que probablemente resultaron de la ausencia de un aplicador al momento de contestar el EXANI correspondiente, en la modalidad *examen desde casa.*

El estudio consistió en analizar el número sustentantes marcados con sospecha de copia (*copiadores*) para cada examinado (*fuente*), utilizando los índices K y K2. Por ejemplo, el sustentante A pudo ser marcado como

fuente de un solo copiador, mientras que el sustentante B podría identificarse con 100 sospechas de copia. Para tener un valor comparable, para cada examinado se estimó la proporción del número de posibles copiadores con relación al total de personas que respondieron el mismo examen. En el ejemplo anterior, y en una aplicación de 101 examinados, el sustentante A tendría una proporción de 0.01, mientras que el B alcanzaría un valor de 1.00 (el total de los aplicados fueron identificados como copiadores del sustentante B).

Después, se obtuvo el promedio de estas proporciones tomando en cuenta a todos los sustentantes y se transformó en porcentaje. Un valor de 100% de este promedio indica que existe sospecha de copia entre todos los examinados de la aplicación analizada; de 50%, que, en promedio, los aplicados tienen sospecha de copia con la mitad de la población, y un valor de 0%, que nadie fue marcado con sospecha de copia en la aplicación. Cabe resaltar que, debido a que el número de sustentantes influye en las estimaciones calculadas por los índices K y K2, sólo es viable comparar aplicaciones con cantidades de aspirantes similares. Para los análisis realizados, estas cantidades se controlaron para las aplicaciones del mismo EXANI en la misma institución usuaria, pero no entre exámenes o instituciones, por lo que los siguientes resultados no son comparables entre exámenes o instituciones distintos.

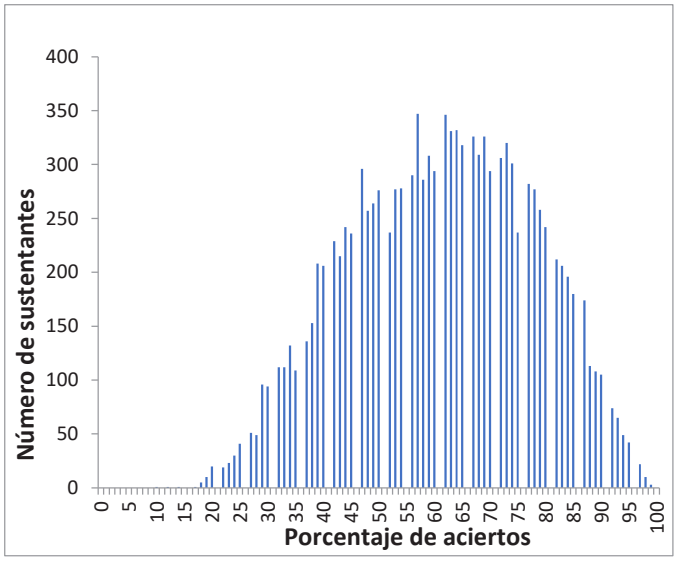
Para el EXANI-I Admisión se comparó una versión aplicada en 2019 de manera presencial en las instalaciones de la institución usuaria, en la modalidad *en línea*, a 10,058 sustentantes, con una aplicación examen *desde casa* a 8,146, en 2020, para la misma institución. Para la primera aplicación se encontraron valores promedio de 0.02% y 0.01% de sustentantes marcados con los índices K y K2, respectivamente. En el caso de la aplicación en modalidad *examen desde casa* se obtuvieron valores promedio de 0.06% para el índice K y 0.04% para el K2. Esto indica que en ambas aplicaciones la detección de sospecha de copia fue muy similar y cercana a 0%, con valores que indican una sospecha en menos de 0.10% de los evaluados.

Por otra parte, se analizó la aplicación del EXANI-II Admisión en esta misma institución, en la que se compararon dos versiones aplicadas en la modalidad *lápiz y papel*, de manera presencial en sus instalaciones, a 3,851 y 3,804 sustentantes en 2019, con una versión contestada de manera remota por 4,796 evaluados en 2020. Para 2019 se encontraron valores promedio de los examinados con sospecha de copia detectada por los índices K y K2 de 0.07% y 0.02%, para ambas versiones, mientras que éstos fueron de 0.12% y 0.05% para la versión de 2020. Los valores encontrados también sugieren una detección de sospecha de copia en cerca de 0% de los aspirantes.

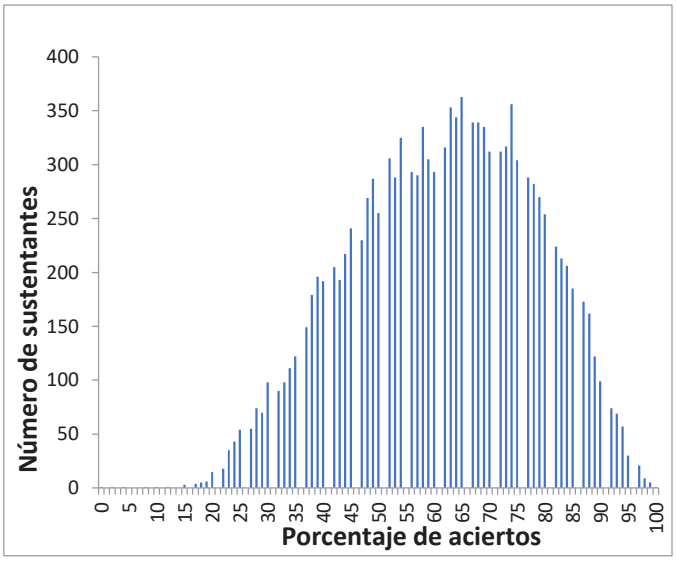
En otro análisis se compararon los valores promedio de sospecha de copia de 15 versiones del EXANI-I Admisión aplicadas remotamente, con los valores de 15 versiones aplicadas *en línea*, pero en la modalidad presencial, en las instalaciones de la misma institución en 2019. En ese año, en promedio, cada versión se aplicó en promedio a 820 sustentantes, mientras que en 2020 a 846. El valor mayor del promedio de sustentantes marcados con el índice K encontrado en una de las 15 versiones de la aplicación presencial fue de 0.21% y de 0.31% para la remota, mientras que los valores mayores calculados con el índice K2 para cada aplicación fueron de 0.01% y 0.10%. Estos valores continúan cerca de 0% y por debajo de 0.50% de la población presentada.

Para verificar que un aumento de 0.10% en el promedio de sujetos marcados con el índice K no impacta en los resultados de la evaluación, se analizaron las distribuciones de los porcentajes de aciertos del total de sustentantes aplicados: 12,306 en 2019 y 12,689 en 2020. A partir de lo anterior, no se observaron diferencias significativas entre el promedio del porcentaje de aciertos, ya que para la aplicación presencial fue de 60.71 y para la de *examen desde casa* de 61.14, y sus distribuciones son muy similares, como se observa en las gráficas 1. 2019 - Presencial, y 2. 2020 - Desde casa. Por lo tanto, este aumento en los valores obtenidos con el índice K no afectó los resultados de la evaluación.

Gráfica 1. 2019 - Presencial



Gráfica 2. 2020 - Desde casa



Los resultados de las aplicaciones analizadas muestran que todos los porcentajes promedio de sospecha de copia estimados con los índices K y K2 son menores de 0.50%, lo que indica que, en promedio, hubo sospecha de copia en menos de este porcentaje de sustentantes en cada aplicación. Además, los valores obtenidos fueron similares entre las aplicaciones en sitio y remotas, lo que sugiere que no hubo una fuga de información importante en la segunda, por lo que no importa si los sustentantes respondieron en presencia o ausencia de un aplicador; en ambos casos los resultados de las evaluaciones son semejantes y válidos.

Lo anterior indica que las medidas implementadas por el Centro para el *examen desde casa* evitaron que los sustentantes intentaran comunicarse con otros para obtener las respuestas a las preguntas. Por lo tanto, ayudaron a mantener la seguridad y calidad de las aplicaciones en esta modalidad como si hubieran sido presenciales. Las medidas fueron el bloqueo de las funcionalidades del equipo de cómputo del sustentante durante la resolución del examen, para que sólo pudiera acceder a éste, sin posibilidad de abrir otro programa o alguna página web al mismo tiempo; la grabación del sustentante y de la pantalla de su computadora durante toda la aplicación del examen; y alterar el orden de la presentación de los reactivos.

Por último, se debe recordar que cualquier índice de copia es únicamente un indicador estadístico de la posible ocurrencia de una conducta indebida. En caso de que alguno de ellos muestre que algún sustentante presenta una alta probabilidad de haber copiado, se requiere obtener más evidencia que respalde la afirmación de que la persona tuvo una conducta incorrecta. Gracias a los avances tecnológicos, esta información se puede obtener de los mismos programas usados para aplicar exámenes de manera remota. Mientras esta tecnología está lista, es posible revisar a detalle las grabaciones de las sesiones de las parejas de examinados marcados por los índices K y K2 con una alta probabilidad de sospecha de copia para recabar dicha evidencia.

La implementación del análisis forense de datos resulta útil para apoyar la preservación de la seguridad de las aplicaciones de los instrumentos del Centro en las distintas evaluaciones que se requieren en nuestro país, sin importar si éstas son en lápiz y papel o en línea, o si se presentan en las instalaciones de las instituciones o en el hogar de los sustentantes. Con esto, el Ceneval refrenda su compromiso de ser una agencia evaluadora de calidad y a la vanguardia, capaz de responder a los retos que se le presenten, como el confinamiento en nuestro país derivado de la pandemia por el COVID-19.

Referencias

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

Cizek, G. J., & Wollack, J. A. (Eds.). (2016). Handbook of quantitative methods for detecting cheating on tests. New York, NY: Taylor & Francis.

Frary, R. B., Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics*, 2(4), 235-256.

Holland, P.W. (1996). Assessing unusual agreement between incorrect answers of two examinees using the K-Index; *Statistical theory and empirical support ETS Research Report* No. 96-97). Princeton, NJ: Educational Testing Service.

International Test Commission (2014). International Guidelines on the Security of Tests, Examinations, and Other Assessments. Recuperado de: https://members.intestcom.org/files/guideline_test_security.pdf

Lewis, C., & Thayer, D. T. (1998). The power of the K-Index (or PMIR) to detect copying; *Statistical theory and empirical support ETS Research Report* No. 98-49. Princeton, NJ: Educational Testing Service.

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement*, 18, 311-314.

Ordoñez Camacho, X. G. (2011). Propiedades de algunos estadísticos empleados para la detección de copia de respuestas en la medición educativa. [Tesis de Doctorado, Universidad Complutense De Madrid].

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.

Sotaridona, L. S., van der Linden, W. J., & Meijer, R. R. (2006). Detecting answer copying using the kappa statistic. *Applied Psychological Measurement*, 30(5), 412-431.

Wesolowsky, G. (2000). Detecting excessive similarity in answers on multiple choice exams. *Journal of Applied Statistics*, *27*(7), 909-921.

Wollack, J. A. (1997). A nominal response model approach to detect answer copying. *Applied Psychological Measurement*, *21*, 307-320.