Análisis de patrones de respuesta en una evaluación estandarizada bajo el marco del análisis forense de datos

Palabras clave: Análisis forense de datos, detección de copia, patrones de respuesta, GBT, Omega de Wollack, evaluación estandarizada, educación

1 Introducción

Si bien existen maneras legítimas de mejorar los resultados en una prueba estandarizada muchas veces diversos actores del sistema educativo se sienten motivados para ayudar a los estudiantes a resolver las pruebas estandarizadas aplicadas a gran escala o incluso a modificar sus respuestas después de haber rendido la evaluación. Cada vez que existan consecuencias asociadas a los resultados de las pruebas habrá mayores amenazas a la seguridad de los test y por lo tanto posible presencia de trampa. Esta situación ha creado la necesidad de contar con estrategias de análisis de datos que ayuden a identificar este tipo de comportamientos irregulares, las cuales en conjunto se conocen como análisis forense de datos (data forensics). Existen diversos indicadores para analizar la posibilidad de copia, una de ellas es el estudio de la similitud de patrones entre estudiantes. Se ha investigado el funcionamiento de diversos índices, encontrando que tanto el omega (ω) de Wollack, como el Generalized Binomial Test (GBT) muestran resultados adecuados. Hay que considerar que un indicador estadísticamente significativo no necesariamente implica copia. Se presenta una aplicación en las evaluaciones censales aplicadas a los alumnos de 2. º grado de primaria en las áreas de Comunicación (Lectura) y Matemática, durante los años 2014, 2015 y 2016 en Perú.

2 Materiales y métodos

2.1 Descripción de los datos

Se trabajó con tres evaluaciones censales de medio millón de estudiantes de 2. º grado de primaria, aplicadas en los años 2014, 2015 y 2016, donde se evaluaron dos áreas: Comunicación (Lectura) y Matemática. En total, estos alumnos rindieron 42 preguntas en la prueba de Matemática y 46 en la de Comunicación.

Además de las variables de respuesta a cada una de las preguntas, se incluyeron variables de gestión educativa (no estatal, estatal), área geográfica (urbano, rural) y la Dirección Regional Educativa (DRE) a la que pertenece la escuela.

2.2 Métodos

El índice ω de Wollack (2003, 2006). Es una aproximación normal a la distribución binomial compuesta para el número de respuestas correctas e incorrectas idénticas entre dos vectores de respuestas; además es un estadístico asimétrico, quiere decir que se puede explorar la similitud al comparar la copia en ambas direcciones (p.ej. X copia de Y, e Y copia de X) y dividir el nivel de alfa (Zopluoglu, 2017). Se utilizó una modificación del paquete CopyDetect2 (Zopluoglu, 2012).

Generalized Binomial Test (GBT) de van der Linden y Sotaridona (2006). Se estima como la probabilidad de observar *m* o más coincidencias en N ítems; además es un indicador simétrico. Se utilizó una modificación del paquete CopyDetect2 (Zopluoglu, 2012)

Agregados a nivel de sección. Se hicieron todas las comparaciones pareadas entre los estudiantes de una sección. El indicador agregado implicó, a nivel de sección, qué porcentaje de esas comparaciones fueron estadísticamente significativas. Para esto también se generó un script en R, utilizando el tidyverse.

Reportes generales y por Dirección Regional Educativa. Se realizaron reportes automatizados en Rmarkdown para mostrar los resultados a nivel nacional, por estratos y por cada una de las 26 Direcciones Regionales Educativas que tiene el país.

3 Resultados

- A nivel nacional, en los tres años estudiados, la correlación de ambos indicadores es fuerte en Lectura (0,85; 0,88 y 0,95) y moderada para Matemática (0,59; 0,84 y 0,84); por lo que se ha optado por presentar los resultados con solo un indicador, GBT.
- En el año 2014, en la prueba de Lectura, a nivel nacional, el 73,5 % de secciones tiene entre 0 y 1% de comparaciones entre estudiantes con un indicador GBT estadísticamente significativo. Este porcentaje es mucho más grande en comparación a lo encontrado en el 2015 (45,3 %) y 2016 (50,2 %).
- La mayoría de las secciones (aproximadamente el 85 %) tiene menos del 20 % de comparaciones entre estudiantes con un indicador GBT estadísticamente significativo en los años 2014 y 2015. Mientras que en el 2016 es el 73 % de secciones.
- En el año 2014, en la prueba de Matemática, a nivel nacional, el 45,2 % de secciones tiene entre 0 y 1% de comparaciones entre estudiantes con un indicador GBT estadísticamente significativo. Este porcentaje es casi el doble en comparación a lo encontrado en el 2015 (22,5 %) y 2016 (24,7 %).
- La mayoría de las secciones (aproximadamente el 75 %) tiene menos del 20% de comparaciones entre estudiantes con un indicador GBT estadísticamente significativo en el 2014. Mientras que en el 2016 es aproximadamente el 64 % y menos en el 2015 (48,3 %).

Referencias

1. van der Linden, W. & Sotaridona, L. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioural Statistics, 31*(3), 283-304. https://doi.org/10.3102/10769986031003283
2. Wollack, J. A. (2003). Comparison of answer copying indices with real data. *Journal of Educational Measurement, 40*(3), 189-205. https://doi.org/10.1111/j.1745-3984.2003.tb01104.x

3. Wollack, J. A. (2006). Simultaneous use of multiple answer copying indexes to improve detection rates. *Applied Measurement in Education, 19*(4), 265-288. https://doi.org/10.1207/s15324818ame1904_3
4. Zopluoglu, C. (2012). CopyDetect: an R package for computing statistical indices to detect answer copying on multiple-choice examinations. *Applied Psychological Measurement*, *37*(1), 93–95. doi:10.1177/0146621612463119
5. Zopluoglu, C. (2017). Similarity, answer copying, and aberrance: understanding the status quo. En G.J. Cizek & J.A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 25-46). New York: Routledge.