Detección de patrones aberrantes en pruebas tipo test: Una aplicación en el Grado en Psicología de la Universidad de Barcelona

Detection of aberrant response patterns in multiple-choice tests: An application in the Degree in Psychology of the University of Barcelona

Georgina Guilera[1], Maite Barrios[1], M. Victòria Carreras Archs[1], Grupo AMERRA[2]
gguilera@ub.edu, mbarrios@ub.edu, vcarreras@ub.edu

[1]Departamento de Psicología Social y Psicología Cuantitativa
Universidad de Barcelona
Barcelona, España

[2]Facultad de Psicología
Universidad de Barcelona
Barcelona, España

[2]Los integrantes del grupo AMERRA son (por orden alfabético): Abdelhamid, G. S. M., Aznar-Casanova, J. A., Barrios, M., Beltran, F. S., Birulés, J., Bono, R., Carreras Archs, M. V., De la Fuente-Arnanz, F. J., Fuentemilla, L., Gómez-Benito, J., González-Gómez, B., Greco, A.M., Guilera, G. (Investigadora Principal), Keil, M., Navarra, J., Núñez-Peña, M. I., Pons, F., y Rojas-Castellanos, M. V.

Resumen- Las pruebas con ítems de respuesta múltiple son una práctica habitual en el contexto universitario. Los patrones de respuesta incoherentes, denominados Patrones Atípicos de Respuesta (PAR), aparecen cuando las respuestas correctas e incorrectas a los ítems no siguen el patrón esperado, i.e., acertar los ítems más fáciles y fallar los más difíciles. Este estudio pretende identificar los PAR en una prueba tipo test de la asignatura Psicometría del Grado en Psicología de la Universidad de Barcelona y explorar su relación con el rendimiento académico. Un total de 214 alumnos matriculados durante el curso 2018-2019 respondieron a una prueba tipo test de 20 ítems. Se calcularon cuatro índices de detección y se identificaron 13 PAR (6,1%), con más presencia en estudiantes con una calificación superior. La detección de PAR debería contribuir a la mejora de los sistemas de evaluación y del proceso de aprendizaje de los alumnos.

Palabras clave: ítems de respuesta múltiple, patrón aberrante de respuesta, educación universitaria, evidencias de validez

Abstract- Multiple-choice tests are a common practice in the university setting. Incoherent response patterns, known as Aberrant Response Patterns (ARP), appear when correct and incorrect responses to items do not follow the expected pattern, i.e., correctly answering the easiest items and failing the most difficult ones. This study aims to identify ARP in a multiple-choice test of the subject Psychometrics of the Degree in Psychology at the University of Barcelona. A total of 214 students enrolled during the 2018-2019 academic year responded to a 20-item test. Four detection indices were computed, and 13 ARP were identified, with more presence in students with a higher academic level. The detection of ARP should contribute to the improvement of the evaluation systems and the learning process of students.

Keywords: multiple-choice items, aberrant response pattern, university education, validity evidence

1. Introducción

Las pruebas con ítems de respuesta múltiple, un tipo de prueba tipo test, han constituido durante años uno de los métodos tradicionales de evaluación del nivel de conocimientos en el contexto universitario y, actualmente, siguen siendo una estrategia de evaluación ampliamente utilizada dada su objetividad e imparcialidad y su facilidad de corrección en grupos numerosos (Opazo Salvatierra, Sepúlveda Obreque, & Pérez Cabaní, 2015). En este tipo de pruebas la puntuación total se establece sumando 1 punto por cada ítem acertado y, habitualmente, penalizando los errores aplicando un factor corrector -1/(k-1), donde k es el número de alternativas de respuesta-. Con esta forma de proceder, la interpretación que hace el profesor de la puntuación total de la prueba es que el estudiante que, por ejemplo, obtiene una puntuación de 6 (sobre 10) domina el 60% de los contenidos y desconoce el 40%, asumiendo que el patrón de respuestas es coherente con la dificultad de los ítems. Es decir, si se ordenan los ítems de más fácil a más difícil, lo esperable es que este estudiante responda correctamente al 60% de los ítems más fáciles y falle el 40% de los ítems más difíciles (i.e., escalograma perfecto de Guttman). Cuando esta lógica no se cumple, es decir, cuando los patrones de respuesta emitidos no se corresponden con el nivel de dificultad de los ítems, aparecen los Patrones Atípicos de Respuesta (PAR) (Meijer & Sitjsma, 2001), los cuales identifican patrones de respuesta incoherentes, aberrantes, inesperados o inconsistentes (Meijer, Niessen, & Tendeiro, 2016).

De este modo, una misma puntuación total en una prueba tipo test puede proceder de patrones de respuesta diferentes, algunos coherentes con el modelo de Gutmann y otros inconsistentes o inesperados (i.e., PAR). Siguiendo con el ejemplo anterior, dos estudiantes que obtienen una puntuación total de 6 pueden haber alcanzado dicha puntuación de forma distinta, acertando los ítems más fáciles y fallando los más difíciles (i.e., ausencia de PAR) o, al contrario, fallando los ítems más fáciles y acertando los más difíciles (i.e., presencia de PAR).

La presencia de PAR, aunque implica una amenaza a la validez de las inferencias realizadas a partir de la puntuación total de la prueba sobre el nivel de conocimientos del estudiante (International Test Commission, 2013; Reynolds, 2010),

DOI: 10.26754/CINAIC.2019.0038

permite identificar la existencia de diferentes estilos de aprendizaje o de estrategias frente a las pruebas tipo test. Así, por ejemplo, los PAR pueden resultar de: a) malas pautas de estudio (e.g., centrar el estudio en los contenidos más complejos y desatender a los más sencillos); b) conductas ilícitas (e.g., copiar respuestas); c) proporcionar respuestas aleatorias; d) ser extremadamente creativo (e.g., reinterpretar los ítems fáciles por ser demasiado simples para ser ciertos); e) ser descuidado o negligente (e.g., no revisar las respuestas a los ítems más fáciles); f) dificultades reales (e.g., problemas de comprensión lectora); entre otras causas (Doval, Riba, Fuentes, & Renom, 2017; Karabatsos, 2003; Meijer, 1996).

En este contexto, la detección de PAR supone una herramienta especialmente útil en los procesos de evaluación formativa, donde el profesorado proporciona retroalimentación al estudiante con la finalidad de identificar carencias en el proceso de aprendizaje por parte del alumnado y de corregirlas de cara a futuras situaciones de evaluación (Sánchez Santamaría, 2011).

2. Contexto

El presente estudio se encuadra en el proyecto 'Ansiedad Matemática, Estilos de Respuesta y Rendimiento Académico' (AMERRA), el cual se está llevando a cabo en la Facultad de Psicología de la Universidad de Barcelona en el curso 2018-2019, en el marco del Programa de Investigación en Docencia Universitaria (REDICE) de la misma universidad. El proyecto pretende explorar la relación entre la ansiedad matemática y ansiedad ante los exámenes, los estilos de respuesta (patrones atípicos de respuesta, tiempo de ejecución y cambios de respuesta en las pruebas tipo test) y el rendimiento académico real y percibido de los estudiantes, en asignaturas con diferente presencia de contenidos matemáticos. Dicho proyecto todavía está en desarrollo, por lo que los resultados que aquí se presentan son parciales, a la espera de terminar el curso y, con ello, la recogida de datos.

Con el actual estudio se pretende identificar los PAR en una prueba tipo test de la asignatura Psicometría del Grado en Psicología de la Universidad de Barcelona mediante cuatro índices no paramétricos y explorar su relación con el rendimiento académico. Los objetivos específicos que se persiguen son los siguientes: a) juzgar la adecuación de las pruebas tipo test utilizadas en la asignatura, atendiendo al número de PAR identificados; b) analizar la concordancia entre procedimientos de detección de PAR; c) explorar la relación de los PAR con el rendimiento académico; y d) valorar la posibilidad de incorporar la detección de PAR de forma rutinaria en el sistema de retroalimentación individualizada al estudiante. La detección de PAR debería contribuir a la mejora tanto de los sistemas de evaluación utilizados en la asignatura Psicometría como del sistema de retroalimentación que se proporciona al estudiante cuando acude a revisar la calificación que ha obtenido en la prueba.

3. Descripción

A. Participantes

La muestra está formada por 214 estudiantes (79,4% mujeres) matriculados en el curso 2018-2019 en tres grupos de turno de mañana de la asignatura Psicometría del Grado en Psicología de la Universidad de Barcelona.

B. Instrumentos

El equipo docente de la asignatura Psicometría dispone de un banco de aproximadamente 500 ítems con cuatro alternativas de respuesta, donde únicamente una alternativa era correcta, clasificados por el contenido que evalúan y la naturaleza teórica o práctica de la pregunta. Cada curso académico, mediante una tabla de especificaciones de los bloques temáticos y conocimientos a evaluar, se seleccionan al azar aquellos ítems que formarán la prueba en cuestión. Este procedimiento se realiza en dos ocasiones para diseñar dos pruebas tipo test, una realizada a mediados del semestre y otra a finales de curso. Un ejemplo de ítem que pretende evaluar el bloque temático *Transformación de puntuaciones* a nivel práctico es el siguiente:

¿Qué puntuación transformada corresponde a una puntuación directa de 30 en un test que tiene una media de 25 y una desviación típica de 4?

a) Puntuación típica de -1,25
b) Puntuación T de 75
c) Penta de 4
d) Estanina de 3

Para este estudio, los alumnos respondieron a una prueba de conocimientos tipo test de 20 ítems, diseñada a partir del banco de ítems. Cada acierto contribuyó con 1 punto a la puntuación total, mientras que los errores penalizaron restando 0,33 puntos. La puntuación se transformó a una escala de 0 a 10 puntos.

C. Procedimiento

En el horario habitual de clase los alumnos respondieron a la prueba, disponiendo de 90 minutos para ello.

D. Análisis de datos

Los PAR se identificaron a través de cuatro índices no paramétricos basados en la Teoría de Respuesta al Ítem: C (Sato, 1975), U3 (van der Flier, 1980), MCI (Harnisch & Linn, 1981) y H^T (Sijtsma, 1986; Sijtsma & Mejer, 1992), al ser índices que funcionan adecuadamente en situaciones diversas (Karabatsos, 2003; Tendeiro & Meijer, 2014). Para el cálculo de los índices se empleó el paquete *PerFit* de R (Tendeiro, 2015; Tendeiro, Meijer, & Niessen, 2016), tratando como errores las respuestas en blanco.

Para cada índice se calculó el punto de corte mediante 1000 simulaciones bootstrap de la distribución muestral del índice en cuestión con un nivel de significación del 5%.

La concordancia entre procedimientos de detección de PAR se analizó obteniendo el coeficiente Kappa de Fleiss para múltiples codificadores (Fleiss, Levin, & Paik, 2003).

4. Resultados

En la Figura 1 se muestra la distribución de puntuaciones en la prueba de conocimientos. La puntuación media fue de 6,65 (DT = 2,00), con un rango comprendido entre 0,69 y 10. Fueron 46 (21,5%) los estudiantes con una calificación de suspenso, 60 de aprobado (28,0%), 87 de notable (40,7%) y 21 de excelente (9,8%).

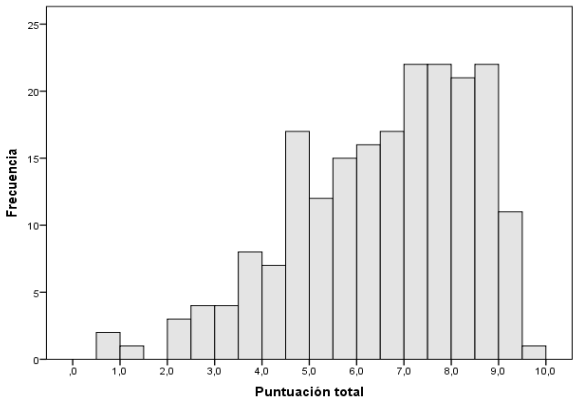

Figura 1. Distribución de la puntuación total de la prueba tipo test.

En la Tabla 1 se muestran los PAR identificados con cada uno de los procedimientos, acompañados del punto de corte empleado. El acuerdo entre índices fue considerable, con un coeficiente Kappa de Fleiss de 0,91.

Tabla 1. Casos identificados con PAR en función del índice.

Índice	Punto de corte	n (%) PAR	Casos con PAR
C	> 1,10	12 (5,6%)	27-60-63-65-104-109-134-141-178-189-193-203
U3	> 0,54	10 (4,7%)	27-60-63-65-134-141-178-189-193-203
MCI	> 0,61	12 (5,6%)	60-63-65-104-109-134-141-161-178-189-193-203
H^T	< -0,05	12 (5,6%)	27-60-63-65-104-109-134-141-178-189-193-203

A modo de ejemplo, la Figura 2 presenta la distribución del índice H^T obtenida con el paquete *PerFit*. La línea vertical señala el punto de corte por debajo del cual los casos son identificados como PAR (i.e., -0,048).

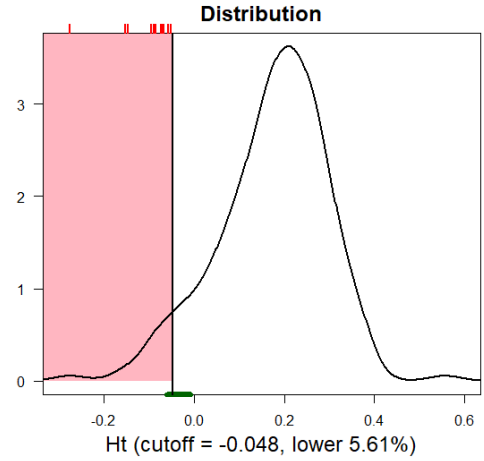

Figura 2. Distribución del índice H^T y punto de corte utilizado en la detección de PAR.

En conjunto, se identificaron 13 PAR, lo que supone un 6,1% de los examinados. Este bajo porcentaje constata que la validez de la prueba tipo test utilizada no se encuentra amenazada por la presencia de PAR. La Figura 3 muestra el patrón de respuesta para el participante 60, identificado como PAR por los cuatro índices empleados. Como puede observarse, su patrón se caracteriza por haber fallado los ítems más fáciles y de dificultad media y, en cambio, haber acertado los ítems más difíciles.

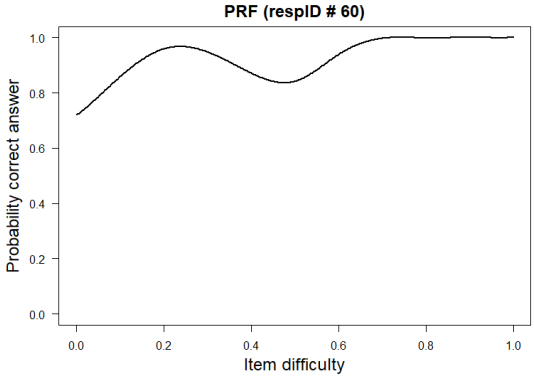

Figura 3. Patrón de respuesta del caso 60.

Al explorar la relación entre los PAR y el rendimiento académico, se observó una mayor presencia de PAR en aquellos alumnos con un nivel de conocimientos más elevado (ver Figura 4), en la línea de los resultados encontrados por otros autores (Petridou & Williams, 2007).

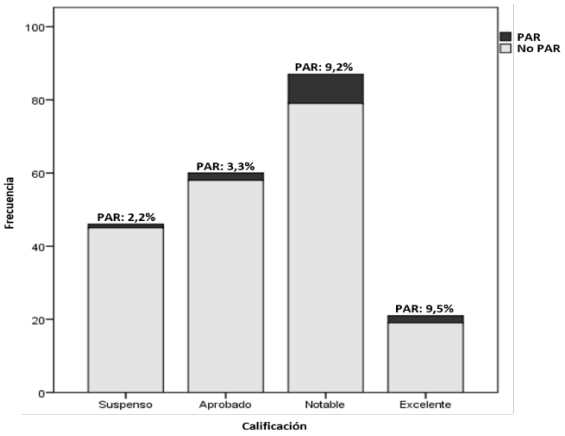

Figura 4. Porcentaje de PAR en función de la calificación.

5. Conclusiones

En este estudio se ha presentado un procedimiento de detección de PAR en la asignatura Psicometría del Grado en Psicología de la Universidad de Barcelona, con la finalidad última de mejorar los sistemas de evaluación (i.e., pruebas tipo test) y de retroalimentación utilizados en dicha asignatura.

A modo de conclusión, y retomando los objetivos específicos del estudio, primero, el bajo porcentaje de PAR detectado constata la adecuación de utilizar el banco de ítems, depurados y mejorados curso tras curso, para crear las pruebas tipo test de la asignatura. Segundo, la concordancia entre procedimientos de detección de PAR es muy elevada, lo que indica que parece irrelevante qué índice emplear y, en consecuencia, en cursos futuros podría utilizarse un único índice de detección para optimizar el coste en el tiempo del análisis. Tercero, aunque el porcentaje de PAR es escaso, existe una asociación entre la presencia de PAR y la calificación obtenida, con más presencia de PAR en aquellos alumnos con un nivel medio-alto de conocimientos. Finalmente, con este estudio se confirma que es factible incorporar el análisis de PAR en la planificación de la asignatura y añadir de forma rutinaria los resultados de dicho análisis en la sesión de retroalimentación que se le ofrece al estudiante para ayudarle a mejorar su proceso de aprendizaje de cara a evaluaciones futuras.

En definitiva, la identificación de PAR aporta información complementaria a la puntuación total obtenida en una prueba tipo test y es especialmente útil en aquellos contextos en los que la evaluación tiene una finalidad formativa en la que se proporciona retroalimentación al estudiante sobre su propio proceso de aprendizaje. Determinar las causas específicas de la presencia de PAR en un alumno y contexto de evaluación concretos requiere de una aproximación cualitativa (e.g., realización de una entrevista con el alumno) para averiguar las estrategias que ha seguido al estudiar los contenidos y al responder a la prueba tipo test (Rupp, 2013). Del mismo modo, conocer qué ítems son los que contribuyen en mayor medida a la presencia de PAR requiere de análisis complementarios (Boixadera, García Rueda, Doval, Riba, Renom, & Fuentes).

El procedimiento de detección de PAR que se ha ejemplificado en este estudio es transferible a otros contextos educativos donde: a) se trabaje con grupos de matrícula numerosos, dado que los índices de detección basados en la Teoría de Respuesta al Ítem requieren tamaños muestrales considerables; b) el proceso de evaluación se realice mediante pruebas con ítems de respuesta múltiple; y c) se persiga una finalidad formativa, proporcionando una retroalimentación individualizada al estudiante. Es recomendable aplicar dicho procedimiento en pruebas con al menos 10 ítems y tamaños muestrales suficientes para garantizar una estimación adecuada de los índices de detección (Meijer et al., 2016; Rupp, 2013).

Agradecimientos

Este estudio ha sido financiado por el *Programa de Recerca en Docència Universitària REDICE-18* (Código de proyecto: REDICE18-2222) de la Universidad de Barcelona.

Referencias

Doval, E., Riba, M. D., Fuentes, M., & Renom, J. (2017). Los patrones atípicos de respuesta. Una fuente de información para evaluar la calidad del proceso enseñanza-aprendizaje. Comunicación oral presentada al *5th International Congress of Educational Sciences and Development*, 25-27 Mayo, Santander, España.

Fleiss, J.L., Levin, B., & Paik, M.C. (2003). *Statistical methods for rates and proportions* (3rd Edition). New York: John Wiley & Sons.

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.

International Test Commission (2014). International Test Commission. (2014). ITC guidelines on quality control in scoring, test analysis, and reporting of test scores. *International Journal of Testing*, *14*(3), 195-217.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*(4), 277-298.

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education*, *9*(1), 3-8.

Meijer, R. R., Niessen, A. S. M., & Tendeiro, J. N. (2016). A practical guide to check the consistency of item response patterns in clinical research through person-fit statistics: Examples and a computer program. *Assessment*, *23*(1), 52-62.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*(2), 107-135.

Opazo Salvatierra, M., Sepúlveda Obreque, A., & Pérez Cabaní, M. L. (2015). Estrategias de evaluación del aprendizaje en la universidad y tareas auténticas: percepción de los estudiantes. *Diálogos Educativos, 15*(29), 19-33.

Petridou, A., & Williams, J. (2007). Accounting for aberrant test response patterns using multilevel models. *Journal of Educational Measurement, 44*(3): 227-247.

Reynolds, C. R. (2010). Measurement and assessment: An editorial view. *Psychological Assessment, 22*, 1-4.

Rupp, A. A. (2013). A systematic review of the methodology for person fit research in item response theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling*, *55*, 3-38.

Sánchez Santamaría, J. (2011). Evaluación de los aprendizajes universitarios: una comparación sobre sus posibilidades y limitaciones en el Espacio Europeo de Educación Superior. *Revista de Formación e Innovación Educativa Universitaria, 4*(1), 40-54.

Sato, T. (1975). *The construction and interpretation of S-P tables*. Tokyo: Meiji Tokyo.

Sijtsma, K. (1986). A coefficient of deviant response patterns. *Kwantitative Methoden, 7,* 131–145.

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,* 149–157.

Tendeiro, J. N. (2015). *PerFit* (version 1.4.3) [Computer software]. University of Groningen. Retrieved from https://CRAN.R-project.org/package=PerFit

Tendeiro, J. N., & Meijer, R. R. (2014), Detection of invalid test scores: The usefulness of simple nonparametric statistics. *Journal of Educational Measurement, 51*, 239-259.

Tendeiro, J. N., Meijer, R. R., & Niessen, A. S. M. (2016). PerFit: An R package for person-fit analysis in IRT. *Journal of Statistical Software*, *74*(5), 1-27.

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