



Identificación de tipologías de Patrones Atípicos de Respuesta en pruebas tipo test
Eduardo Doval y M.Dolors Riba, Universitat Autònoma de Barcelona, eduardo.doval@uab.cat

Resumen:

Las pruebas tipo test para la evaluación de los conocimientos están ampliamente aceptadas por su objetividad y facilidad en la aplicación y corrección. Sin embargo, para asegurar la validez de sus puntuaciones es necesario verificar que éstas han sido obtenidas después de responder a la prueba de una forma prevista por el evaluador. Los patrones de respuesta que se alejan de lo previsto se consideran Patrones Atípicos de Respuesta (PAR).

En este trabajo se presenta una metodología encaminada a identificar la presencia de PAR así como su tipología. Dado que los diferentes tipos de PAR pueden identificar déficits en el proceso de enseñanza-aprendizaje, esta metodología pretende ser de utilidad para evaluaciones formativas.

Palabras clave:

Pruebas tipo test, patrones atípicos de respuesta, evaluación formativa

Esta investigación se ha realizado gracias a la ayuda de la Dirección General de Investigación y Gestión del Plan Nacional de I+D+I, del Ministerio de Economía y Competitividad, a través del proyecto EDU2013-41399-P. Agradecemos al Consell Superior d'Avaluació de Catalunya su colaboración en este proyecto.

1. Objetivos o propósitos:

El objetivo del trabajo es presentar una metodología de tratamiento de las respuestas a pruebas de evaluación tipo test a fin de a) identificar la presencia de Patrones Atípicos de Respuesta (PAR), y b) una vez identificados los estudiantes que responden a la prueba de una forma atípica, proporcionar un instrumento que ayude a establecer el tipo de perfil de respuesta en relación con la dificultad de los ítems.

2. Marco teórico:

Las pruebas tipo test constituyen una forma objetiva e imparcial de evaluar conocimientos, siempre que se diseñen de forma adecuada (Lane, Haladyna y Raymond, 2016) y las preguntas se redacten siguiendo las pautas que preserven

Organizado por:





su validez (Haladyna y Rodríguez, 2013; Moreno, Martínez y Muñiz, 2015). Estas características, junto con su facilidad en la administración y corrección, han hecho de las pruebas tipo test uno de los formatos elegidos en la mayoría de las evaluaciones a gran escala, como las pruebas PISA, TIMMS y PIRLS, a nivel internacional, o las pruebas de evaluación de las competencias básicas que realiza el Instituto Nacional de Evaluación Educativa (INNE) y las diversas comunidades autónomas, a nivel nacional.

La forma más simple de obtener un indicador del conocimiento evaluado a partir de las respuestas dadas a las preguntas de elección múltiple consiste en otorgar un punto a cada respuesta correcta (y cero puntos a las incorrectas), y sumar el número de puntos obtenidos. La interpretación de esa puntuación depende del número de preguntas que tenga la prueba, de manera que, para obtener un indicador más fácil de interpretar, suele ser habitual calcular la proporción de preguntas contestadas correctamente o multiplicar esa proporción por 10 para obtener una puntuación con valores entre 0 y 10.

Con el mismo criterio, la proporción de respuestas correctas que el grupo de alumnos evaluado ha dado a una determinada pregunta de la prueba es un indicador inverso del nivel de dificultad de dicha pregunta (cuanto más baja sea la proporción de respuestas correctas más difícil será la pregunta).

Cuando a partir de la puntuación obtenida por un alumno se infiere su nivel de conocimientos en la materia evaluada, se está dando por hecho que ese alumno ha seguido una pauta de respuestas coherente con la dificultad de las preguntas. Así pues, si por ejemplo contestara la mitad de las preguntas de la prueba correctamente, cabría esperar que responda correctamente a las preguntas más fáciles y falle las más difíciles. Sin embargo, la puntuación obtenida en la prueba no informa sobre cuál ha sido la pauta de respuestas seguida, de manera que dos alumnos pueden obtener la misma puntuación a partir de patrones de respuesta completamente diferentes, e incluso en algunos casos, incompatibles con la lógica antes descrita. Sin embargo, si se llegara a dar este caso, dado que obtienen la misma puntuación concluiríamos que poseen el mismo nivel de conocimientos. Las pautas de respuestas a las preguntas de una prueba que no se corresponden con los niveles de dificultad previstos para las preguntas que se responden bien (por ejemplo, contestar sistemáticamente mejor a las preguntas difíciles que a las fáciles), se denominan Patrones Atípicos de Respuesta.

Para la identificación de PAR se han propuesto numerosos índices (Karabatsos, 2003; Meijer y Sitjsma, 2001), algunos paramétricos, como los que toman como referencia de patrón de respuesta el previsto por algún modelo de Teoría de

Organizado por:





Respuesta a los ítems (TRI), y otros no paramétricos. La mayoría de ellos puede calcularse fácilmente con el paquete PerFit de R (Tendeiro, 2015). Uno de los índices no paramétricos que ofrecen un buen rendimiento en la detección de PAR (Karabatsos, 2003) es el índice de Precaución Modificado (MCI; Harnisch y Linn (1981), y que compara el patrón de respuestas estudiado con los patrones de respuestas esperados si la persona evaluada hubiera respondido de acuerdo a la dificultad de las preguntas y a su nivel de conocimiento (es decir, si hubiera respondido correctamente a las preguntas más fáciles e incorrectamente a las difíciles), o si hubiera respondido completamente en desacuerdo con dicha dificultad (es decir, si hubiera respondido correctamente a las preguntas más difíciles e incorrectamente a las más fáciles). El MCI proporciona valores entre 0 (patrón de respuestas esperado) y 1 (patrón de respuesta completamente contrario al esperado), y puede considerarse que los valores superiores a 0.35 corresponden a presencia de PAR.

Igual que ocurre con los otros índices, el MCI permite identificar la presencia de PAR, pero no ofrece información acerca del tipo concreto de respuestas que ha dado lugar a ese patrón atípico.

3. Metodología:

El método propuesto consiste en ordenar las preguntas de la prueba en función de su índice de dificultad (proporción de respuestas correctas), calcular los centiles 33 y 66 de los valores de dificultad de las preguntas y utilizarlos como puntos de corte para definir tres categorías de preguntas en función de su nivel de dificultad (dificultad baja, B; dificultad media, M; dificultad alta, A). Posteriormente, y a nivel individual de cada estudiante, identificar los patrones atípicos de respuesta con el índice MCI. El último paso consiste en calcular, para cada alumno, el porcentaje de respuestas correctas obtenido en cada bloque de preguntas (B-M-A, por ese orden), con el objeto de identificar su perfil de respuestas.

4. Discusión de los datos, evidencias, objetos o materiales

[Insertar texto]

Organizado por:

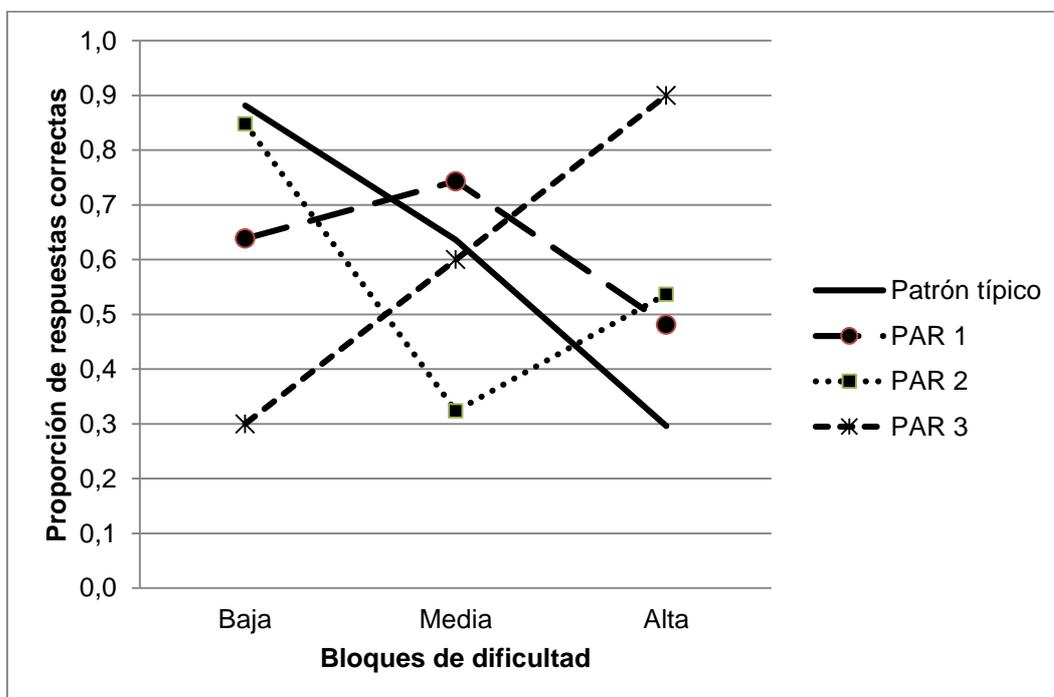




5. Resultados y/o conclusiones

Un patrón no atípico debería tener un perfil descendente o plano, pero nunca creciente en ninguna de sus partes. Así pues, los PAR estarán asociados a una tendencia de respuestas correctas totalmente ascendente o a una tendencia creciente entre los bloques B-M o entre los bloques M-A. La Figura 1 muestra estas tres tendencias de respuesta prototípicas para los PAR. En relación con el patrón típico, claramente descendente en este caso, el patrón atípico identificado con el número 1 tiene menor proporción de respuestas correctas de lo esperado en el bloque de preguntas fáciles, el patrón atípico 2 tiene una proporción de respuestas correctas más baja de lo esperado en el bloque de preguntas de dificultad media, y el patrón atípico 3 muestra un perfil completamente inesperado al responder mejor a los bloques de mayor dificultad.

Figura 1: Patrones de respuestas, típico y atípicos, definidos por la proporción de respuestas correctas a los tres bloques de preguntas definidas por su nivel de dificultad.



Organizado por:





6. Contribuciones y significación científica de este trabajo:

La identificación de los PAR es necesaria para realizar evaluaciones más válidas de los conocimientos (AERA, APA y NCME, 2014). En caso de presencia de PAR, la identificación de las causas de atipicidad en las respuestas es imprescindible para ir más allá de una evaluación puramente sumativa y aprovechar la información que ofrece la manera en que se ha respondido a la prueba para realizar evaluaciones formativas. Por su sencillez conceptual y de aplicación, pretendemos que la metodología propuesta resulte de utilidad a los docentes que utilizan pruebas tipo test como método de evaluación.

7. Bibliografía

- AERA, APA y NCME (2014). *The Standards for Educational and Psychological Testing*. Washington, DC: AERA.
- Haladyna, T.M., y Rodriguez, M.C. (2013). *Developing and validating test items*. New York, NY: Routledge
- Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*, 18, 133–46.
- Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education* 16, 277–298.
- Lane, S., Haladyna, T.M. y Raymond, M. (2016). *Handbook of test development* (2nd ed.). New York, NY: Routledge.
- Meijer, R.R. and Sitjima, K. (2001). Methodology Review: Evaluating Person Fit. *Applied Psychological Measurement*, 25(2), pp. 107–135.
- Moreno, R., Martínez, R.J. y Muñoz, J. (2015). Guidelines based on validity criteria for the development of multiple choice items. *Psicothema*, 27(4), 388-394.
- Tendeiro, J.N. (2015). Package 'PerFit' [Programa informático]. University of Groningen. Disponible en <http://cran.r-project.org/web/packages/PerFit>

Organizado por:

