Pruebas tipo test como instrumentos de evaluación diagnóstica y formativa

Riba, M.Dolors
Universitat Autònoma de Barcelona
Departament de Psicobiologia i de Metodologia de les Ciències de la Salut
Facultat de Psicologia
Edifici B, Carrer de la Fortuna, 08193 Bellaterra (Cerdanyola del Vallés)
Dolors.Riba@uab.cat

Doval, Eduardo
Universitat Autònoma de Barcelona
Departament de Psicobiologia i de Metodologia de les Ciències de la Salut
Facultat de Psicologia
Edifici B, Carrer de la Fortuna, 08193 Bellaterra (Cerdanyola del Vallés)
Eduardo.Doval@uab.cat

Fauquet, Jordi
Universitat Autònoma de Barcelona
Departament de Psicobiologia i de Metodologia de les Ciències de la Salut
Facultat de Psicologia
Edifici B, Carrer de la Fortuna, 08193 Bellaterra (Cerdanyola del Vallés)
Jordi.Fauquet@uab.cat

1. **RESUMEN:** Se propone un método para identificar patrones atípicos de respuesta (PAR) en pruebas tipo test. Se analizan los resultados de una evidencia de evaluación de una asignatura del grado de Psicología de la UAB aplicada durante el curso 2015-16 a 417 estudiantes. Se identifican cuatro tipos diferentes de PAR (25% de los estudiantes). La identificación de PAR proporciona información del proceso de aprendizaje de utilidad en evaluaciones diagnósticas y formativas.

2. **ABSTRACT:** A method for identifying atypical patterns of response (ARP) on multiple choice tests is proposed. Results of a test of the degree of Psychology of the UAB applied during the course 2015-16 to 417 students are analyzed. Four different types of ARP (25 % of students) are identified. ARP provides information of the learning process useful in diagnostic and formative assessments.

ISSN: 2385-6203

3. **PALABRAS CLAVE:** Pruebas tipo test, patrones atípicos de respuesta, evaluación diagnóstica, evaluación formativa.

KEYWORDS: Assessment, Multiple-Choice tests,Aberrant Response Patterns, Educational Assessment, Formative Assessment.

4. **DESARROLLO:**

Introducción

Las pruebas con ítems de elección múltiple, o tipo test, han constituido durante años uno de los métodos tradicionales de evaluación en el ámbito de la educación superior. A raíz de la adaptación de las titulaciones al Espacio Europeo de Educación Superior (EEES) el repertorio de formas de evaluación ha ido creciendo considerablemente y métodos como el portafolio o la evaluación entre pares, por indicar sólo dos, han ganado protagonismo en el terreno de la evaluación. Como consecuencia, las pruebas tipo test se han visto desplazadas e incluso descartadas como pruebas de evaluación válidas para los requisitos que marca el EEES, de manera que cuando se utilizan los docentes suelen mostrar cierto complejo que refleja perfectamente Gil Maciá (2013): “Las pruebas tipo test quizás no sean las más originales para la evaluación continua, quizás tampoco sean la forma más fiel de comulgar con el Plan Bolonia, y probablemente tampoco sean la forma más fidedigna de poder evaluar los conocimientos del alumnado (…) Pero, con todo, dados los medios y recursos de los que disponemos no se nos ocurrió ninguna otra alternativa mejor”.

Fierro (2000, p 294) señala las principales ventajas de las pruebas tipo test al afirmar que, “si los ítems están bien formulados (lo que no es fácil, pero tampoco imposible), en pruebas de este género parecen quedar a salvo la objetividad e imparcialidad, así como también la economía de tiempo de corrección cuando son muchos los examinados”. La objetividad e imparcialidad se garantizan con un buen diseño de la estructura de la prueba (Lane, Haladyna y Raymond, 2016) ycon enunciados y opciones de respuesta redactados correctamente (Haladyna y Rodríguez, 2013; Moreno, Martínez y Muñiz, 2015). Una vez conseguida una prueba con estas características, la facilidad en la administración y corrección la convierten en casi insustituible cuando se tienen que evaluar grupos numerosos (Sánchez Santamaría, 2011).

A pesar de lo que a menudo se ha señalado como una limitación, las preguntas de elección múltiple permiten evaluar niveles competenciales superiores al memorístico, como la capacidad de aplicar conocimientos, e incluso de evaluar y analizar información (Rodriguez, 2016). El nivel de competencia evaluada no lo impone tanto el tipo de prueba como la manera en que ha sido diseñada.

Si este tipo de pruebas presentan claras ventajas y no tantos inconvenientes como parece, ¿por qué, entonces, hay ciertas reticencias a utilizarlas en los nuevos contextos evaluativos? El

modelo que propone Sánchez Santamaría (2011) para entender la evaluación de los aprendizajes universitarios en el EEES nos proporciona una posible respuesta a esta pregunta. En él se definen tres ejes bipolares. Uno de ellos contrapone los conocimientos a las competencias, otro la evaluación sumativa a la formativa y un tercero la evaluación finalista a la evaluación continua. Los primeros elementos de estas dimensiones se asocian a la educación tradicional y a evaluaciones superficiales, y los segundos a las propuestas del EEES y a evaluaciones más auténticas. Por asociación, las pruebas tipo test se vinculan a métodos tradicionales de evaluación, y por tanto alejadas de la posibilidad de ofrecer informaciones útiles para evaluar el proceso de enseñanza-aprendizaje.

Si la aplicación de este tipo de pruebas se acompaña de un análisis individualizado de las respuestas, tanto de los aciertos como de los errores junto con su localización a lo largo de la prueba, podemos acercarnos al proceso que ha seguido el estudiante para dar sus respuestas. Este tipo de evidencia, en forma de patrones de respuesta, permite complementar la que proporciona la puntuación obtenida como suma de respuestas correctas.

Objetivo

Presentamos un método de análisis de las respuestas a preguntas tipo test con el objetivo de identificar diferentes formas de contestar a la prueba. Se centra, no en la puntuación final, si no en el proceso que cada estudiante ha seguido para llegar a la misma (AREA, APA, NCME, 2014), por lo que pretende ser de utilidad en evaluaciones diagnósticas y formativas.

Método

Tras ordenar las preguntas de la prueba en función de su índice de dificultad (proporción de respuestas correctas), éstas se distribuyen en tres categorías o bloques, fáciles (F), de dificultad moderada (M) y difíciles (D), incluyendo cada una de ellas alrededor del 33% de las preguntas de la prueba.

Posteriormente, se calcula para cada estudiante el índice de precaución modificado (IPM) de Harnisch y Linn (1981), que permite señalar la existencia de patrones atípicos de respuesta (PAR). El IPM se calcula de la siguiente manera:

$$IPM= [cov(x^*,p) - cov(x,p)] / [cov(x^*,p) - cov(x',p)]$$

Donde, estando las preguntas ordenadas de más fácil a más difícil, p es el vector de dificultad de las preguntas, X es el patrón observado de respuestas (1=aciertos y 0=errores), X* es el patrón, con la misma puntuación total que X pero que acumula todas las respuestas correctas en los ítems más fáciles y X' es el patrón que acumula todas las respuestas correctas en los

ISSN: 2385-6203

ítems más difíciles. A las pautas observadas X que coinciden con los patrones X* les corresponde un IPM=0. Dichas pautas se conocen como patrones Guttman perfectos. A las pautas observadas X que coinciden con los patrones X' les corresponde un IPM=1. Dichas pautas se conocen como patrones anti-Guttman. Son patrones de respuesta totalmente atípicos. En nuestro estudio se considera un patrón atípico si tiene un valor de IPM≥0.35.

Para profundizar en los casos identificados como atípicos, se han definido cuatro perfiles de respuesta a partir de la comparación de los porcentajes de respuestas correctas de cada categoría de dificultad con su adyacente. Una persona cuyas respuestas se distribuyan de la forma esperada tendría un patrón no ascendente, es decir, más preguntas correctas (o igual) en la categoría fácil que en la moderada y más en ésta (o igual) que en difícil. En el caso concreto de alumnos con todas las respuestas correctas, el patrón sería plano. En general, este perfil de respuesta (perfil 1) está asociado a valores de IPM bajos.

Un perfil ascendente (perfil 2) es, claramente, contrario a lo esperado puesto que se contestan correctamente más preguntas difíciles que fáciles. En otros casos, puede darse un perfil descendente entre las dos primeras categorías pero ascendente entre las dos últimas (perfil 3) o bien ascendente entre las dos primeras y descendente entre las dos últimas (perfil 4).

Este método se ha aplicado a una prueba tipo test de la asignatura Análisis de Datos del grado de Psicología de la UAB realizada en el curso académico 2015-16. La asignatura tiene 6 créditos y se imparte durante el primer semestre del segundo curso. Para la evaluación de la asignatura se programan 6 evidencias, dos de ellas presenciales con pruebas tipo test. Las respuestas analizadas corresponden a la primera evidencia tipo test a la que se presentaron 417 estudiantes de los 431 matriculados. La prueba estaba formada por 31 preguntas con cuatro opciones de respuesta. Ninguna de las preguntas exigía una respuesta basada en el conocimiento memorístico, en cambio sí se evaluaba la capacidad de comprensión, aplicación de conocimientos y análisis e interpretación de resultados. Para la ejecución de la prueba, los estudiantes podían disponer de material de apoyo elaborado por ellos mismos.

Resultados

La figura 1 muestra la proporción de respuestas correctas a las preguntas de la prueba ordenadas por dificultad. En ella se pueden observar las tres categorías de preguntas identificadas por su nivel de dificultad (F y D con 10 preguntas, y M con 11).

De acuerdo con las propiedades esperadas para el IPM, sus valores son independientes de la puntuación obtenida en la prueba (r=-,005). Tras la aplicación del punto de corte adoptado(IPM≥0,35), 315 estudiantes (el 75,5% de los casos) presentan un patrón de respuestas no atípico. Como se observa en la figura 2, en este patrón la proporción de

respuestas correctas desciende a medida que los bloques contienen preguntas más difíciles. En el 24,5% restante se observan todo tipo de perfiles de respuestas atípicas aunque con una distribución variada: el 5,8% de las puntuaciones provienen de un perfil con una proporción de respuestas correctas decreciente con respecto a la dificultad de las preguntas pero con menos respuestas correctas de las que cabría esperar en el bloque F y más preguntas correctas de las esperadas en el bloque M (perfil 1), en el 1% de los casos se responde mejor a preguntas más difíciles que fáciles (perfil 2), en el 7,2% se presenta una menor proporción de respuestas correctas en el bloque M que el bloque D (perfil 3), y en el 10,6% se presenta una menor proporción de respuestas correctas en el bloque F que en el bloque M (perfil 4).

Una misma puntuación en la prueba puede obtenerse con patrones de respuesta diferentes. A modo de ejemplo, en la tabla 1 se muestran diferentes patrones de respuestas de alumnos que han obtenido 5,81 puntos sobre 10 en la prueba. Con esa puntuación, correspondiente a 18 respuestas correctas a las 31 preguntas de la prueba, el patrón esperado implicaría contestar correctamente el 100% de las preguntas de la categoría F, el 70% de las de la categoría M y ninguna de la categoría D. El caso 92 se aproxima a ese perfil y por ese motivo el valor del IPM no es elevado. Sin embargo, el resto de los casos seleccionados tienen un valor de IPM superior a 0,35, lo que indica una pauta atípica, pero cada uno de ellos corresponde a un perfil de respuesta determinado. Para esta puntuación en concreto no se ha observado ningún perfil del tipo 2.

En el conjunto de los datos analizados, el tipo de perfil no ha guardado relación con la puntuación obtenida en la prueba ($F(3,98)=0,116$, $p=0.950$).

Conclusiones

El método de análisis de las respuestas a preguntas tipo test ha permitido identificar, además del patrón de respuesta no atípico, cuatro tipos de patrones atípicos de respuesta. El patrón 1 se aleja poco de la pauta de referencia, por lo que, aunque detectado como patrón atípico, no parece que en la práctica docente requiera de mucha atención. El patrón 2, en cambio, es totalmente inesperado porque es ascendente, aunque en nuestro estudio ha sido realmente infrecuente. Al igual que ocurre con los patrones 3 y 4, estos tres perfiles atípicos pueden ser característicos de personas que copian algunas respuestas, las de dificultad moderada y/o de alta dificultad, aunque también puede tratarse de personas que siendo capaces de contestar bien las preguntas más difíciles, no responden bien las de menor dificultad, probablemente por algún motivo relacionado con el proceso de enseñanza-aprendizaje (por ejemplo, porque han centrado más su interés en aspectos complejos y han desatendido los más sencillos).

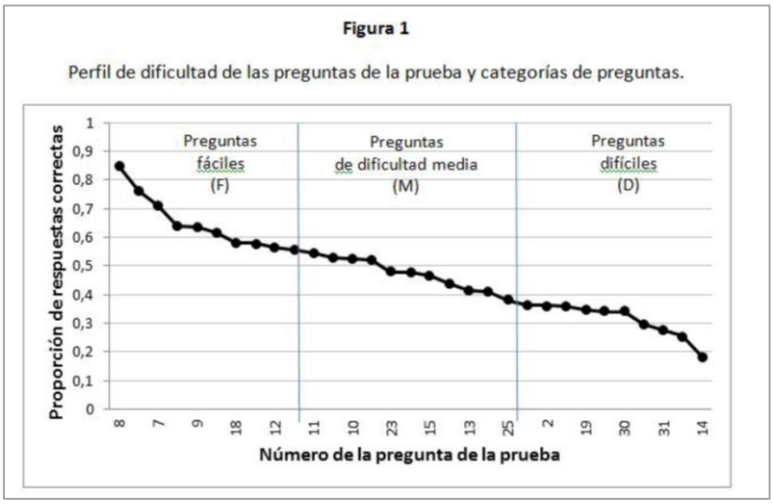
El análisis de los patrones de respuesta, y la detección de patrones atípicos, no es una práctica habitual, aunque desde luego, consideramos igual que Haladyna y Rodriguez (2013) que es

muy recomendable. En primer lugar porque los patrones atípicos ponen en duda la validez de la puntuación de la prueba. Desde luego, parece razonable preguntarse si los cuatro casos que se muestran en la Tabla 1 demuestran tener el mismo nivel de competencias aunque hayan obtenido la misma puntuación. En segundo lugar, porque proporciona una evidencia del proceso de respuesta seguido por la persona que contesta a los ítems, y esta evidencia puede ser de mucha utilidad para identificar y corregir carencias o errores en el proceso de enseñanza-aprendizaje y que en la actualidad no está incorporado en el sistema de evaluación. En este sentido, el método propuesto resulta adecuado para realizar evaluaciones diagnósticas y formativas, aunque deba complementarse con entrevistas con los estudiantes para intentar averiguar las causas de dichos perfiles de respuestas atípicos y de esa forma poder incidir sobre ellas. Por último, aunque no menos importante, el método propuesto puede ser aplicado a la evaluación de competencias no necesariamente memorísticas, en evaluaciones continuas y en grupos numerosos.

4.1. FIGURA O IMAGEN 1

Figura 1

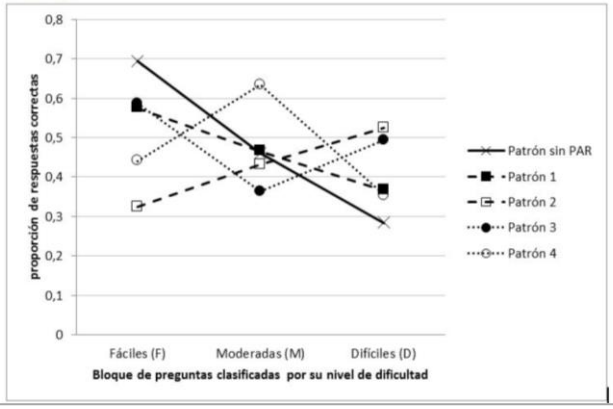
Perfil de dificultad de las preguntas de la prueba y categorías de preguntas.



4.2. FIGURA O IMAGEN 2

Figura 2

Patrones de respuestas a los ítems identificados a partir de la proporción de respuestas correctas promedio en cada bloque de preguntas definidas por su nivel de dificultad.



4.3. FIGURA O IMAGEN 3

Tabla 1

Casos identificados con la misma puntuación en la prueba, 5,81, y diferentes patrones de respuestas correctas.

caso	IPM	Atípico	Proporción de respuestas correctas			Perfil
			Bloque Fácil	Bloque Moderado	Bloque Difícil	
ideal	0	No	1	0,7	0	1
92	0,1467	No	1	0,5	0,2	1
137	0,3501	Sí	0,8	0,5	0,4	1
297	0,4088	Sí	0,6	0,5	0,6	3
384	0,4145	Sí	0,5	0,8	0,4	4

5. REFERENCIAS BIBLIOGRÁFICAS

AERA, APA y NCME (2014). The Standards for Educational and Psychological Testing.Washington, DC: AERA.

Fierro, A. y Fierro-Hernández, C. (2000). Formatos de examen y objetividad en las calificaciones académicas. Revista de educación, 322, 291-304.

Gil Maciá, L. (2013). El sistema de evaluación en el EEES: la experiencia en el área de Sistema Fiscal. XI Jornadas de Redes de Investigación en Docencia Universitaria: Retos de futuro en la enseñanza superior: docencia e investigación para alcanzar la excelencia académica [Recurso electrónico disponible en http://web.ua.es/es/ice/jornadas-redes/documentos/2013-comunicaciones-orales/335028.pdf]

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 EducationalMeasurement, 18, 133 -46.

Lane, S., Haladyna, T.M. y Raymond, M. (2016).Handbook of test development (2nd ed.). New York, NY: Routledge.

Moreno, R., Martínez, R.J. y Muñiz, J. (2015). Guidelines based on validity criteria for the development of multiple choice items. Psicothema, 27(4), 388-394.

Rodriguez, M.C. (2016). Selected-response item formats. In S. Lane, T.M. Haladyna, & M. Raymond (Eds.), Handbook of test development (2nd ed.). New York, NY: Routledge.

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