Revista
Torreón Universitario
www.faremcarazo.unan.edu.ni / ISSN 2410-5708 / e-ISSN 2313-7215
Año 7 | Núm. 18 | pág. 35 - pág. 48 | febrero - mayo 2018

Explicaciones a la presencia de patrones atípicos de respuesta en una prueba de evaluación de conocimientos en la FAREM-Carazo

Dr. Eduardo Doval
Universidad Autònoma de Barcelona, UAB
Especialista en evaluación psicométrica
eduardo.doval@uab.cat

MSc. Pedro Silvio Conrado González
UNAN-Managua, FAREM-Carazo
Máster en Didácticas Especificas con Administración del Currículo
pedrosilvioc081180@hotmail.com

Dra. Marta Fuentes Agustí
Universidad Autònoma de Barcelona, UAB
Especialista en estrategias de enseñanza y aprendizaje
marta.fuentes@uab.cat

Dra. M. Dolors Riba
Universidad Autònoma de Barcelona, UAB
Especialista en análisis estadístico
dolors.riba@uab.cat

Dr. Jordi Renom
Universitat de Barcelona, UB
Especialista en evaluación psicométrica
jrenompinsach@ub.edu

DOI: https://doi.org/10.5377/torreon.v7i18.7715

Palabras Claves: *Patrones atípicos de respuesta, índice de precaución modificado, prueba de evaluación de conocimientos, validez.*

RESUMEN

Las pruebas o exámenes tipo test proporcionan de buena forma los conocimientos académicos adquiridos por los estudiantes. Aunque las pruebas estén diseñadas de forma correcta, con contenidos representativos de los conocimientos que se desean evaluar, los resultados obtenidos por estudiantes que contestan de forma atípica a las preguntas pueden ser indicadores sesgados de sus niveles de conocimientos. Esta posible invalidez de algunas puntuaciones individuales puede estudiarse identificando Patrones Atípicos de Respuesta (PAR).

Sin embargo, la identificación de PAR no aporta información acerca de las causas del mismo. El objetivo de este trabajo es identificar algunas de estas posibles causas. Para ello, se han analizado las respuestas de una misma prueba de 136 estudiantes de tres carreras impartidas en la Facultad Regional Multidisciplinaria de Carazo, de UNAN-Managua (FAREM-Carazo). Veinte y seis de los estudiantes contestaron de forma atípica, gracias a las entrevistas voluntarias realizadas con 16 de ellos; se pudo identificar como posibles explicaciones a la presencia de PAR a falta de estudio y las consecuentes respuestas aleatorias a preguntas consideradas difíciles o incluso, a la presencia de copia de respuesta. Todos estos motivos justifican la duda acerca de la validez de las puntuaciones obtenidas por esos estudiantes que dieron respuestas a la prueba basadas en aspectos diferentes a los de su conocimiento en la materia evaluada.

INTRODUCCIÓN

En el contexto académico, es frecuente realizar la evaluación de conocimientos mediante pruebas de evaluación con un número determinado de preguntas, el mismo para todos los alumnos evaluados. Las pruebas denominadas tipo test, con opciones de respuesta concretas y limitadas, constituye uno de los formatos de pruebas más populares por su relativa facilidad en la elaboración y su objetividad en la corrección. Los docentes que desarrollan este tipo de pruebas están preocupados, y con razón, en que el contenido de las mismas, tanto en lo que se refiere a los enunciados de las preguntas como a las opciones de respuesta, sea relevante respecto a los contenidos evaluados y que estos contenidos estén representados al máximo por las preguntas y respuestas planteadas. Sobre estos aspectos, el docente dispone de ayudas (Haladyna y Rodríguez, 2013; Lane, Haladyna y Raymond, 2016; Moreno, Martínez y Muñiz, 2015), y siguiéndolas, puede validar el contenido de la prueba para realizar una buena evaluación. Sin embargo, con esto no queda garantizada la validez de la prueba, ya que otros aspectos pueden amenazarla. Uno de estos aspectos concierne a la forma en que el alumno evaluado emite sus respuestas.

Se supone, que un estudiante ha de responder a un examen única y exclusivamente basándose en el nivel de conocimientos que posee sobre la materia evaluada. Si es así, las puntuaciones resultantes, normalmente son la suma de respuestas correcta y podrán interpretarse en los términos previstos: un estudiante con un nivel alto de conocimientos contestará muchas preguntas correctamente y por tanto, obtendrá una puntuación alta, mientras que en caso contrario, un estudiante con bajo nivel de conocimientos no será capaz de contestar bien muchas preguntas y eso se reflejará en una puntuación baja. Con la misma lógica, cabría esperar que un estudiante que conteste bien algunas preguntas y mal otras, no siga una pauta de respuestas cualquiera, sino que lo que cabría esperar es que contestase correctamente preguntas más fáciles y fallase preguntas más difíciles. Esto plantea un dilema, cuando un estudiante contesta de una forma ilógica desde este punto de vista, por ejemplo, acertando las preguntas más difíciles y

fallando las más fáciles. De esta manera, dos estudiantes podrían tener la misma puntuación (por ejemplo, 5 puntos obtenidos en una prueba de 10 preguntas), pero uno contestando bien las cinco preguntas más fáciles y el otro, las cinco más difíciles. Diferencias como la planteada, generan una serie de dudas acerca de la validez de las puntuaciones de la prueba: dado que los dos estudiantes han obtenido la misma puntuación, ¿se puede deducir que han adquirido el mismo nivel de conocimientos? ¿Qué es lo que explica que un estudiante que es capaz de contestar preguntas de alto nivel de dificultad no muestre competencia a la hora de contestar preguntas muy fáciles?

La primera pregunta puede contestarse mediante el análisis de lo que se conoce como Patrones Atípicos de Respuesta, o formas inesperadas de responder a la prueba como la planteada con base al nivel de dificultad de sus preguntas. Para identificar la presencia de PAR se han desarrollado numerosos índices (Karabatsos, 2003; Meijer y Sitjsma, 2001). La mera identificación de PAR, sin embargo, no permite justificar su presencia, y por eso, es necesario indagar sobre los motivos con otros procedimientos complementarios, como, por ejemplo, preguntando directamente a los alumnos sobre la forma en que han contestado al examen (Petridou y Williams, 2010), algo necesario para dar respuesta a la segunda de las preguntas planteadas.

Con este estudio se pretende encontrar explicación a la presencia de patrones atípicos de respuesta en una prueba de evaluación de conocimientos académicos.

MÉTODO

Sujetos

De 136 alumnos de FAREM-Carazo evaluados, 46 son de primer año de las carreras de Administración Turística y Hotelera, 45 de Ciencias de la educación con mención en Física Matemática, 45 de Ciencias de la educación con mención en Lengua y Literatura. Todos ellos cursaban la asignatura Geografía e Historia de Nicaragua que se imparte en el primer año de esas carreras. La prueba administrada evaluaba conocimientos de las materias impartidas en las dos primeras unidades de la asignatura: Unidad I. Introducción al estudio de la Geografía y la Historia de Nicaragua para la formación ciudadana y profesional y la Unidad II. Identidades territoriales e identidades culturales de Nicaragua.

Instrumentos

La prueba de evaluación estaba formada por 30 preguntas; 15 de elección múltiple, con cuatro alternativas de respuesta, y 15 de respuesta verdadero/falso. La puntuación máxima en la prueba era 25 puntos, que correspondían al 25 % de la calificación correspondiente al acumulado previo al examen final. El otro 75 % se consiguió con otra prueba corta y 2 trabajos escritos.

La pregunta 21 presentó un problema y fue retirada del examen, por lo que para el análisis fueron examinadas 29 preguntas.

Con posterioridad a la realización de la prueba de evaluación, se entrevistó a algunos alumnos. El guion de la entrevista incluía las siguientes preguntas.

1. ¿Cuál fue la estrategia didáctica para resolver el examen?
2. ¿Considera que la preparación previa al examen fue suficiente para poderlo resolver?
3. Al momento de contestar el examen ¿Cuáles fueron las principales dificultades que encontró en el mismo?
4. Después de la lectura del examen, ¿al momento de responder contestó alguna pregunta al azar?
5. Después de la lectura del examen, ¿al momento de responder copió alguna respuesta de sus compañeros?
6. ¿Considera que la extensión del examen era adecuada para evaluar los contenidos?
7. ¿Considera que el tiempo asignado para responder el examen era suficiente para responder todas las preguntas?

Procedimiento

La prueba de evaluación se realizó en tiempo y hora previstos por la Facultad. Las respuestas de los estudiantes a las preguntas de la prueba se codificaron como aciertos (1) y errores (0). Se analizaron las respuestas individuales a la prueba con el fin de identificar patrones atípicos de respuesta. Esta identificación se realizó de dos maneras: calculando el Índice de Precaución Modificado (IPM) (Harnish y Linn, 1981) y comparando el perfil de aciertos observados (O) con el que cabría esperar según el modelo (M) determinista de Guttman (Doval y Riba, 2016; Doval, Riba, García-Rueda y Renom, 2016; Riba, Doval, Renom y Fuentes, 2017).

En ambos índices, la referencia de patrón de respuestas correctas es el modelo de Guttman (1950). Este modelo determina que en una prueba de K preguntas, una persona que obtenga una puntuación X (siendo X<K), debería haber contestado correctamente los X ítems más fáciles y contestar incorrectamente los K-X ítems más difíciles.

El IPM compara las pautas de respuestas observadas con el patrón Guttman perfecto (contestar correctamente a las K preguntas más fáciles) y con el patrón Guttman inverso (contestar correctamente a las K preguntas más difíciles), todas ellas ponderadas por las dificultades de las preguntas. Proporciona valores entre 0 (patrón de respuestas esperado) y 1 (patrón de respuesta completamente contrario al esperado). En este estudio se calculó con

el paquete *Perfit* de R (Tendeiro, 2015) y se consideraron los posibles indicadores de PAR y los valores de IPM igual o superiores a 0.30 (Karabatsos, 2003).

El perfil de aciertos (Doval y Riba, 2016; Doval, Riba, García-Rueda y Renom, 2016) se obtiene como sigue. Las preguntas se dividen, según el centil asignado a su índice de dificultad, en tres grupos: dificultad baja (centil igual o inferior a 33), dificultad media (centil entre 33 y 66) y dificultad alta (centil superior a 66). A continuación, se calcula, dentro de cada bloque, el porcentaje de preguntas contestadas correctamente. La representación gráfica de dichos porcentajes es el perfil de aciertos observados (O: ver figura 1). Por otra parte, se calcula el porcentaje de respuestas correctas en cada bloque teniendo en cuenta el modelo de Guttman (1950). Concretamente, en una prueba de 30 preguntas (10 de dificultad baja, 10 de dificultad media y 10 de dificultad alta), una persona que haya contestado correctamente 15 preguntas, según el modelo de Guttman debería haber contestado los 10 ítems más fáciles (100 % del bloque de dificultad baja) y también los 5 ítems siguientes en dificultad (50 % del bloque de dificultad media) y contestar incorrectamente los 10 ítems más difíciles (0 % del bloque de dificultad alta).

La representación gráfica de estos porcentajes conforma el perfil de aciertos según el modelo (M: ver figura 1). La distancia euclídea entre los perfiles O y M se utilizó como indicador de la presencia de PAR (Riba, Doval, Renom y Fuentes, 2017). Una distancia euclídea igual o superior a 0.50 fue considerada indicador de posible presencia de PAR.

Las pautas que cumplían los dos criterios anteriores (IPM>0.30 y distancia euclídea>0.5) fueron consideradas PAR.

Con el fin de profundizar en los motivos de la presencia de PAR, se pidió a los alumnos que, de forma voluntaria y sin consecuencias en el resultado de la evaluación, asistieran a una entrevista individual con el profesor de la asignatura.

El tipo de PAR fue identificado comparando, mediante diferencia, el perfil observado (O) y el perfil del modelo (M), lo que proporciona un nuevo perfil (O-M) que ilustra el desvío de las respuestas observadas respecto a las modelizadas. A la derecha de la figura 1 se muestra el perfil O-M resultante de comparar los perfiles O y M que se muestran a la izquierda. El caso representado como muestra un perfil de desvío relevante en los bloques de dificultades baja y media (menos respuestas correctas que las modelizadas) y en el de dificultad alta (más respuestas correctas que las modelizadas).

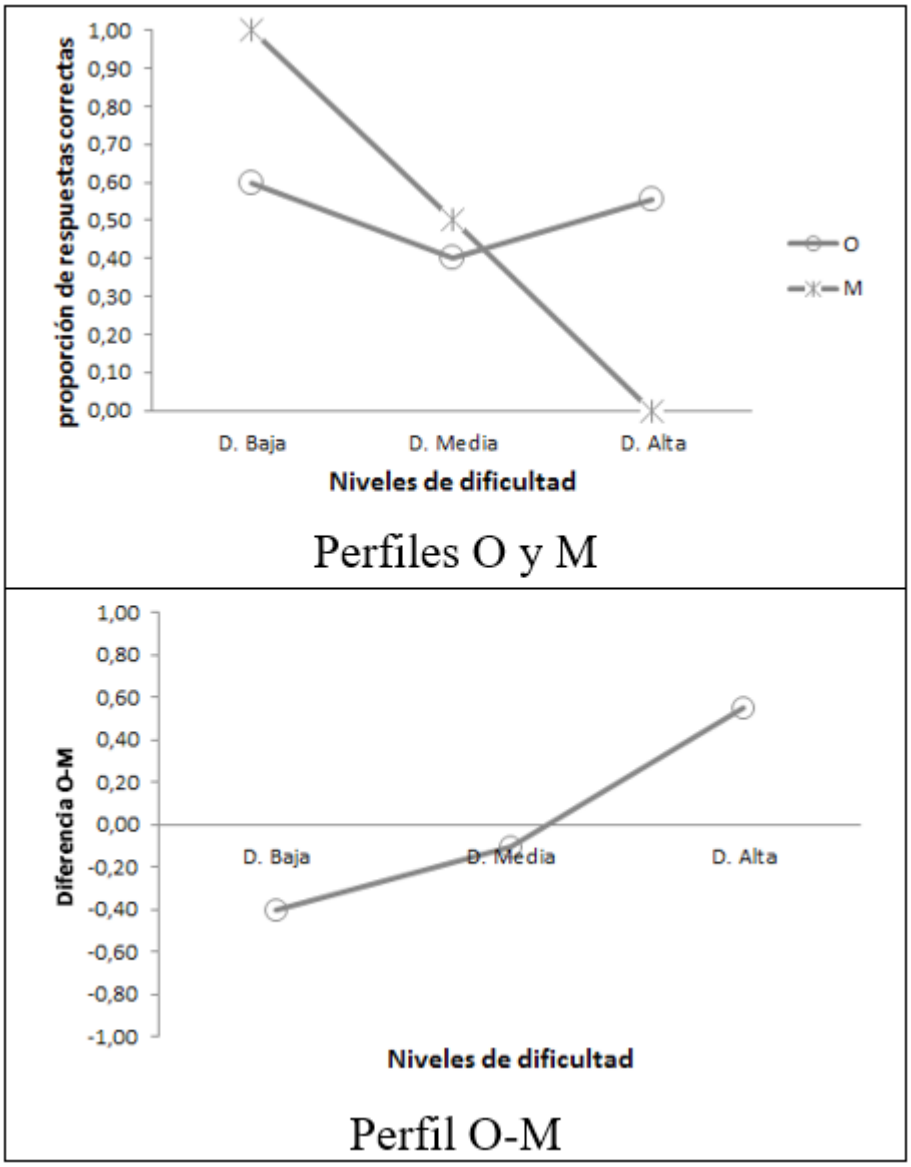


Figura 1. Perfiles de aciertos observados (O) y modelizado (M) y perfil diferencia (O-M)

RESULTADOS

El perfil de dificultad del examen puede verse en la figura 2. Las 10 preguntas más fáciles conformaron el bloque de preguntas de dificultad baja, las 10 siguientes el bloque de preguntas de dificultad media y las 9 más difíciles, el bloque de preguntas de dificultad alta.

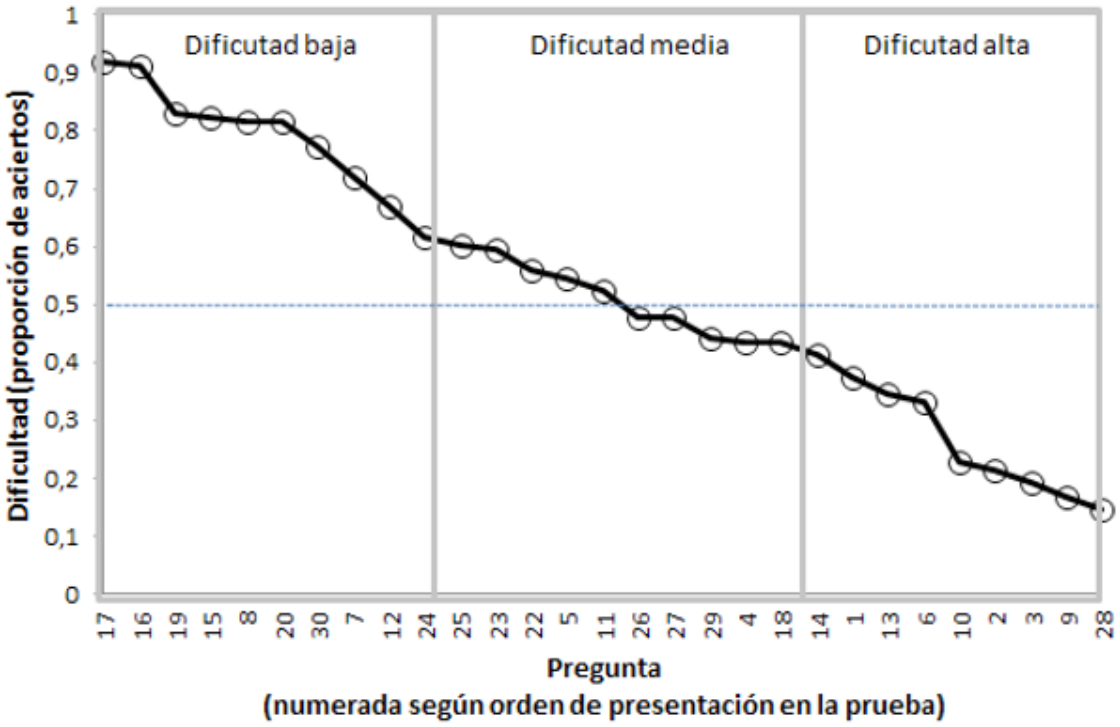


Figura 2. Perfil de dificultad de la prueba

La distribución de puntuaciones (min=5, máx=21, M=15,4, DE=2.92, Asimetría=-.50) se muestra en la figura 3. La distribución de los índices IPM (M=.48, DE=.16, Asimetría=.29) y distancia euclídea entre perfiles (M=.22, DE=.08, Asimetría=.35) muestra que la mayoría de las pautas de respuesta no fueron atípicas. Fueron identificados 28 PAR, un 20,6 % del total de patrones de respuestas. La distribución de puntuaciones de esos alumnos es similar a la del conjunto (min=8, máx=21, M=16, DE=3.22, Asimetría= -.56).

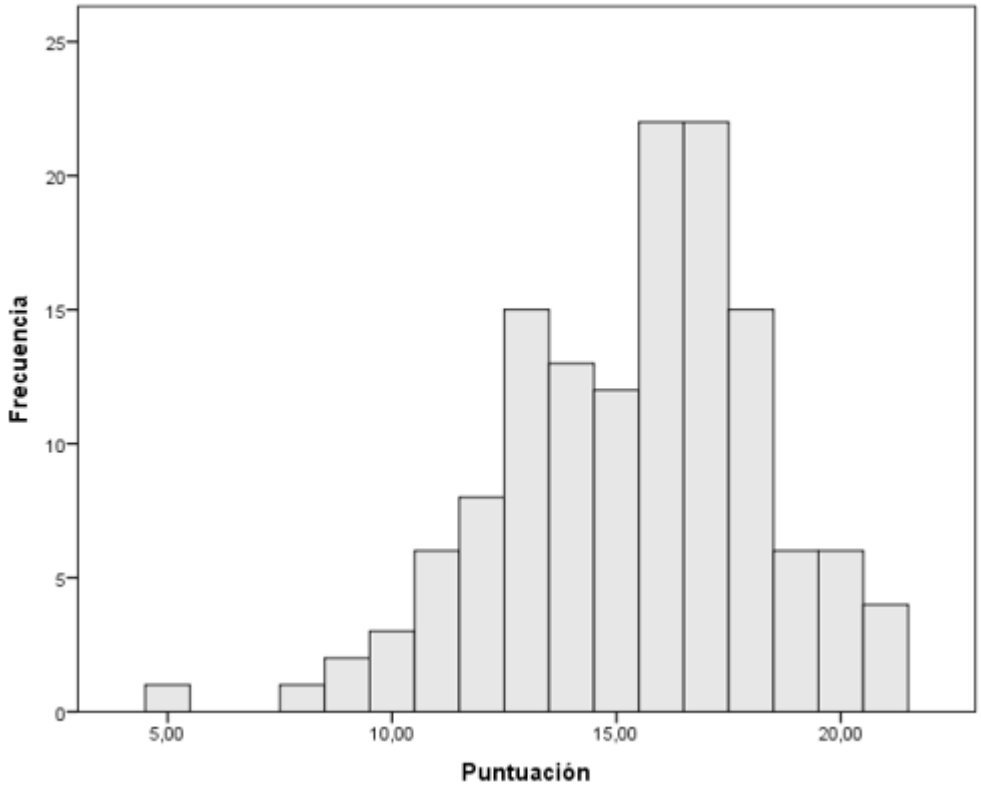


Figura 3. Distribución de puntuaciones en la prueba

De los 28 alumnos que contestaron de forma atípica, 16 aceptaron ser entrevistados por el profesor. Diez de estos alumnos (62,5 %) indicaron que el tamaño de la prueba era adecuado, y el resto (37,5 %) opinaron que era muy larga, pero casi todos ellos (87,5 %) consideraron que el tiempo que tuvieron para realizar la prueba era suficiente y solo dos (12,5 %) dijeron que habían tenido que contestar rápido. Estos dos alumnos también opinaron que la prueba era larga.

Del total de los alumnos, 9 (56,25 %) indicaron que habían estudiado poco, cuatro alumnos (25 %) dijeron haber estudiado parcialmente, otros dos (12,5 %) dijeron haber estudiado lo necesario y solo uno (6,25 %) afirmó haber estudiado mucho.

Con respecto a la prueba, siete alumnos (43,75 %) indicó que las preguntas le resultaron confusas, tres (18,75 %) dijeron que las preguntas eran difíciles y otros tres (18,75 %), que el problema era que no habían estudiado suficiente. Dos estudiantes (12,5 %) opinaron que la prueba tenía demasiadas preguntas y solo uno (6,25 %) dijo que consideraba que la prueba era normal en dificultad y longitud.

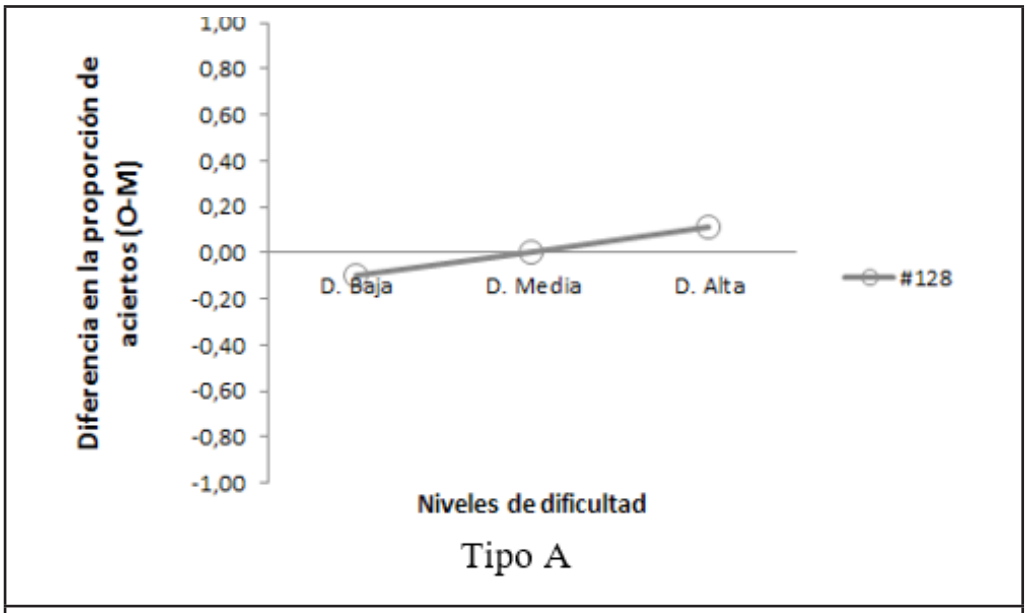
Con respecto a las estrategias seguidas para responder las preguntas, 9 (56,25 %) de los alumnos contestó primero las preguntas que consideraron más fáciles y dejar para el final las más difíciles; 4 dijo contestar las preguntas de corrido, respetando su orden de presentación (25 %). Un estudiante (6,25 %) indicó que sus respuestas se basaron en la lectura y análisis detallado de las preguntas. Dos alumnos (12,5 %) afirmaron que en la mayoría de las preguntas buscaron

claves o pistas para elegir la respuesta. Más de la mitad de los estudiantes evaluados dijeron haber contestado algunas preguntas al azar (11: 68,75 %). Dos alumnos (12,5 %) afirmaron haber copiado algunas preguntas.

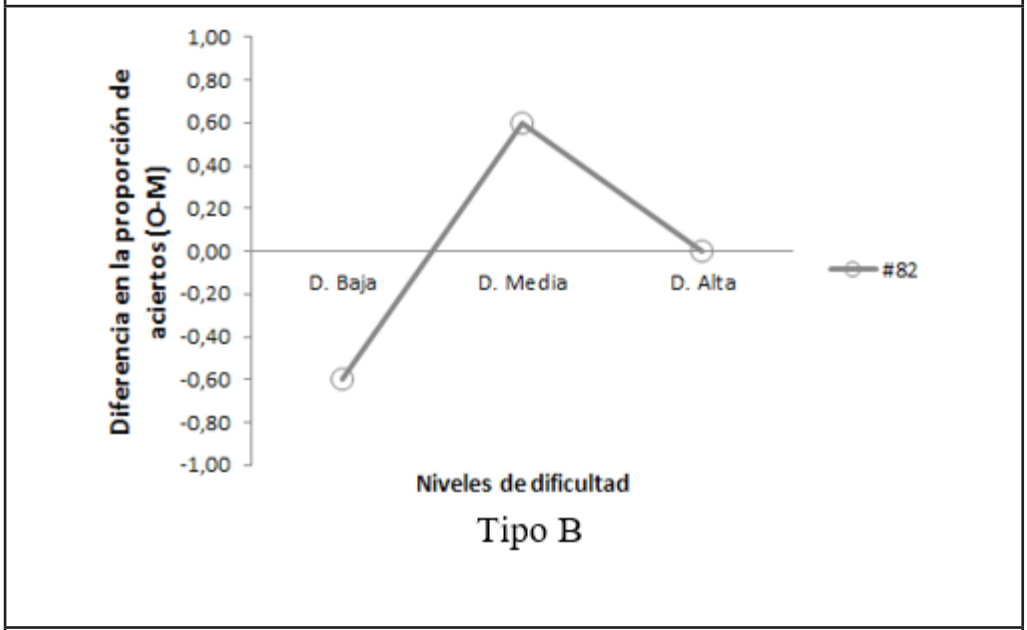
Considerando en conjunto los tres aspectos (dominio/no dominio, respuestas al azar/no azar y copia/no copia), el perfil más frecuente ha sido el de los estudiantes que afirmaron no tener dominio y contestar al azar a algunas preguntas (6: 37.5 %), seguido por el de los que dijeron dominar la materia, pero, aun así, haber contestado algunas preguntas al azar (4: 25 %) y los que dijeron no dominar la materia, aunque indicaron que no contestaron a las preguntas de manera inválida (3:18.75 %). De manera más residual, dos estudiantes (12.5 %) afirmaron no dominar la materia, y uno de ellos dijo haber copiado algunas respuestas y otro, además, haber contestado algunas preguntas al azar.

La figura 4 contiene diferentes ejemplos de perfiles O-M observados en las respuestas de los alumnos evaluados. El perfil A corresponde a un alumno sin PAR. Como se observa, el desvío entre el porcentaje de respuestas correctas observadas y modelizadas es pequeño o nulo en cada uno de los tres niveles de dificultad. El resto de perfiles pueden ser identificativos de PAR, dependiendo de lo relevante de los desvíos. En el perfil tipo B, las respuestas correctas en el bloque de preguntas fáciles son mucho menos de lo que cabría esperar, mientras que en el bloque de preguntas de dificultad media ocurre lo contrario. En el perfil tipo C, es en las preguntas difíciles en las que se observan más respuestas correctas de lo que habría que esperar, mientras que en el perfil tipo D estas respuestas correctas inesperadas se observan en el bloque de preguntas de dificultad media y alta. Finalmente, en el perfil tipo E, en los bloques de preguntas de dificultad baja y media se observan menos respuestas correctas de lo esperado y, sin embargo, es en el bloque de respuestas más difíciles en el cual se observan, inesperadamente, más respuestas correctas.

El 81,25 % (13) de los 16 alumnos entrevistados respondieron según un patrón de respuestas tipo E y el resto (3: 18,75 %) con un patrón de respuestas tipo D.



Tipo A



Tipo B

Tabla continúa en página siguiente

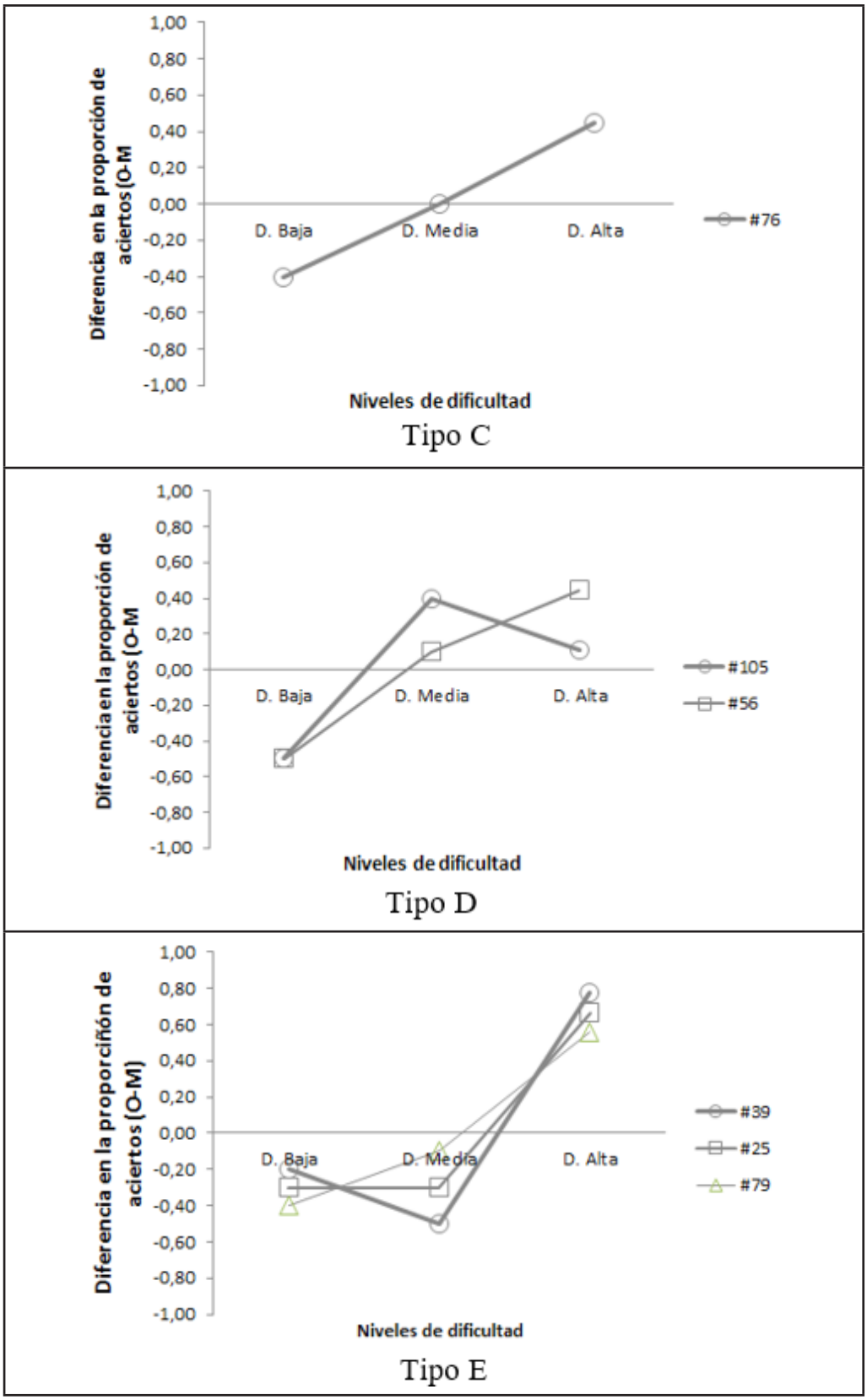


Figura 4. Diferentes tipos de diferencias O-M. Los tipos B, C, D y E pueden representar tipos de PAR

DISCUSIÓN Y CONCLUSIONES

Resulta significativo que la mayoría de los alumnos entrevistados, cuyas pautas de respuesta habían sido previamente identificadas como patrones atípicos, manifestasen haber respondido que no habían contestado a todas las preguntas basándose únicamente en sus conocimientos sobre la materia evaluada. De hecho, también la mayoría de estos alumnos confesaron no haberse preparado suficientemente en la materia. Con este contexto, no es de extrañar que los alumnos entrevistados contestaran mejor a las preguntas más difíciles que a las más fáciles, especialmente cuando ellos mismos afirman que ante el desconocimiento de las respuestas correctas optaron mayoritariamente por contestar al azar, y en menor caso, por copiar las respuestas de algún compañero con más conocimiento.

Contestar al azar es una conducta que, hasta cierto punto, es aceptada en el contexto académico, pero copiar constituye una conducta reprobable. Quizás este hecho justifique que los alumnos entrevistados hayan confesado más respuestas al azar que copia. El procedimiento seguido, que aseguraba a los alumnos que no se tomarían represalias en las calificaciones, no garantiza sin embargo la total confianza en las respuestas en las entrevistas. A pesar de ello, de las justificaciones recogidas en las entrevistas se ha obtenido suficiente evidencia que las puntuaciones obtenidas por los alumnos identificados con PAR, no son válidas para identificar el verdadero nivel de conocimientos adquiridos en la materia. Y, aunque la presencia de PAR puede indicar una infravaloración o sobrevaloración del nivel de conocimiento, los motivos aducidos por los estudiantes entrevistados indican, que, en este caso, las puntuaciones obtenidas por la gran mayoría de ellos sobrevaloraron sus conocimientos. Solo en un caso, el alumno manifestó haberse preparado bien en la materia y consideró, que la prueba no era difícil y que el tiempo para contestarla era adecuado. Su puntuación fue elevada (19 puntos) y a pesar de ello, contestó de manera atípica a las preguntas fáciles pues falló en una respuesta de ellas y en cambio, contestó más preguntas correctas en el bloque de preguntas difíciles que de dificultad media. Si realmente había estudiado la materia, es posible que el problema en este caso fuese que era capaz de contestar preguntas difíciles y, por tanto, el problema sería averiguar por qué no contestó correctamente más preguntas de dificultad media. Es posible que este sea un caso de una puntuación que infravalore la verdadera capacidad del alumno.

Sea como fuere, el análisis realizado ha permitido detectar a partir de la forma de responder a la prueba, a un grupo de alumnos que han sido evaluados incorrectamente puesto que las inferencias, que sobre su conocimiento se pueden realizar a partir de las puntuaciones obtenidas, tienen una dudosa validez.

Se considera que, la validez de las puntuaciones debe garantizarse en todas las evaluaciones educativas (AERA, APA, NCME, 2014) y que el método descrito puede ser de gran utilidad para

identificar posibles fuentes de invalidez. Además, este método también es aplicable a exámenes con preguntas de respuesta abierta, puesto que su corrección en términos de respuesta correcta o respuesta incorrecta, también define un patrón de respuestas que puede ser analizado. Sin embargo, el análisis de PAR no constituye un método infalible para identificar patrones que invaliden las puntuaciones obtenidas. Por las conclusiones que puedan extraerse de él, deben tomarse con cautela, y a ser posible, antes de que el docente tome medidas al respecto, se asegure de ampliar las evidencias que apunten a una infravaloración o sobrevaloración de las puntuaciones obtenidas en la prueba.

AGRADECIMIENTOS

Esta investigación se ha beneficiado de los estudios realizados mediante la financiación de la Dirección General de Investigación y Gestión del Plan Nacional de I+D+i, del Ministerio de Economía y Competitividad de España (Proyecto EDU2013-41399-P) y los proyectos de cooperación FSXXVIII, AECID Ref. 1/041892/11 y FSXXXIII.

REFERENCIAS

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for educational and psychologicaltesting*. Washington, DC: American Educational Research Association

Doval, E. y Riba, M.D. (2016). Identificación de tipologías de Patrones atípicos de respuesta en pruebas tipo test. *V Congreso internacional multidisciplinar de investigación educativa, CIMIE16*. Sevilla (España).

Doval, E., Riba, M.D., García-Rueda, R. y Renom, J. (2016). Comparison of the capacity of three nonparametric person -fit indexes to detect different aberrant response patterns on real data.*VII European Congress of Methodology*. Palma de Mallorca (España).

Guttman, L.A. (1950). The basis for scalogram analysis. In Stouffer, S.A., Guttman, L.A., y Schuman, E.A., *Measurement and prediction*. Volume 4 of Studies in social psychology in World War II. Princeton: Princeton University Press.

Haladyna, T.M., y Rodríguez, M.C. (2013). *Developing and validating test items*. New York, NY: Routledge

Harnisch, D. L., y 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. y Sitjsma, K. (2001). Methodology Review: Evaluating Person Fit.Applied Psychological Measurement, 25 (2), pp. 107–135.

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.

Petridou, A. y Williams, J. (2010). Accounting for unexpected test responses through examinees' and their teachers' explanations. *Assessment in Education: Principles, Policy & Practice, 17*:4, 357-382.

Riba, M.D., Doval, E., Renom, J. y Fuentes, M. (2017). Propuesta para detectar patrones atípicos de respuestas en contextos reales de evaluación. *XV Congreso de metodología de las ciencias sociales y de la salud*. Barcelona (España).

Tendeiro, J. N. (2015). Package 'Per Fit' [Software]. University of Groningen. Available at http: //cran.r-project.org/web/packages/PerFit