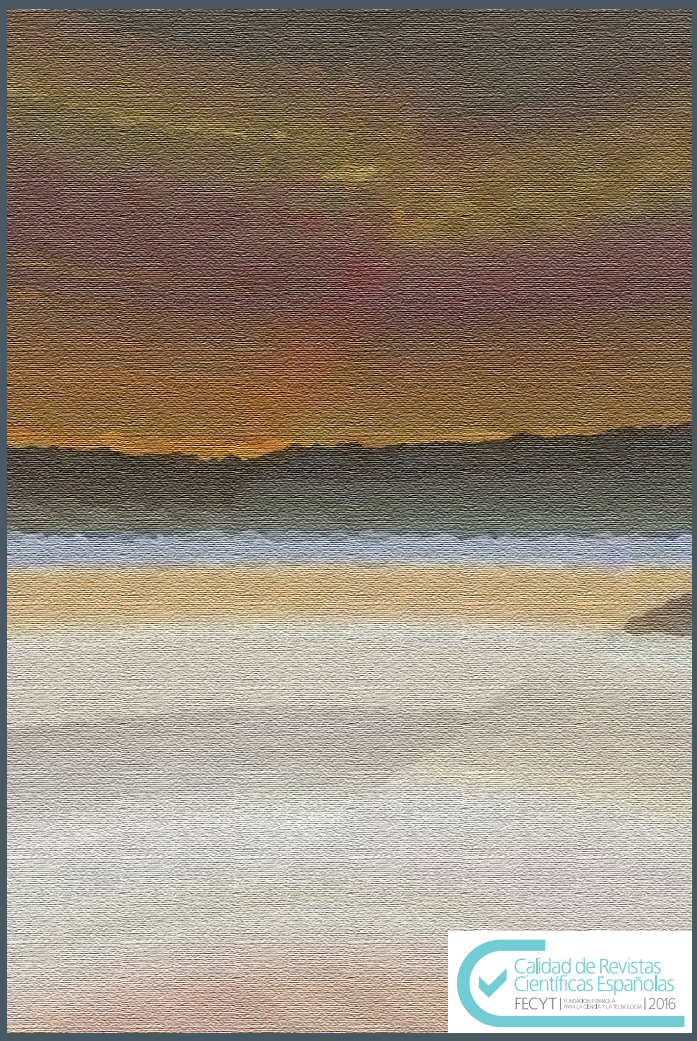
revista de
eDUCACIÓN
Nº 386 OCTUBRE-DICIEMBRE 2019

Las Rúbricas No neutralizan el Efecto de los correctores: Una estimación con el modelo de facetas múltiples de Rasch

Rubrics do not neutralize Raters' effects: A many-faceted Rasch model estimation

Rubén Fernández-Alonso
Pamela Woitschach
José Muñiz



GOBIERNO DE ESPAÑA
MINISTERIO DE EDUCACIÓN Y FORMACIÓN PROFESIONAL

Calidad de Revistas Científicas Españolas
FECYT | FUNDACIÓN ESPAÑOLA PARA LA CIENCIA Y LA TECNOLOGÍA | 2016

Las Rúbricas No neutralizan el Efecto de los correctores: Una estimación con el modelo de facetas múltiples de Rasch

Rubrics do not neutralize Raters' effects: A many-faceted Rasch model estimation

DOI: 10.4438/1988-592X-RE-2019-386-428

Rubén Fernández-Alonso
Pamela Woitschach
José Muñiz
Universidad de Oviedo

Resumen

Los **ítems de** respuesta construida son ampliamente utilizados en todo tipo de evaluaciones educativas. A pesar de la utilización de rúbricas muy específicas para su corrección, la influencia de los correctores está bien documentada en la literatura, afectando a los resultados de la evaluación. El objetivo central del presente trabajo es la estimación de los efectos de los correctores y de las rúbricas en una tarea de expresión escrita. Se utilizaron 13 correctores que revisaron 375 producciones escritas de estudiantes de sexto curso. Los correctores fueron asignados a los ensayos escritos siguiendo un cuadrado Youden de 13 bloques, un diseño de bloques incompletos balanceado. En el análisis de datos se empleó el modelo de Rasch de facetas múltiples con tres facetas: corrector, rúbricas y dificultad de paso de los mismos. Se compararon diferentes modelos y se analizaron los efectos del corrector y las características de las rúbricas. Los resultados ponen de manifiesto las diferencias entre los correctores en cuanto la severidad y la exactitud de sus juicios. Se concluye que no incluir el efecto del corrector en la estimación de los resultados del alumnado puede introducir un componente claro de inequidad en las evaluaciones.

Palabras clave: Evaluación educativa, correctores, rúbricas, Modelos MFRM

Recibido: 16-11-2018 Aceptado: 21-06-2019

Abstract

Constructed response items are widely used in all types of educational evaluations. Despite the use of very specific rubrics for scoring the items, the influence of raters is well documented in the literature, affecting the results of the evaluations. The main goal of the present study is the estimation of the effects of the raters and the rubrics in the assessment of written expression. We used 13 raters that reviewed 375 written productions of sixth grade students. The raters were assigned to the written essays following a 13-block Youden square, a balanced incomplete block design. In the data analysis, the many-faceted Rasch model was used with three facets: corrector, rubrics and difficulty of passing them. Different models were compared and the effects of the raters and the characteristics of the rubrics were analyzed. The results reveal the differences between the raters in terms of the severity and accuracy of their judgments. It is concluded that not considering the effect of the rater in the estimation of student outcomes can introduce a clear component of inequity in the evaluations.

Key words: Educational assessment, rater, scoring rubrics, MRFM models

Introducción

Los sistemas educativos europeos contemplan la realización de exámenes de alto impacto, tales como pruebas de titulación, certificación, acceso a diferentes estudios y evaluación de centros, los cuales incluyen ítems en los que los efectos del corrector pueden influir en la calificación del alumnado (European Commission/EACEA/Eurydice, 2009*)*. Dada la importancia que estas evaluaciones tienen tanto para el propio alumnado, como para la sociedad y los distintos agentes implicados, su objetividad ha sido ampliamente investigada (Congdon y McQueen, 2000; Engelhard, 1992; Gyagenda y Engelhard, 2009; Leckie y Baird, 2011; Lunz y Stahl, 1990; Lunz, Wright y Linacre, 1990; Park, 2010; Wolfe, 2004). Según Eckes (2009) hay dos amplias fuentes de error que afectan a la objetividad de las evaluaciones: distales y proximales.

Los factores de carácter distal se refieren a aspectos tales como las características de los evaluadores, del alumnado y del contexto de la evaluación. En el caso de los evaluadores se han encontrado variaciones asociadas al género, edad, trayectoria profesional y campo disciplinar del corrector, así como a sus actitudes hacia el género, nacionalidad o etnia

Recibido: 16-11-2018 Aceptado: 21-06-2019

de las personas evaluadas (Amengual, 2004; Congdon y McQueen, 2000; Linacre, Engelhard, Tatum y Myford, 1994; Lopes Toffoli, de Andrade y Bornia, 2015; Mahmood, Dagnæs, Bube, Rohrsted y Konge, 2017). También se hallaron diferencias vinculadas al proceso de corrección, tales como experiencia previa, orden de corrección, entrenamiento, familiaridad con las rúbricas y estrategias cognitivas de los correctores (Bejar, 2012; Congdon y McQueen, 2000; Crisp, 2012; Jonsson y Svingby, 2007; Leckie y Baird, 2011; Linacre et al., 1994; McNamara, 1996; Suto, 2012).

Los factores proximales hacen alusión al constructo medido, la dificultad de la tarea y de los criterios de corrección, la estructura de los niveles de calificación y los efectos del corrector (Eckes, 2009). Cuxart-Jardí (2000), Cuxart-Jardí, Martí-Recober y Ferrer-Juliá (1997) encontraron mayor grado de acuerdo en exámenes de materias científico-matemáticas que en materias socio-lingüísticas; por su parte Jonsson y Svingby (2007) informan de diferencias de fiabilidad en función del número de categorías empleadas y de la naturaleza de la tarea, concluyendo que los índices de acuerdo aumentan al disminuir el número de categorías y que las pruebas físicas y la resolución de casos presentan índices de fiabilidad más altos que las producciones orales y escritas; finalmente, Baird, Meadows, Leckie y Caro (2017), Jonsson y Svingby (2007) y Kuo (2007) coinciden al señalar que en las pruebas de alto impacto las rúbricas analíticas son preferibles a las holísticas, aunque para ello es necesario que los procedimientos de puntuación y los criterios de separación entre los niveles de ejecución estén claramente detallados y ejemplificados. De entre los factores proximales el que más atención ha recibido son los efectos del corrector (severidad/permisividad, halo, restricción del rango, consistencia, etc.), que han sido estudiados desde diferentes perspectivas metodológicas, tales como la aproximación clásica (OECD, 2014; Saal, Downey y Lahey, 1980; Woitschach et al., 2018), teoría de la Generalizabilidad (Sudweeks, Reeve y Bradshaw, 2005), análisis multinivel (Congdon y McQueen, 2000; Leckie y Baird, 2011), o modelos de Teoría de Respuesta al Ítem (TRI; Adams y Wu, 2010; Lunz et al., 1990; Myford y Wolfe, 2003, 2004). Si bien cada una de estas estrategias metodológicas tienen sus ventajas e inconvenientes (Baird, Hayes, Johnson, Johnson y Lamprianou, 2013; Gyagenda y Engelhard, 2009; Sudweeks et al, 2005), los modelos derivados de la TRI ofrecen soluciones integrales, ya que permiten validar las rúbricas (Lallmamode, Mat Daud y Abu Kassim,

2016), analizar los efectos de las diversas fuentes de variación (Linacre et al., 1994), y manejar conjuntamente dichos efectos para estimar los resultados del alumnado (Eckes, 2009).

En concreto el modelo de facetas múltiples de Rasch (MFRM, por sus siglas en inglés Many-Facet Rasch Measurement; Linacre et al., 1994) permite parametrizar en una única escala todas las facetas del modelo: competencia del sujeto, características del corrector, dificultad, parámetros de paso de las categorías, y cuantas facetas pueda incluir el análisis (Eckes, 2009; Myford y Wolfe, 2003, 2004; Prieto-Adánez, 2011). Por otra parte, Linacre et al. (1994) y Lunz et al. (1990) mostraron que los MFRM pueden identificar los efectos del corrector (severidad/permisividad, calificaciones inconsistentes, sobre-consistencias…) y diagnosticar los criterios y rúbricas de corrección (ajuste, dimensionalidad de la escala, parámetros de paso…). Con todo, la principal ventaja a nivel práctico de los MFRM es haber demostrado la necesidad de incluir los parámetros del modelo (en especial las variaciones en la severidad de los correctores) en la estimación de los resultados del alumnado en la prueba, ya que en caso contrario esto podría afectar a las decisiones sobre el nivel de competencia del alumnado y, por ende, a la validez de las conclusiones del estudio (Eckes, 2005, 2009; Prieto-Adánez, 2015; Shackleton, 2018).

Dentro de este contexto, el objetivo general del presente trabajo es analizar los efectos introducidos por los correctores en la evaluación de la expresión escrita, y su interacción con las rúbricas utilizadas. Este objetivo general se desglosa en cuatro específicos: (a) Elegir el modelo de Rasch de facetas múltiples que mejor se ajuste a los datos; (b) Comprobar el ajuste de los correctores al modelo y analizar los efectos introducidos por los correctores; (c) Estudiar los efectos de las rúbricas y su ajuste al modelo; y (d) Estimar la cuantía de los sesgos introducidos por los correctores en las puntuaciones del alumnado. Estos objetivos tienen especial interés ya que apenas existen trabajos que estudien las consecuencias que tiene para el alumnado el hecho de que sus producciones sean asignadas a correctores con niveles de severidad diferentes. En España los trabajos realizados hasta ahora se centran en analizar los efectos del corrector y el ajuste de los modelos, sin disponer de una estimación de casos erróneamente clasificados, al no incluir los efectos del juez en el cálculo de la competencia del alumnado. Creemos que en ese sentido nuestro trabajo es pionero, además de aportar novedades metodológicas en el diseño y análisis de la asignación de correctores a la materia evaluada.

Recibido: 16-11-2018 Aceptado: 21-06-2019

Método

Muestra

Se utilizaron 13 correctores que fueron asignados a los textos siguiendo un diseño cuadrado Youden de 13 bloques. Se trata de un cuadrado latino incompleto balanceado que desarrolla las 4 primeras réplicas del cuadrado latino 13 x 13 (Ato y Vallejo, 2007; Cochran y Cox, 1974; Frey, Hartig y Rupp, 2009; Organisation for Economic Co-operation and Development [OECD], 2014). Para lograr este arreglo las expresiones escritas de las 18 aulas se agruparon en 13 tribunales de calificación (bloques en términos de diseño experimental). La tabla I muestra los 13 tribunales y el número de expresiones escritas asignadas a cada uno. De promedio a cada tribunal le correspondieron 28 expresiones escritas, si bien los tribunales TB09 a TB13 corrigieron más ejercicios ya que incluían las producciones de dos grupos-aula.

TABLA I. Composición de los 13 tribunales: aulas asignadas, número de expresiones escritas a corregir y correctores que conforman cada tribunal

Tribunales (Bloques de expresiones escritas)	Aula(s) asignadas al tribunal	N° expresiones a corregir por tribunal	Correctores que conformarán cada tribunal			
TB01	Aula 01	28	C01	C09	C12	C10
TB02	Aula 02	26	C02	C01	C08	C11
TB03	Aula 03	25	C03	C02	C09	C05
TB04	Aula 04	25	C04	C03	C01	C06
TB05	Aula 05	24	C05	C11	C04	C12
TB06	Aula 06	23	C06	C05	C10	C08
TB07	Aula 07	23	C07	C13	C05	C01
TB08	Aula 08	22	C08	C12	C13	C03
TB09	Aula 09 y Aula 18	34	C09	C08	C07	C04
TB10	Aula 10 y Aula 17	36	C10	C04	C02	C13
TB11	Aula 11 y Aula 16	35	C11	C10	C03	C07
TB12	Aula 12 y Aula 15	36	C12	C07	C06	C02
TB13	Aula 13 y Aula 14	37	C13	C06	C11	C09

Nota: Elaboración propia

Las cuatro últimas columnas de la tabla recogen la identificación de los cuatro correctores asignados a cada tribunal y permiten comprobar la consistencia, balanceo y eficiencia del diseño Youden. La consistencia del diseño hace que cada corrector sea asignado a cuatro tribunales y que cada expresión escrita será corregida por cuatro correctores. Por ejemplo, el Corrector 1 (C1) forma parte de los tribunales TB01, TB02, TB04 y TB07 y, de igual modo, las producciones asignadas al TB01 serán corregidas por C1, C9, C12 y C10. El diseño está completamente balanceado ya que cada corrector coincidirá en una única ocasión con el resto de los correctores a la hora de calificar un bloque o grupo de expresiones escritas. Por ejemplo, C1, coincide con C09, C12 y C10 en el tribunal TB01, con C02, C11 y C08 en el TB02, con C03, C04 y C06 en el TB04 y con los tres correctores restantes en el TB07. El resto de las coincidencias pueden apreciarse en la tabla 1. La eficiencia del diseño radica en que, de promedio, cada corrector revisará poco más de 100 expresiones escritas (por ejemplo, a C12 le corresponden las 110 producciones de los TB01, 05, 08 y 12) y no las 375 que serían necesarias si se hubiera empleado el cuadrado latino completo. En el conjunto del estudio se realizarán 1500 correcciones y no las casi 5000 que obligaría el diseño completo, ahorrando aproximadamente el 60% de las correcciones posibles sin que las conclusiones del análisis pierdan validez. En definitiva, esta distribución de correctores a expresiones escritas asegura el doble control propio del cuadrado latino (Fernández-Alonso y Muñiz, 2011) de tal modo que la posible severidad o benevolencia de los jueces no puede ser imputada al hecho de que le fuera asignado un grupo de alta o baja competencia en la expresión escrita.

Las expresiones escritas empleadas en este estudio provienen de la *Evaluación Final de Educación Primaria* (prevista en la Ley Orgánica 8/2013 de Mejora de la Calidad Educativa) realizada en una Comunidad Autónoma en el curso 2015/16 donde se recogieron 6653 ejercicios provenientes de 403 grupos aula de 6º de Educación Primaria. Empleando un muestro sistemático y aleatorio se seleccionaron 18 aulas con una probabilidad proporcional a su tamaño (OECD, 2014), y se revisaron todas las producciones realizadas por el alumnado de los grupos elegidos: en total 375 ejercicios.

Recibido: 16-11-2018 Aceptado: 21-06-2019

Instrumentos

El estímulo empleado en la evaluación era una lectura que contenía información turística de tres ciudades. El alumnado debía elegir una de ellas y escribir un texto para convencer a sus compañeros de que esa ciudad era el mejor destino para el viaje de estudios. La rúbrica de corrección evaluaba tres constructos o procesos cognitivos recogidos en el currículo oficial de España. La tabla II muestra la organización de los procesos y estándares evaluados. Las puntuaciones de los estándares se resumieron en los tres constructos que fueron codificados en una escala de tres niveles con el siguiente significado: 0 puntos = No logrado; 1 = Logrado; y 2 = Consolidado. La prueba, sus especificaciones y rúbricas de corrección están disponibles en Ministerio de Educación, Cultura y Deporte (2016, pp. 97-117) y puede consultarse en: https://sede.educacion.gob.es/publiventa/descarga.action?f_codigo_agc=18314

TABLA II. Constructos o procesos cognitivos evaluados y estándares de la rúbrica de corrección

Constructos o procesos cognitivos	Estándares de la rúbrica
Coherencia	Organizar las ideas con claridad y progresión temática; y expresar opiniones, reflexiones y valoraciones coherentemente.
Cohesión	Usar conectores y signos de puntuación para cohesión al texto; y sustituciones pronominales y sinónimos para evitar reiteraciones.
Adecuación y presentación	Respetar las normas gramaticales y ortográficas; usar un registro adecuado al interlocutor y asunto tratado; y presentar un escrito con limpieza, claridad, precisión y orden.

Nota: Elaboración propia

Procedimiento

Inicialmente se gestionó el permiso para que la Consejería de Educación competente hiciera una cesión parcial del fichero que contenía las

pruebas de las aulas seleccionadas de acuerdo con la Ley Orgánica 15/1999 de Protección de Datos de Carácter Personal. Cada expresión escrita, que originalmente se encontraba en formato lápiz y papel, fue escaneada; anonimizada mediante un código de identificación; editada con el programa Adobe Acrobat Prof DC® para eliminar cualquier marca de la corrección original; y guardada en un fichero nombrado con el código de identificación.

Adicionalmente, se prepararon 13 plantillas de corrección (una por corrector) que contenían la relación de las expresiones escritas asignadas a cada corrector y se diseñó una plataforma online donde se cargaron los 375 ficheros con las expresiones escritas en la que cada corrector solo tenía acceso a sus ficheros asignados. Los correctores calificaron y codificaron en la plataforma online los resultados de su corrección y, una vez descargadas las correcciones, se verificaron los códigos introducidos para descartar errores de transcripción en la plataforma.

El grupo de correctores, estudiantes de postgrado de Educación, recibió seis horas de formación en procedimientos de corrección: dos horas sobre conceptos básicos de corrección mediante rúbricas y cuatro de trabajo con el instrumental de la prueba, incluyendo entrenamientos de corrección con producciones similares que no se emplearían en el estudio, pero que sirvieron para familiarizarse con el material y el procedimiento de corrección.

Análisis de los datos

Se utilizó el MFRM para datos politómicos, el cual asume que la puntuación del alumnado es una función logística que incluye, además de la competencia del estudiante, tres facetas más: corrector (*rater*), procesos cognitivos (*criteria*) y dificultad de dichos procesos (*step*). Ello permite definir un modelo de efectos principales y modelos adicionales de interacción entre facetas y realizar una comparación jerárquica de modelos. Inicialmente se ajustó el modelo de efectos principales, a continuación, tres modelos de interacción de dos facetas (modelos de interacción de segundo nivel: *criteria x step; rater x step; y criteria x rater*); y finamente un modelo que incluyó la interacción de las tres facetas conjuntamente (modelo *criteria x step x rater*). Para comprobar qué modelo presenta el mejor ajuste se empleó la prueba de χ^2,

Recibido: 16-11-2018 Aceptado: 21-06-2019

comparando la reducción de la *deviance* de cada nuevo modelo con el aumento del número de parámetros necesarios para ajustar los modelos de complejidad creciente (Adams y Wu, 2010).

Seleccionado el modelo, y para cumplir con el segundo objetivo del estudio, se comprobó la adecuación de las calificaciones mediante los valores de ajuste de los residuales de las medias cuadráticas (MNSQ) de los correctores, siguiendo los criterios establecidos por Eckes (2005) y Wolfe y McVay (2012). Los efectos del corrector (dificultad, halo y centralidad) se estudiaron mediante la comparación de los estimadores de severidad de los jueces, el análisis del índice de separación entre jueces, y los valores de ajuste MNSQ de los correctores según los criterios de Eckes (2005, 2009), Lunz et al., (1990), Engelhard (1994) y Wolfe y McVay (2012).

La dimensionalidad de la escala, que se basa en el supuesto de la correlación positiva entre las puntuaciones de los tres procesos cognitivos evaluados, se verificó atendiendo a los valores de ajuste MNSQ de los procesos cognitivos al modelo propuesto. Igualmente se analizaron los valores de posición y el índice de separación entre los procesos cognitivos.

Finalmente, para cumplir con el cuarto objetivo del estudio, se compararon las puntuaciones clásicas de los tribunales (que solo consideran las calificaciones de los correctores) con las estimaciones de competencias ofrecidas por los modelos MFRM, donde la calificación de los estudiantes está corregida por el efecto de severidad de los correctores y de la dificultad de los procesos cognitivos evaluados. Todos los análisis se realizaron con el programa ConQuest 2.0 (Wu, Adams, Wilson y Haldane, 2007).

Resultados

Ajuste del modelo

La tabla III muestra los valores de la *deviance* y el número de parámetros de cada modelo y permite concluir que el modelo 4 es el que mejor se ajusta a los datos, al modelizar las diferencias entre los correctores en la interpretación de las rúbricas de los procesos cognitivos que, como se verá más adelante, es uno de los principales problemas de la rúbrica empleada.

TABLA III. Comparación de modelos: valores *deviance* y número de parámetros

	Deviance	N parámetros
Modelo 1: efectos principales	7376,3	17
Modelo 2: Interaction *criteria* x *step*	7339,1	20
Modelo 3: Interaction *rater* x *step*	7268,0	30
Modelo 4: Interaction *rater* x *criteria*	6800,6	41
Modelo 5: Interaction *rater* x *criteria* x *step*	7045,9	56

Nota: Elaboración propia

Efectos de los correctores

La tabla IV muestra el nivel de severidad de cada corrector y el error típico de dicha estimación; los valores de ajuste (MNSQ Uw-fit y MNSQ W-fit) con sus intervalos de confianza; y el valor del T-test.

TABLA IV. Efectos de los correctores

Id. Revisor	Severidad	ET de la Severidad	Ajuste sin ponderar (Uw-fit)		Ajuste ponderado (W-fit)	
			MNSQ IC	T	MNSQ IC	T
C01	-0,385	0,120	1,43 (0,73 - 1,27)	2,8	1,52 (0,71 - 1,29)	3,1
C02	0,758	0,100	1,19 (0,75 - 1,25)	1,5	1,21 (0,76 - 1,24)	1,6
C03	0,114	0,134	0,92 (0,73 - 1,27)	-0,5	0,93 (0,74 - 1,26)	-0,5
C04	0,634	0,107	0,92 (0,75 - 1,25)	-0,6	0,90 (0,75 - 1,25)	-0,8
C05	-1,050	0,132	0,80 (0,72 - 1,28)	-1,4	0,77 (0,68 - 1,32)	-1,6
C06	0,270	0,108	1,18 (0,75 - 1,25)	1,3	1,13 (0,75 - 1,25)	1,0
C07	2,398	0,120	1,32 (0,76 - 1,24)	2,4	1,36 (0,75 - 1,25)	2,6
C08	-0,961	0,124	1,20 (0,73 - 1,27)	1,4	1,14 (0,72 - 1,28)	1,0
C09	0,149	0,110	1,07 (0,75 - 1,25)	0,6	1,06 (0,75 - 1,25)	0,5
C10	-1,105	0,121	1,30 (0,75 - 1,25)	2,2	1,27 (0,73 - 1,27)	1,9
C11	-0,279	0,103	1,54 (0,75 - 1,25)	3,7	1,67 (0,74 - 1,26)	4,4
C12	-0,042	0,131	1,12 (0,74 - 1,26)	0,9	1,12 (0,74 - 1,26)	0,9
C13	-0,501*		1,66 (0,74 - 1,26)	4,3	1,65 (0,74 - 1,26)	4,2
*Parámetro fijo; ET: error típico; IC: intervalo de confianza						

Nota: Elaboración propia

Recibido: 16-11-2018 Aceptado: 21-06-2019

Los jueces presentan diferencias significativas en su grado de severidad, $\chi^2(12) = 723{,}1$; $p < .001$, siendo la fiabilidad del índice de separación entre correctores muy alta (R = 0,985). El rango de severidad es de 3,5 *logits* ($\lambda_{C07} = 2{,}4$ vs. $\lambda_{C10} = -1{,}1$), mientras que la desviación típica de la distribución latente, expresada como la raíz cuadrada de la varianza muestral, es de 1,6 *logits*. Incluso eliminado al corrector C07 las diferencias siguen siendo grandes, ya que la distancia entre el resto de correctores más severos (C02 y C04) y los más permisivos (C10, C05 y C08) está en torno a 1,9 *logits*. Por tanto, dependiendo del rigor del juez la calificación de un mismo estudiante podría variar más de una desviación típica de la distribución latente.

Los datos no permiten concluir la existencia de los efectos halo y centralidad en los correctores. El valor de ajuste más bajo corresponde al C05 (MNSQ-Uw = 0,80; MNSQ-W = 0,77), si bien está claramente por encima del valor crítico MNSQ = 0,5. Por tanto, parece descartarse el efecto de sobre-consistencia dentro de los jueces como consecuencia de un manejo inadecuado de la rúbrica analítica. Por el contrario, los índices de ajuste señalan que tres correctores (C01, C11 y C13) han realizado un manejo inconsistente de las rubricas. En los tres casos los valores MNSQ están por encima de 1,5 superando el límite del criterio más permisivo de los señalados por Eckes (2005). En todo caso, el desajuste es pequeño hasta el punto que aplicando el criterio de Wolfe y McVay (2012), sólo los valores del C13 aparecen desajustados, al mostrar sus puntuaciones una mayor variación de lo esperable por el modelo.

Efectos de las rúbricas

La tabla V muestra los parámetros de los procesos cognitivos (o criterios de corrección) y señala que los correctores han sabido discriminar entre ellos, siendo las diferencias globales entre procesos significativas, $\chi^2(2) = 532{,}6$; $p < .001$, y la fiabilidad del índice de separación entre criterios muy alta (R = 986).

TABLA V. Efectos de las rúbricas

Criterios	Dificultad	ET de la dificultad	Ajuste sin ponderar (Uw-fit)		Ajuste ponderado (W-fit)	
			MNSQ IC	T	MNSQ IC	T
Coherencia	0,415	0,042	1,11 (0,86 - 1,14)	1,5	1,10 (0,86 - 1,14)	1,3
Cohesión	0,931	0,045	1,03 (0,86 - 1,14)	0,4	1,05 (0,86 - 1,14)	0,6
Presentación	-1,345*		1,42 (0,86 - 1,14)	5,1	1,51 (0,84 - 1,16)	5,4
*Parámetro fijo; ET: error típico; IC: intervalo de confianza						

Nota: Elaboración propia

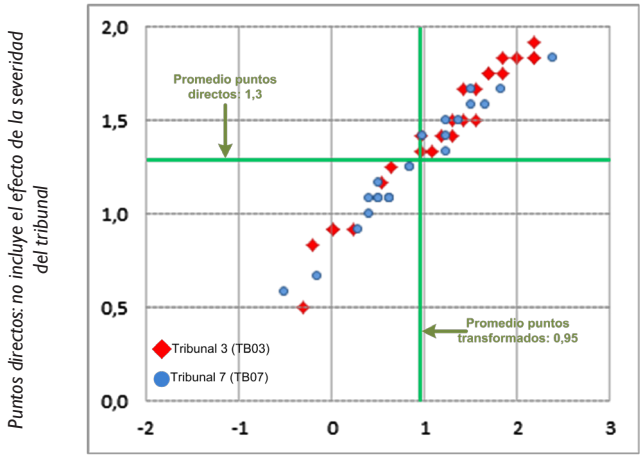
Los valores de posición indican que el proceso cognitivo más difícil fue Cohesión ($\lambda = 0,9$) y el más fácil Adecuación y Presentación ($\lambda = -1,3$). Este rango es razonable ya que los estándares de evaluación que subyacen a este último proceso (texto persuasivo-argumentativo, limpieza, respecto márgenes, etc.) son cuestiones consolidadas para la mayoría de los estudiantes de 12 años. No obstante, el proceso Adecuación y Presentación aparece desajustado (MNSQ > 1,3), indicando que los correctores no han hecho un uso coherente de las rúbricas de este proceso. Los altos valores MNSQ señalan que este proceso es menos discriminante que los otros dos, y que los correctores tienden a otorgar calificaciones polarizadas hacia los extremos (0 puntos o 2 puntos), advirtiéndose especialmente que las calificaciones están inesperadamente sesgadas hacia el nivel alto (2 puntos) que es la causa de que el valor MNSQ esté por encima de 1,3 puntos. El análisis de componentes principales inicial señaló que el peso factorial de los procesos Organización y Cohesión duplicaba el peso del proceso Adecuación y Presentación, lo que parece indicar que probablemente este último proceso, desde el punto de vista de la dimensionalidad de la escala, esté evaluando cuestiones diferentes a los otros dos. De hecho, al eliminar el proceso Adecuación y Presentación el ajuste del modelo mejora considerablemente, desapareciendo incluso los valores desajustados de los correctores. Estas evidencias parecen señalar que los desajustes encontrados en los correctores pueden deberse a una falta de especificación correcta de las rúbricas de corrección, antes que a un uso inconsistente de las rúbricas.

Recibido: 16-11-2018 Aceptado: 21-06-2019

Sesgos introducidos por los correctores

Como ya se advirtió el efecto del corrector más claro son las discrepancias en el grado de severidad de los jueces. Dado que para asignar correctores a tribunales se empleó un diseño matricial es fácil predecir que algunos tribunales habrían sido especialmente severos y otros más benévolos. Los gráficos I y II ilustran las consecuencias de no considerar el efecto de la severidad de los tribunales en la estimación de los resultados del alumnado. En los gráficos cada punto de la nube señala la posición de un estudiante comparando su promedio expresado en puntos directos sin corregir por el efecto del tribunal (eje vertical), con la puntuación *logit* estimada para el mismo estudiante, pero teniendo en cuenta la severidad del tribunal al que fue asignado (eje horizontal). Para facilitar la interpretación de los resultados en ambos gráficos se han trazado las medias de la muestra en términos clásicos (1,3 puntos) y en la puntuación transformada y corregida por la severidad del tribunal (0,95 *logits*), quedando el espacio cartesiano dividido en cuatro cuadrantes.

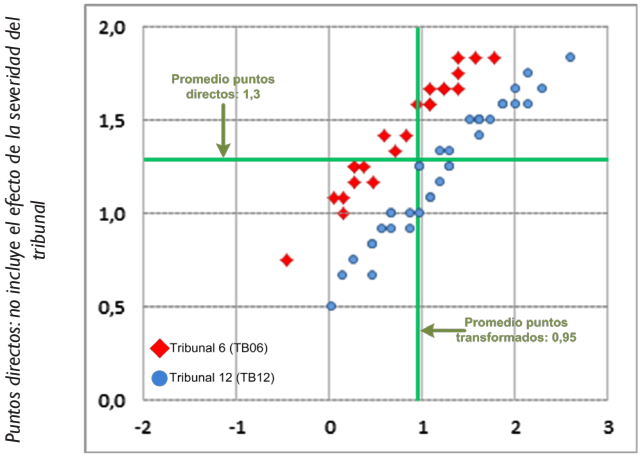
GRÁFICO I. Comparación de puntos directos y transformados del alumnado evaluado por TB03 y TB07



Nota: Elaboración propia

Recibido: 16-11-2018 Aceptado: 21-06-2019

GRÁFICO II. Comparación de puntos directos y transformados del alumnado evaluado por TB06 y TB12



Nota: Elaboración propia

El gráfico I compara las puntuaciones de dos tribunales 3 y 7, donde el promedio del nivel de severidad de sus cuatro correctores fue muy similar (λ_{TB03} = 0,013 y λ_{TB07} = 0,006). Las dos nubes de puntos aparecen superpuestas, apreciándose que los estudiantes del cuadrante inferior izquierdo, es decir, aquellos con una puntuación clásica por debajo de la media también se encuentran por debajo del promedio en el modelo que tiene en cuenta la severidad del tribunal. De igual modo los puntos situados en el cuadrante superior derecho corresponden con estudiantes por encima de la media en puntos directos que a su vez también presentan resultados por encima de la media cuando la estimación tiene en cuenta la severidad del tribunal. En el gráfico II se comparan los resultados de los dos tribunales más dispares: TB06 y TB12, que fueron respectivamente los tribunales más permisivo y más severo. La mayoría del alumnado del TB06 está sobre la media o por encima de la misma cuando se atiende a las puntuaciones clásicas, mientras que en el caso del TB12 la situación es la inversa: la mayoría de los estudiantes salen

Recibido: 16-11-2018 Aceptado: 21-06-2019

claramente perjudicados debido a que en el sorteo fueron asignados a un tribunal donde los cuatro correctores en su conjunto fueron los más severos. Sin embargo, cuando se tiene en cuenta el efecto de la severidad del corrector la situación cambia significativamente. Por ejemplo, el grupo de puntos rojos ubicados en el cuadrante superior izquierdo del gráfico II representa los casos que serían claramente favorecidos por la permisividad del TB06, ya que en términos clásicos superan el promedio de la prueba. No obstante, obtener 1,3 puntos en el TB06 no es lo mismo que obtenerlos en el TB12, por ello cuando el modelo corrige por la dificultad del tribunal se observa que los estudiantes que lograron 1,3 puntos en el tribunal permisivo se encuentran por debajo de la media en el modelo transformado (aproximadamente 0,3 *logits*), mientras que los estudiantes que obtuvieron el mismo promedio en el tribunal más severo lograrían entre 1 y 1,3 *logits*, es decir, estarían sobre la media o por encima de la media de la puntuación transformada.

Al tomar como punto de corte las medias clásica y corregida (1,3 puntos y 0,95 *logits* respectivamente) para decidir la competencia o no del alumnado, se encontró que 41 de los 375 casos valorados, es decir, el 15,4% serían susceptibles de modificar el diagnóstico. En 29 casos los estudiantes presentan una puntuación clásica por encima de la media, pero que cae por debajo del promedio de puntos transformados. Serían aquellos que se han beneficiado por la benevolencia del tribunal asignado. En los 12 casos restantes (8 de los cuales se concentran en el TB12) la situación es la contraria: la severidad del tribunal los coloca por debajo de la media en términos clásicos, mientras que su puntuación corregida por la severidad del tribunal está por encima de la media.

Discusión y Conclusiones

Uno de los retos de las evaluaciones educativas de alto impacto es asegurar la objetividad y equidad en las calificaciones. La corrección objetiva y rigurosa de los ensayos y de los ítems de respuesta construida constituye un verdadero desafío. Para lograr este objetivo es necesario implementar soluciones metodológicas diversas. En primer lugar hay que dar un entrenamiento sistemático a los correctores, y en segundo lugar utilizar diseños de anclaje adecuados para asignar correctores a tribunales (Frey, et al., 2009; OECD, 2014). Este trabajo ha mostrado

el potencial de los diseños incompletos y balanceados para controlar el efecto de asignación de los correctores a los tribunales. Fernández-Alonso y Muñiz (2011) señalan las ventajas en inconvenientes del uso de diseños completos e incompletos. No obstante, el arreglo Youden de 13 bloques empleado en este estudio es una solución que aúna la robustez y consistencia propia del cuadrado latino con el equilibrio y eficiencia de los diseños incompletos balanceados. Desde un punto de vista psicométrico, los modelos MFRM utilizados muestran grandes posibilidades para realizar un diagnóstico integral de las rúbricas de corrección, encontrando resultados significativos relacionados con los cuatro objetivos planteados en el estudio. Con respecto al primer objetivo el modelo de interacción entre correctores y procesos cognitivos (criterios) fue el que presentó mejor ajuste, señalando diferencias entre los correctores a la hora de interpretar las rúbricas de los procesos cognitivos a evaluar; este resultado es convergente con los presentados por Adams y Wu (2010) y Eckes (2005), en los que los modelos más complejos ajustan mejor que el modelo más parsimonioso de efectos principales.

En relación al segundo objetivo, el análisis con MFRM ha demostrado su capacidad para detectar sesgos de corrección, encontrándose una fuerte variación en los niveles de severidad de los correctores, que se estima en más de una desviación típica en la escala de puntuaciones, aún después de descontar el efecto de los correctores más extremos. Estos datos se alinean con trabajos precedentes que indican que las diferencias entre-jueces son la norma y no la excepción (Congdon y McQueen, 2000; Eckes, 2005; Lopes et al., 2015; Lunz et al., 1990; Prieto-Adánez, 2011; Prieto-Adánez y Nieto, 2014; Wolfe, 2004; Wolfe y McVay, 2012). Se identificaron desajustes y puntuaciones inconsistentes en tres de los trece correctores, si bien bajo algunos de los criterios establecidos por la literatura (Wolfe y McVay, 2012) los valores estarían dentro de los límites del ajuste esperados, salvo en un caso. No obstante, se han detectado indicios de que esta anomalía podría estar motivada, no tanto por una baja fiabilidad intra-juez, como por un problema estructural de la propia rúbrica.

El análisis de las rúbricas de los criterios de evaluación (tercer objetivo) señala un desajuste en proceso Adecuación y Presentación, apareciendo como el proceso menos discriminante y con puntuaciones más polarizadas e inesperadamente sesgadas hacia el extremo superior de la escala. Los

Recibido: 16-11-2018 Aceptado: 21-06-2019

datos del análisis exploratorio de componentes principales ya apuntaban a que la inclusión de este proceso podría ser problemática y sus rúbricas deberían ser revisadas, ya que cuando se elimina del análisis todos los parámetros del modelo (incluido el ajuste de los correctores) muestran valores adecuados. En la literatura están documentados este tipo de problemas, ya sea porque los criterios sean reiterativos (Adams y Wu, 2010), o bien por la falta de consistencia entre los criterios (Basturk, 2008; Lunz et al., 1990), lo que demuestra las posibilidades del análisis MFRM para detectar errores de diseño en las rúbricas de corrección.

El último objetivo del trabajo pretendía mostrar la ventaja práctica de los MFRM, en especial la posibilidad de ofrecer puntuaciones del alumnado corregidas por la severidad de los jueces o tribunales. Estudios previos muestran que al incluir el efecto del corrector en las puntuaciones los resultados de algunos estudiantes pueden variar significativamente (Lunz et al., 1990; Wang y Yao, 2013). En nuestro caso, se han presentado datos que confirman que estudiantes con idéntica puntuación podrían estar separados por aproximadamente un *logit* cuando el modelo considera la severidad del tribunal. También se encuentra que el diagnóstico en un 15,4% de los casos podría variar según se tenga en cuenta o no el efecto de la severidad del tribunal. Estos resultados son similares a los presentados por Eckes (2005), el cual reanaliza los resultados del TestDAF (una prueba de certificación de nivel de idioma en Alemania) y estima que el 13,5% de los casos en expresión escrita y el 17,1% en expresión oral deberían cambiar de resultado en la prueba (certificar o no certificar) si se hubiese considerado el efecto del corrector.

Los datos presentados en este trabajo tienen importantes implicaciones educativas. Por un lado, muestran las posibilidades del análisis de los modelos MFRM para diagnosticar los efectos del corrector y posibles defectos en la definición de la variable a estudiar o en las rúbricas de corrección empleadas. Igualmente se ha mostrado que cuando se tienen en cuenta las diferencias de los jueces en grado de severidad del tribunal los resultados pueden variar, en algunos casos muy significativamente. En España existen exámenes con alto impacto como serían las pruebas de acceso a la universidad o las certificaciones de idiomas donde es probable que los efectos de la severidad del juez estén incidiendo en los resultados del alumnado. En ambos casos existen estudios centrados en analizar la consistencia de los correctores (Amengual y Méndez García, 2012; Cuxart- Jardí, 2000; Cuxart-Jardí et al., 1997; Grau, Cuxart y Martí-

Recober, 2002; Prieto-Adánez, 2011; Prieto-Adánez y Nieto, 2014; Suárez-Álvarez, González-Prieto, Fernández-Alonso, Gil y Muñiz, 2014), sin embargo, y hasta donde alcanza nuestra información, no existe ningún estudio centrado en analizar las consecuencias prácticas que tiene para el alumnado el hecho de que sus calificaciones sean asignadas por correctores que probablemente tengan niveles de severidad variables.

A la vista de los datos aportados en este estudio y teniendo en cuenta el alto impacto de este tipo de pruebas en el futuro académico y personal del alumnado parece necesario explorar con detalle sus procesos de corrección y calificación. También sería importante analizar el posible funcionamiento diferencial de los ítems inducido por correctores y rúbricas (Gómez-Benito, Sireci, Padilla, Hidalgo y Benítez, 2018; Woitschach, Zumbo y Fernández-Alonso, 2019), así como comprobar la generalización de los resultados a evaluaciones de alumnado con distintas características personales (Amor, Verdugo, Calvo, Navas y Aguayo, 2018). Al lado de los citados análisis, un aspecto fundamental es el desarrollo de programas sistemáticos de formación y entrenamiento de los evaluadores, para así minimizar las diferencias individuales relativas a las estrategias de corrección y calificación seguidas. Se trata, en suma, de complementar las medidas de entrenamiento y formación de los evaluados con los procedimientos de análisis de los resultados, utilizando los modelos psicométricos y estadísticos más actuales y sofisticados. Cabe esperar que esa doble combinación pre-post reduzca a un mínimo los errores cometidos en la evaluación rigurosa y objetiva del alumnado, una exigencia ética y deontológica ineludible.

Como no puede ser de otro modo, el trabajo no está exento de limitaciones, por lo que los resultados hay que tomarlos con prudencia, y no se puede realizar su generalización universal sin más. Hay tres fuentes claras de error que pueden afectar los resultados de este tipo de trabajos, en primer lugar, el tipo de evaluador, cuyas características personales, de formación, entrenamiento, e implicación pueden variar de unos casos a otros. La materia objeto de la evaluación impone también algunas restricciones que pueden modular los resultados. Finalmente, el diseño de las propias rúbricas puede influir en los resultados de la evaluación, por ejemplo, minimizando o maximizando la varianza de las puntuaciones obtenidas.

Recibido: 16-11-2018 Aceptado: 21-06-2019

Referencias bibliográficas

Adams, R. y Wu, M. (2010). *The analysis of rater effects*. Recuperado de: https://www.acer.org/files/Conquest-Tutorial-3-RaterEffects.pdf

Amengual, M. (2004). *Análisis de la fiabilidad en las puntuaciones holísticas en ítems abiertos* [Reliability analysis in holistic scores on constructed response ítems]. Universidad Complutense de Madrid: Tesis doctoral no publicada. Recuperado de: http://biblioteca.ucm.es/tesis/fll/ucm-t26663.pdf

Amengual, M. y Méndez García, M. C. (2012). Implementing the oral English task in the spanish university admission examination: An international perspective of the language. *Revista de Educación*, 357, 105-127.

Amor, A. M., Verdugo, M. A., Calvo, M. I., Navas, P. y Aguayo, V. (2018). Psychoeducational assessment of students with intellectual disability: professional-action framework analysis. *Psicothema, 30*, 39-45. doi: 10.7334/psicothema2017.175

Ato, M. y Vallejo, G. (2007). *Diseños experimentales en psicología*. Madrid: Pirámide.

Baird, J. A., Hayes, M., Johnson, R., Johnson, S. y Lamprianou, I. (2013). *Marker effects and examination reliability. A comparative exploration from the perspectives of generalizability theory, Rasch modelling and multilevel modelling (Ofqual/13/5261)*. Coventry: Office of Qualifications and Examinations Regulation. Recuperado de: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/378059/2013-01-21-marker-effects-and-examination-reliability.pdf

Baird, J. A., Meadows, M., Leckie, G. y Caro, D. (2017). Rater accuracy and training group effects in Expert- and Supervisor-based monitoring systems. *Assessment in Education: Principles, Policy and Practice*, *24*(1), 44-59. doi: 10.1080/0969594X.2015.1108283

Basturk, R. (2008). Applying the many-facet Rasch model to evaluate PowerPoint presentation performance in higher education. *Assessment y Evaluation in Higher Education*, *33*(4), 431- 444. doi: 10.1080/02602930701562775

Bejar, I. I. (2012): Rater cognition: Implications for validity. *Educational Measurement*, *31*(3), 2-9.

Crisp, V. (2012). An investigation of rater cognition in the assessment of projects. *Educational Measurement*, *31*(3), 10-20.

Cochran, W.G. y Cox, G.M. (1974). *Diseños experimentales*. México: Trillas. (orig. 1957).

Congdon, P. J. y McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, *37*(2), 163-178. doi: 10.1111/j.1745-3984.2000.tb01081.x

Cuxart-Jardí, A. (2000). Modelos estadísticos y evaluación: tres estudios en educación. [Statistical models and assessment: three studies in education]. *Revista de Educación*, 323, pp. 369-394

Cuxart-Jardí, A., Martí-Recober, M. y Ferrer-Juliá, F. (1997). Algunos factores que inciden en el rendimiento y la evaluación en los alumnos de las pruebas de aptitud de acceso a la universidad [Some factors that affect the students' performance in the university entrance exams]. *Revista de Educación*, 314, 63-88.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A multi-faceted Rasch analysis. *Language Assessment Quarterly*, *2*(3), 197-221. doi: 10.1207/s15434311laq0203_2

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Section H). Strasbourg, France: Council of Europe/Language Policy. Recuperado de: https://rm.coe.int/1680667a23#search=eckes

European Commission/EACEA/Eurydice (2009*). National testing of pupils in Europe: Objectives, organization and use of results*, Luxembourg: Publications Office of the European Union. doi: 10.2797/18294

Engelhard, G. (1992). The measurement of writing ability with a multi-faceted Rasch model. *Applied Measurement in Education*, *5*(3), 171-191. doi: 10.1207/s15324818ame0503_1

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a multi-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93-112. doi: 10.1111/j.1745-3984.1994.tb00436.x

Fernández-Alonso, R. y Muñiz, J. (2011). Diseño de cuadernillos para la evaluación de las competencias básicas [Booklet designs for the evaluation of basic skills]. *Aula Abierta*, *39*(2), 3-34.

Frey, A., Hartig, J. y Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28*(3), 39-53.

Recibido: 16-11-2018 Aceptado: 21-06-2019

Gómez-Benito, J., Sireci, S., Padilla, J. L., Hidalgo, M. D. y Benítez, I. (2018). Differential item functioning: Beyond validity evidence based on internal structure. *Psicothema, 30*, 104-109. doi: 10.7334/psicothema2017.183

Grau, R. M., Cuxart, A. y Martí-Recober, M. (2002). La calidad en el proceso de corrección de las pruebas de acceso a la universidad: variabilidad y factores [Quality in the scoring process of university entrance exams: variability and factors]. *Revista de Investigación Educativa*, *20*(1), 209-224.

Gyagenda, I. y Engelhard, G. (2009). Using classical and modern measurement theories to explore rater, domain, and gender influences on student writing ability. *Journal of Applied Measurement*, *10*(3), 225-246.

Jonsson, A. y Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130-144.

Kuo, S. A. (2007): Which rubric is more suitable for NSS liberal studies? Analytic or holistic? *Educational Research Journal*, *22*(2), 179-199.

Lallmamode, S. P., Mat Daud, N. y Abu Kassim, N. L. (2016). Writing Development and initial argument-based validation of a scoring rubric used in the assessment of L2 writing electronic portfolios. *Assessing Writing, 30*, 44-62. http://dx.doi.org/10.1016/j.asw.2016.06.001

Leckie, G. y Baird, J. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, *48*(4), 399-418. doi: 10.1111/j.1745-3984.2011.00152.x

Ley Orgánica 15/1999 de Protección de Datos de Carácter Personal. *Boletín Oficial del Estado de 14 de diciembre de 1999.* Recuperado de: https://www.boe.es/buscar/doc.php?id=BOE-A-1999-23750

Ley Orgánica 8/2013 para la Mejora de Calidad Educativa. *Boletín Oficial del Estado de 10 de diciembre de 2013*. Recuperado de: https://www.boe.es/buscar/pdf/2013/BOE-A-2013-12886-consolidado.pdf

Linacre, J. M., Engelhard, G., Tatum, D. S. y Myford, C. M. (1994) Measurement with judges: Many-faceted conjoint measurement. *International Journal of Educational Research*, *21*(6), 569-577. doi: 10.1016/0883-0355(94)90011-6

Lopes Toffoli, S. F., de Andrade, D. F. y Bornia, A. C., (2015): Evaluation of open items using the many-facet Rasch model, *Journal of Applied Statistics*, doi: 10.1080/02664763.2015.1049938

Lunz, M. E. y Stahl, J. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions*, *13*(4), 425-444. doi: 10.1177/016327879001300405

Lunz, M. E., Wright, B. D. y Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, *3*(4), 331-345. doi: 10.1207/s15324818ame0304_3

Mahmood, O., Dagnæs, J., Bube, S., Rohrsted, M. y Konge, L. (2017). Nonspecialist raters can provide reliable assessments of procedural skills. *Journal of Surgical*. Recuperado de: http://dx.doi.org/10.1016/j.jsurg.2017.07.003

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman

Ministerio de Educación, Cultura y Deporte (2016): *Pruebas de la evaluación final de Educación Primaria. Curso 2015-2016*. Madrid: Instituto de Evaluación

Myford, C. M. y Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*, 386-422

Myford, C. M. y Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*, 189-227

Organisation for Economic Co-operation and Development [OECD]. (2014). *PISA 2012 Technical Report*. Paris: OECD Publishing. Recuperado de: https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

Park, T. (2010). *An investigation of an ESL placement test of writing using multi-faceted Rasch measurement*. Teachers College, Columbia University Working Papers in TESOL and Applied Linguistics, 4(1), 1-19

Prieto-Adánez, G. (2011). Evaluación de la ejecución mediante el modelo Many Facet Rasch Measurement [Performance assessment using the Many-Facet Rasch Measurement]. *Psicothema*, *23*, 233-238

Prieto-Adánez, G. (2015). Análisis de un test de desempeño en expresión escrita mediante el modelo de MFRM [Analysis of a writing test with the MFRM model]. *Actualidades en Psicología, 29*(119), 1-17. http://dx.doi.org/10.15517/ap.v29i119.19822

Prieto-Adánez, G. y Nieto, E. (2014). Analysis of rater severity on written expression exam using Many Faceted Rasch Measurement. *Psicológica*, 35, 385-397

Recibido: 16-11-2018 Aceptado: 21-06-2019

Saal, F. E., Downey, R. G. y Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*(2), 413-428

Shackleton, C. (2018). Linking the University of Granada CertAcles listening test to the CEFR. *Revista de Educación, 381*, 35-60. doi: 10.4438/1988-592X-RE-2017-381-380

Suárez-Álvarez, J., González-Prieto, C., Fernández-Alonso, R., Gil, G. y Muñiz, J. (2014). Psychometric assessment of oral expression in English language in the University Entrance Examination. *Revista de Educación*, 364, 93-118. doi: 10.4438/1988-592X-RE-2014-364-256

Sudweeks, R. R., Reeve, S. y Bradshaw, W. S. (2005). A comparison of generalizability theory and many facet measurement in an analysis of college sophomore writing. *Assessing Writing, 9*, 239-261.

Suto, I. (2012). A critical review of some qualitative research methods used to explore rater cognition. *Educational Measurement, 31* (3), 21-30.

Wang, Z. y Yao, L. (2013). The effects of rater severity and rater distribution on examinees' ability estimation for constructed-response items. *ETS Research Report Series*, i-22. doi:10.1002/j.2333-8504.2013.tb02330.x

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, *46*(1), 35-51.

Wolfe, E. W. y McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement*, *31* (3), 31-37.

Woitschach, P., Díaz-Pérez, C., Fernández-Argüelles, D., Fernández-Castañón, J., Fernández-Castillo, A., Fernández-Rodríguez, L., ... y Fernández-Alonso, R. (2018). Efectos del corrector en las evaluaciones educativas de alto impacto. [Rater effects in high-impact educational assessments]. *REMA Revista Electrónica de Metodología Aplicada, 23*(1), 12-27.

Woitschach, P., Zumbo, B. D. y Fernández-Alonso, R. (2019). An ecological view of measurement: Focus on multilevel model explanation of differential item functioning. *Psicothema, 31*(2), 194-203. doi:10.7334/psicothema2018.303

Wu, M. L., Adams, R. J., Wilson, M. R. y Haldane, S. A. (2007). *ACER ConQuest 2.0: generalised item response modelling software*. Camberwell, Victoria: Australian Council for Educational Research.

Recibido: 16-11-2018 Aceptado: 21-06-2019

Dirección de contacto: Rubén Fernández-Alonso. Consejería de Educación y Cultura del Gobierno del Principado de Asturias y Universidad de Oviedo, Facultad de Formación del Profesorado y Educación, Departamento de Ciencias de la Educación. Calle Aniceto Sela, 1, 33005 Oviedo, Asturias. E-mail: fernandezaruben@uniovi.es

Recibido: 16-11-2018 Aceptado: 21-06-2019

Rubrics do not neutralize rater effects: A many-faceted Rasch model estimation[1]

Las Rúbricas No neutralizan el Efecto de los correctores: Una estimación con el modelo de facetas múltiples de Rasch

DOI: 10.4438/1988-592X-RE-2019-386-428

Rubén Fernández-Alonso
Pamela Woitschach
José Muñiz
Universidad de Oviedo

Abstract

Constructed response items are widely used in all types of educational evaluations. Despite the use of very specific rubrics for scoring items, the influence of raters is well documented in the literature, affecting the results of evaluations. The main goal of the present study is to estimate the effects of raters and rubrics in the assessment of written expression. We used 13 raters who reviewed 375 pieces of writing by sixth grade students. The raters were assigned to the written essays following a 13-block Youden square, a balanced incomplete block design. In the data analysis, the many-faceted Rasch model was used with three facets: corrector, rubrics and difficulty of passing. Different models were compared, and the effects of the raters and the characteristics of the rubrics were analyzed. The results show the differences between raters in terms of severity and accuracy of evaluation. We conclude that not considering the effect of the rater in the estimation of student outcomes can introduce a clear component of inequity in the evaluations.

Key words: Educational assessment, rater, scoring rubrics, MRFM models

(1) This research was funded by the Ministry of Economy and Competitiviness of the Spanish Government. References: PSI2017-85724-P, BES2012-053488.

Received: 16-11-2018 Accepted: 21-06-2019

Resumen

Los ítems de respuesta construida son ampliamente utilizados en todo tipo de evaluaciones educativas. A pesar de la utilización de rúbricas muy específicas para su corrección, la influencia de los correctores está bien documentada en la literatura, afectando a los resultados de la evaluación. El objetivo central del presente trabajo es la estimación de los efectos de los correctores y de las rúbricas en una tarea de expresión escrita. Se utilizaron 13 correctores que revisaron 375 producciones escritas de estudiantes de sexto curso. Los correctores fueron asignados a los ensayos escritos siguiendo un cuadrado Youden de 13 bloques, un diseño de bloques incompletos balanceado. En el análisis de datos se empleó el modelo de Rasch de facetas múltiples con tres facetas: corrector, rúbricas y dificultad de paso de los mismos. Se compararon diferentes modelos y se analizaron los efectos del corrector y las características de las rúbricas. Los resultados ponen de manifiesto las diferencias entre los correctores en cuanto la severidad y la exactitud de sus juicios. Se concluye que no incluir el efecto del corrector en la estimación de los resultados del alumnado puede introducir un componente claro de inequidad en las evaluaciones.

Palabras clave: Evaluación educativa, correctores, rúbricas, Modelos MFRM

Introduction

European education systems include high-stakes exams, such as exams for qualifications, certifications, access to further study and school evaluations. They include items where rater effects can influence a student's score (European Commission/EACEA/Eurydice, 2009*)*. Given how important these evaluations are to the student, society, and the various stakeholders involved, their objectivity has been widely researched (Congdon & McQueen, 2000; Engelhard, 1992; Gyagenda & Engelhard, 2009; Leckie & Baird, 2011; Lunz & Stahl, 1990; Lunz, Wright & Linacre, 1990; Park, 2010; Wolfe, 2004). According to Eckes (2009) there are two broad sources of error that affect the objectivity of evaluations: distal and proximal.

Distal factors refer to aspects such as the characteristics of the evaluator, the student, and the context of the evaluation. Evaluator variation has been found associated with gender, age, professional history, the field the rater works in, as well as in raters' attitudes towards the gender,

Received: 16-11-2018 Accepted: 21-06-2019

nationality or ethnicity of the person being evaluated (Amengual, 2004; Congdon & McQueen, 2000; Linacre, Engelhard, Tatum & Myford, 1994; Lopes Toffoli, de Andrade & Bornia, 2015; Mahmood, Dagnæs, Bube, Rohrsted & Konge, 2017). Differences linked to the grading process have also been found, such as previous experience, the grading order, entertainment, familiarity with rubrics, and rater cognitive strategies (Bejar, 2012; Congdon & McQueen, 2000; Crisp, 2012; Jonsson & Svingby, 2007; Leckie & Baird, 2011; Linacre et al., 1994; McNamara, 1996; Suto, 2012).

Proximal factors refer to the construct being measured, the difficulty of the task and the grading criteria, the structure of the scoring banks and rater effects (Eckes, 2009). Cuxart-Jardí (2000), Cuxart-Jardí, Martí-Recober and Ferrer-Juliá (1997) found better agreement in scientific and mathematics subjects than in socio-linguistic subjects. Jonsson and Svingby (2007) reported differences in reliability according to the number of categories used and the nature of the task, concluding that indexes of agreement increase as the number of categories decreases, and that physical tests and case studies exhibit higher reliability indexes than oral or written production. Finally, Baird, Meadows, Leckie and Caro (2017), Jonsson and Svingby (2007) and Kuo (2007) agree when they indicate that in high-stakes tests analytical rubrics are preferable to holistic rubrics, although the procedures, scoring and criteria for discerning between performance levels need to be clearly detailed with examples. The proximal factors that have received most attention include rater effects (strictness/leniency, halo, range restriction, consistency, etc.), which have been studied from various methodological perspectives such as the classical approach (OECD, 2014; Saal, Downey & Lahey, 1980; Woitschach et al., 2018), generalizability theory (Sudweeks, Reeve & Bradshaw, 2005), multilevel analysis (Congdon & McQueen, 2000; Leckie & Baird, 2011), and Item Response Theory models (TRI; Adams & Wu, 2010; Lunz et al., 1990; Myford & Wolfe, 2003, 2004). While each strategy has its strengths and weaknesses (Baird, Hayes, Johnson, Johnson & Lamprianou, 2013; Gyagenda&Engelhard, 2009; Sudweeks et al, 2005), TRI models offer integrated solutions, as they allow rubrics to be validated (Lallmamode, Mat Daud & Abu Kassim, 2016), they examine the effects of various sources of variation (Linacre et al., 1994), and they manage those effects as a whole in order to estimate student results (Eckes, 2009).

Received: 16-11-2018 Accepted: 21-06-2019

The Many-Facet Rasch Measurement (MRFR; Linacre et al., 1994) allows the parameterization in one scale of all of the facets of the model; subject competence, rater characteristics, difficulty, category parameters, and how many facets the analysis could include (Eckes, 2009; Myford & Wolfe, 2003, 2004; Prieto-Adánez, 2011). In addition, Linacre et al. (1994) and Lunz et al. (1990) demonstrated that MFRM could identify rater effects (severity/leniency, inconsistent scoring, over-consistency, etc.) and diagnose issues with grading rubrics and criteria (fit, scale dimensionality, category parameter, etc.). The main practical advantage of MFRM is that it has demonstrated the need to include model parameters (in particular variations in rater severity) in the estimation of student test results, as otherwise this could affect decisions about levels of student competence, and ultimately the validity of study conclusions (Eckes, 2005, 2009; Prieto-Adánez, 2015; Shackleton, 2018).

Within this context, the objective of this study is to examine rater effects in the evaluation of writing, and its interaction with the rubrics used. This general objective can be separated into four specific goals: (a) Choose the multi-faceted Rasch model that best fits the data; (b) Examine the fit of the raters to the model and analyze the effects introduced by the raters; (c) Examine the effects of the rubrics and their fit to the model; and (d) Estimate the extent of bias added by the raters to the students' grades.

These objectives are particularly interesting as there are very few studies that have looked at the consequences for students of their work being assigned to raters with differing levels of severity. In Spain, research up to now has focused on examining rater effects and the fit of models, without estimating wrongly graded cases as they have not included the effects of the rater in determining student competence. We believe that our research is breaking new ground in this sense, in addition to providing novel methodologies in the design and analysis of assigning raters.

Method

Sample

We used 13 raters who were assigned texts following a 13-block Youden square design. This is an incomplete balanced Latin square which uses the first four replicas of the 13x13 Latin square (Ato & Vallejo, 2007;

Received: 16-11-2018 Accepted: 21-06-2019

Cochran & Cox, 1974; Frey, Hartig & Rupp, 2009; Organization for Economic Co-operation and Development [OECD], 2014). To do this the writings from 18 classes were grouped in 13 grading groups (blocks in experimental design terms). Table I shows the 13 groups and the number of writings assigned to each. On average, each group had 28 writings, although groups TB09 to TB13 graded more as they had writings from two classes each.

TABLE I. Makeup of the 13 grading groups: classes assigned, number of writings to grade, and raters in each grading group

Grading groups (blocks of writings)	Class(es) assigned to the group	N° of writings to grade	Raters in each grading group			
TB01	Aula 01	28	C01	C09	C12	C10
TB02	Aula 02	26	C02	C01	C08	C11
TB03	Aula 03	25	C03	C02	C09	C05
TB04	Aula 04	25	C04	C03	C01	C06
TB05	Aula 05	24	C05	C11	C04	C12
TB06	Aula 06	23	C06	C05	C10	C08
TB07	Aula 07	23	C07	C13	C05	C01
TB08	Aula 08	22	C08	C12	C13	C03
TB09	Aula 09 & Aula 18	34	C09	C08	C07	C04
TB10	Aula 10 & Aula 17	36	C10	C04	C02	C13
TB11	Aula 11 & Aula 16	35	C11	C10	C03	C07
TB12	Aula 12 & Aula 15	36	C12	C07	C06	C02
TB13	Aula 13 & Aula 14	37	C13	C06	C11	C09

Note: Authors' data

The final four columns in the table identify the four raters assigned to each group and allow us to check the consistency, balance and efficiency of the Youden design. The consistency of the design means that each rater is assigned four groups and that each writing is graded by four raters. For example, Grader 1 (C1) is in groups TB01, TB02, TB04 and

Received: 16-11-2018 Accepted: 21-06-2019

TB07. Similarly, the writings assigned to TB01 will be graded by C1, C9, C12 and C10. The design is completely balanced as each rater will grade the same set of writings with every other rater on one occasion. For example, C1 will grade with C09, C12 and C10 in group TB01, with C02, C11 and C08 in group TB02, with C03, C04 and C06 in group TB04 and the remaining three raters in group TB07. The remaining groupings can be seen in table 1.

The efficiency of the design is that each rater will, on average, grade a little over 100 writings (for example rater C12 will grade the 110 writings in sets TB01, TB05, TB08 and TB12) rather than the 375 that a full Latin square would need. In total 1500 writings will be graded in this study rather than the almost 5000 that the full design would need, saving about 60% of possible gradings without the conclusions of the analysis losing validity. In short, this distribution of raters and writings ensures the dual control of the Latin square (Fernández-Alonso & Muñiz, 2011) such that the possible severity or leniency of the raters cannot be imputed to being assigned a skilled or less-skilled writing group.

The writings in this study come from the final evaluation of primary education (the *Evaluación Final de Educación Primaria*; required by organic law 8/2013 on Education Quality Improvement) carried out in a Spanish Autonomous community in academic year 2015/16, which collected 6653 writings from 403 class groups of sixth grade primary education. Using systematic, random sampling, 18 classes were selected with a probability proportional to their size (OECD, 2014). All of the students' writings from each chosen class were graded, 375 in total.

Instruments

The prompt used in the evaluation was a text containing tourist information for three cities. The students had to choose one city and write a text to convince their classmates that it would be the best destination for a class trip. The grading rubric evaluated three constructs or cognitive processes in the official Spanish curriculum. Table II shows how the standards and processes are organized. The scores in the standards are summarized in the three constructs which were coded in a three-level scale as follows; 0 points = not achieved; 1 = achieved; and 2 = Consolidated. The test, specifications and rubrics are available from the Ministry of Education,

Received: 16-11-2018 Accepted: 21-06-2019

Culture and Sport (2016, pp. 97-117) and may be found at: https://sede.educacion.gob.es/publiventa/descarga.action?f_codigo_agc=18314

TABLE II. Constructs or cognitive processes evaluated and standards in the grading rubric

Constructs or cognitive processes	Rubric standards
Coherence	Organize ideas clearly and with thematic progression; express opinions, thoughts and evaluations coherently.
Cohesion	Use connectors and punctuation to produce a cohesive text; use pronoun substitution and synonyms to avoid repetition.
Appropriateness and presentation	Use appropriate grammar and spelling; use an appropriate register for the reader and topic; present a tidy writing that is clear, precise and ordered.

Procedure

Permission was first obtained so that the relevant education department could partly release the files containing the tests from the selected classes, in accordance with Spanish data protection laws (Organic law 15/1999 on protection of personal data). Each piece of writing, originally in pencil on paper, was scanned, anonymized via an identification code, edited with the Adobe Acrobat Prof DC® program to remove original grading marks, and saved in a file labelled with the identification code.

In addition, 13 grading templates were prepared (one for each rater) containing the data about the writings assigned to each rater. An online platform was designed where all 375 writings were uploaded in which each rater only had access to their assigned files. The raters recorded the results of their grading on the online platform, and once the grades were uploaded, the codes were verified to avoid transcription errors.

The raters, postgraduate Education students, received six hours of training in grading procedures, two hours of training in basic concepts of grading via rubrics, and four hours of work with the test instruments, including training in grading similar writings to those used in the study to familiarize themselves with the material and the grading process.

Data analysis

We used the MFRM for polytomous data, which assumes that each student's score is a logistical function which includes three facets in addition to student competence: grader (*rater*), cognitive processes (*criteria*) and difficulty of those processes (*step*). This lets us define a model of principal effects and additional models of interaction between facets, and lets us perform a hierarchical comparison of models. Initially, the model of principal effects is created, then three models of interaction of two facets (models of second level interaction: *criteria x step; rater x step; and criteria x rater*), and finally a model which includes the interaction between the three facets together (*criteria x step x rater*). To see which model demonstrated the best fit, we used the χ^2 test, comparing the reduction in *deviance* in each new model with the increase in the number of parameters necessary to produce models of growing complexity (Adams & Wu, 2010).

Once the model was selected, and to meet the second study objective, the suitability of the scores was checked via the values of the raters' mean square residuals (MNSQ), following the criteria from Eckes (2005) and Wolfe and McVay (2012). Rater effects (difficulty, halo and central tendency) were examined via comparison of the estimators of rater severity, analysis of separation indices between raters and the MNSQ fit values for the raters according to the criteria from Eckes (2005, 2009), Lunz et al., (1990), Engelhard (1994) and Wolfe and McVay (2012).

The dimensionality of the scale, based on the supposed positive correlation between the scores of the three cognitive processes being evaluated, was verified by considering the MNSQ values of fit of the cognitive processes to the proposed model. Similarly, the values of position and indices of separation between the cognitive processes were analyzed.

Finally, to meet the fourth study objective, we compared the classical scores of the rater groups (which only consider the rater grades) with estimations of student competency offered by the MFRM models, in which student scores are corrected for the effect of rater severity and the difficulty of the cognitive processes being evaluated. All analysis was performed using ConQuest 2.0 software (Wu, Adams, Wilson & Haldane, 2007).

Received: 16-11-2018 Accepted: 21-06-2019

Results

Model fit

Table III gives the values of *deviance* and the number of parameters in each model. Model four exhibits the best fit to the data, modelling the differences between raters in the interpretation of the rubrics of the cognitive processes which, as will be seen later, is one of the main problems of the rubric used.

TABLE III. Comparison of models: *deviance* values and number of parameters

	Deviance	**N° of parameters**
Model 1: Principal effects	7376.3	17
Model 2: Interaction *criteria* x *step*	7339.1	20
Model 3: Interaction *rater* x *step*	7268.0	30
Model 4: Interaction *rater* x *criteria*	6800.6	41
Model 5: Interaction *rater* x *criteria* x *step*	7045.9	56

Note: Authors' data

Rater effects

Table IV shows the levels of each rater's severity and the standard error of this estimation, the values of fit (MNSQ Uw-fit and MNSQ W-fit) and their confidence intervals, and the result of the T-test.

Received: 16-11-2018 Accepted: 21-06-2019

TABLE IV. Rater effects

Rater ID	Severity	SE of severity	Unweighted fit (Uw-fit)		Weighted fit (W-fit)	
			MNSQ (CI)	T	MNSQ (CI)	T
C01	-0.385	0.120	1.43 (0.73 - 1.27)	2.8	1.52 (0.71 - 1.29)	3.1
C02	0.758	0.100	1.19 (0.75 - 1.25)	1.5	1.21 (0.76 - 1.24)	1.6
C03	0.114	0.134	0.92 (0.73 - 1.27)	-0.5	0.93 (0.74 - 1.26)	-0.5
C04	0.634	0.107	0.92 (0.75 - 1.25)	-0.6	0.90 (0.75 - 1.25)	-0.8
C05	-1.050	0.132	0.80 (0.72 - 1.28)	-1.4	0.77 (0.68 - 1.32)	-1.6
C06	0.270	0.108	1.18 (0.75 - 1.25)	1.3	1.13 (0.75 - 1.25)	1.0
C07	2.398	0.120	1.32 (0.76 - 1.24)	2.4	1.36 (0.75 - 1.25)	2.6
C08	-0.961	0.124	1.20 (0.73 - 1.27)	1.4	1.14 (0.72 - 1.28)	1.0
C09	0.149	0.110	1.07 (0.75 - 1.25)	0.6	1.06 (0.75 - 1.25)	0.5
C10	-1.105	0.121	1.30 (0.75 - 1.25)	2.2	1.27 (0.73 - 1.27)	1.9
C11	-0.279	0.103	1.54 (0.75 - 1.25)	3.7	1.67 (0.74 - 1.26)	4.4
C12	-0.042	0.131	1.12 (0.74 - 1.26)	0.9	1.12 (0.74 - 1.26)	0.9
C13	-0.501*		1.66 (0.74 - 1.26)	4.3	1.65 (0.74 - 1.26)	4.2
*Fixed parameter; SE: standard error; CI: confidence interval						

Note: Authors' data

The raters exhibited significant differences in severity, $\chi^2(12) = 723.1$; $p < .001$, with the reliability of the separation index between raters being very high (R = 0.985). The range of severity is 3.5 *logits* ($\lambda_{C07} = 2.4$ vs. $\lambda_{C10} = -1.1$), while the standard deviation of the latent distribution, expressed as the square root of the sample variance, is 1.6 *logits*. Even after eliminating rater C07 the differences are still large, as the distance between the other most severe raters (C02 and C04) and the most lenient (C10, C05 and C08) is around 1.9 *logits*. So, depending on the severity of the rater, the same student's score may vary by more than one standard deviation of the latent distribution.

The data do not indicate the existence of halo or central tendency effects among the raters. The lowest value of fit is for C05 (MNSQ-Uw = 0.80; MNSQ-W = 0.77) although it is much higher than the critical value of MNSQ = 0.5. Therefore we may discount over-consistency in the raters as a consequence of poor management of the rubric. On the contrary, the indices of fit indicate that three raters (C10, C11 and C13) had inconsistent handling of the rubrics. In those three cases MNSQ is

Received: 16-11-2018 Accepted: 21-06-2019

over 1.5, higher than the most lenient of Eckes' (2005) criteria. In each case, the level of poor fit is small, so much so that when the criteria from Wolfe and McVay (2012) are applied, only the values from C13 seem to have poor fit, with their scores exhibiting greater variation than expected by the model.

Rubric effects

Table V shows the parameters of the cognitive processes (or grading criteria) and indicates that the raters had been able to discriminate between them. The overall differences between processes were significant, $\chi^2(2) = 532.6$; $p < .001$, and the reliability of the separation index between criteria was very high (R = 986).

TABLE V. Rubric effects

Criteria	Difficulty	SE of difficulty	Unweighted fit (Uw-fit)		Weighted fit (W-fit)	
			MNSQ (CI)	T	MNSQ (CI)	T
Coherence	0.415	0.042	1.11 (0.86 – 1.14)	1.5	1.10 (0.86 – 1.14)	1.3
Cohesion	0.931	0.045	1.03 (0.86 – 1.14)	0.4	1.05 (0.86 – 1.14)	0.6
Presentation	-1.345*		1.42 (0.86 – 1.14)	5.1	1.51 (0.84 – 1.16)	5.4
*Fixed parameter; SE: standard error; CI: confidence interval						

Note: Authors' data

The values indicate that the most difficult cognitive process was cohesion ($\lambda = 0.9$), while the easiest was appropriateness and presentation ($\lambda = -1.3$). This range is reasonable given that the evaluation standards underlying the latter process (persuasive-argument text, tidy, following margins etc.) are well consolidated in the majority of 12-year-old students. Nonetheless, this process seems to have a poor fit (MNSQ > 1.3), which

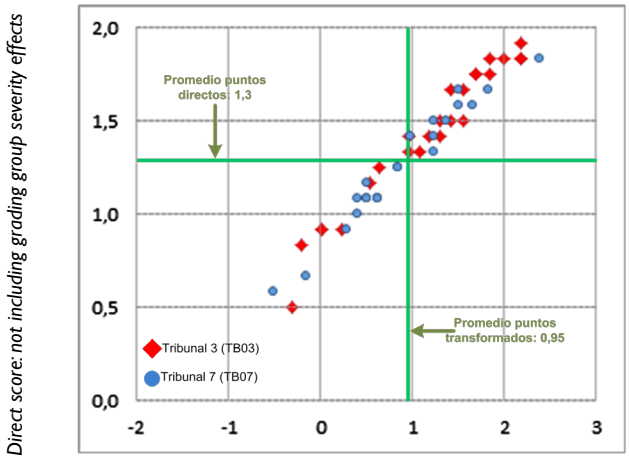
indicates that the raters did not make consistent use of the rubric for this process. The high MNSQ values indicate that this process is less discriminant than the other two, and that the raters tended to give grades which were at the extremes (0 points or 2 points). In particular the grades were unexpectedly biased high (2 points), which is why the MNSQ value is over 1.3 points. The initial principal component analysis indicated that the factorial weights of the Organization and Cohesion processes were double that of the appropriateness and presentation process. This would seem to indicate that the latter process, from a dimensional perspective of the scale, is evaluating different questions to the others. On removing the appropriateness and presentation scale, the fit of the model improves greatly, and even the values of poor fit for the raters disappear. This seems to indicate that the poor fit found in the raters could be due to a fault in the correct specification of the grading rubrics rather than inconsistent use of those rubrics.

Rater-added bias

The clearest rater effects, as previously stated, are the differences in rater severity. As raters were assigned to rating groups using a matrix design, it is easy to say that some groups have been stricter, and others more lenient. Figures I and II illustrate the results of not considering the effect of group severity on student grades. In the figures, each point indicates the position of a student, comparing their score expressed as points directly without correcting for the effect of the grading group (vertical axis) with the *logit* score for the same student, considering the severity of the group assigned to grade that student (horizontal axis). To make it easier to interpret the results, the means in both figures have been included for the classical scores (1.3 points) and in a transformed score, corrected for group severity (0.95 *logits*), dividing the cartesian space into four quadrants.

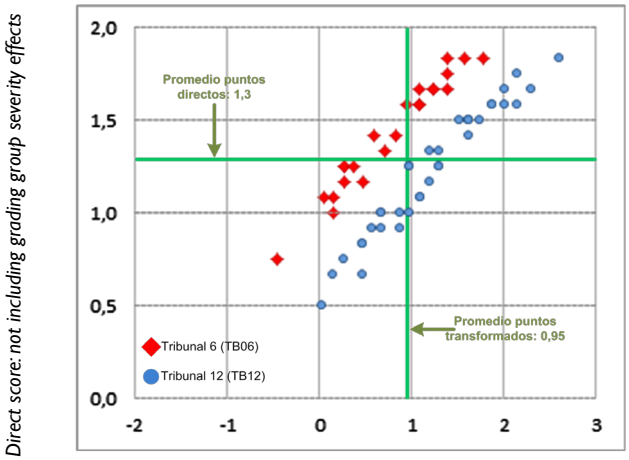
Received: 16-11-2018 Accepted: 21-06-2019

FIGURE I. Comparison of direct and transformed scores for students graded by TB03 and TB07



Note: Authors' data

FIGURE II. Comparison of direct and transformed scores for students graded by TB06 and TB12



Note: Authors' data

Figure I compares the scores of grading groups 3 and 7, in which the level of severity of the correctors was very similar (λ_{TB03} = 0.013 and λ_{TB07} = 0.006). The two sets of points appear to be superimposed, indicating that the students in the bottom-left quadrant, those with a classical score below the mean, also have below-average scores in the model which corrects for the grading group severity. Similarly, the points in the upper-right quadrant are those students with higher than average direct scores who also have higher than average results when grading group severity is taken into account. Figure II compares the results of two more varied grading groups: TB06 and TB12, which were the strictest and the most lenient of the groups respectively. Most students in TB06 are at or above average in classical scores, while in TB12 the situation is the opposite, most students are clearly punished by being assigned to a group which the four raters were collectively the strictest. However, when the effect of the grader is taken into account the situation changes dramatically. The red points in the top-left quadrant in figure II are those cases who have been clearly advantaged by the leniency of TB06, as in classical terms they have scored above the test average. However, scoring 1.3 in TB06 is not the same as in TB12, when the model corrects for grading group difficulty, students who score 1.3 points in the lenient grading group find themselves below the average in the transformed models (by approximately 0.3 *logits*), while students with similar scores in the stricter group score between 1 and 1.3 *logits*, in other words they are at or above the mean transformed score.

Taking the cutoff points from the classical and corrected means (1.3 points and 0.95 *logits* respectively) to decide the students' competency, we find that 41 of the 375 cases, or 15.4%, would have had their decision changed. In 29 cases students had above-average classical scores, but were below the average in transformed points. They had benefited from the leniency of the group they were assigned to. The 12 remaining cases (8 of whom were in TB12) were the opposite, the severity of the grading group put them below average in classical terms while their corrected scores were above average.

Received: 16-11-2018 Accepted: 21-06-2019

Discussion and conclusions

One of the difficulties of high-stakes educational evaluations is ensuring objectivity and fairness in the scores. Objective and rigorous grading of writings and constructed item responses is a true challenge. Various methodological solutions are needed in order to achieve this objective. Firstly, systematic training needs to be given to the raters and secondly, appropriate linked block design needs to be used to assign raters to groups (Frey, et al., 2009; OECD, 2014). This study demonstrates the potential of balanced, incomplete designs to control the effect of assigning raters to groups. Fernández-Alonso & Muñiz (2011) indicated the advantages and disadvantages of using complete and incomplete designs. However the 13-block Youden square used in this study brings together the robustness and consistency of the Latin square with the balance and efficiency of balanced, incomplete designs.

From a psychometric point of view, the MFRM models used exhibit enormous possibilities for carrying out integrated diagnostics of the grading rubrics, with significant results related to the four study objectives. With respect to the first objective, the interaction model between raters and cognitive processes (criteria) demonstrated the best fit, identifying differences between raters in interpreting the rubrics of the cognitive processes being evaluated. This result is in line with Adams and Wu (2010) and Eckes (2005), in which the more complex models had better fit than the more parsimonious principal effects model.

In terms of the second objective, analysis with MFRM demonstrated its capacity to detect grading bias, finding wide variation in the levels of rater severity, estimated at more than one standard deviation in the scoring scale, even when discounting the effect of the most extreme raters. This data is in line with previous research indicating that between-rater differences are the norm and not the exception (Congdon & McQueen, 2000; Eckes, 2005; Lopes et al., 2015; Lunz et al., 1990; Prieto-Adánez, 2011; Prieto-Adánez & Nieto, 2014; Wolfe, 2004; Wolfe & McVay, 2012). Poor fit and inconsistent scores were identified in three of the thirteen raters, although according to some of the criteria in the literature (Wolfe & McVay, 2012) they would be within the expected limits of poor fit, except for one case. Nonetheless, we found indications that this anomaly may not be due so much to low intra-rater reliability, but rather to a structural problem with the rubric itself.

The analysis of the evaluation criteria rubrics (the third objective) indicated a mismatch in the process Appropriateness and Presentation, which was the least discriminant process and had the most polarized scores which were unexpectedly biased towards the top end of the scale. The exploratory analysis of the principal components indicated that the inclusion of this process could be problematic and that its rubrics should be reviewed, as when this analysis was removed, all of the model parameters (including rater fit) gave appropriate values. The literature has documented this kind of problem, either because the criteria are reiterative (Adams & Wu, 2010), or for a lack of consistency between them (Basturk, 2008; Lunz et al., 1990), which shows the possibilities offered by MFRM analysis for detecting design errors in grading rubrics.

The final study objective was to demonstrate the practical advantage of MFRM, and in particular the possibility of having student scores corrected for severity in raters or rating groups. Previous research has shown that including rater-effects can result in significant variation in some students' scores (Lunz et al., 1990; Wang & Yao, 2013). In our study, the data confirms that students with an identical score can be separated by approximately one *logit* when the model considers rating group severity. We also found that the result of the test in 15.4% of cases could change depending on whether the effect of rater-group severity is taken into account or not. This result is similar to Eckes (2005), who reanalyzed the results of the TestDAF (a language certification test in Germany) and estimated that 13.5% of the writings and 17.1% of the speaking tests would have had the result of the test changed (certified or not certified) had the rater-effect been taken into account.

The data in this study have significant educational implications. On the one hand, they demonstrate the possibilities of MFRM models for identifying rater effects and possible defects in the definition of the variable to study or in grading rubrics. They also show that when differences in rater severity are taken into account, the results may vary, occasionally significantly. In Spain, there are various high-stakes exams such as university entrance exams and language certification in which the effects of rater severity will have an impact on student results. There is research looking at rater consistency in both of these types of exams (Amengual & Méndez García, 2012; Cuxart- Jardí, 2000; Cuxart-Jardí et al., 1997; Grau, Cuxart & Martí-Recober, 2002; Prieto-Adánez, 2011; Prieto-Adánez & Nieto, 2014; Suárez-Álvarez, González-Prieto,

Received: 16-11-2018 Accepted: 21-06-2019

Fernández-Alonso, Gil & Muñiz, 2014), however, as far as we are aware, there is no research that has focused on the practical consequences to the student of having their grading assigned to raters who probably have varied levels of severity.

Given the data provided by this study and bearing in mind the high impact of these types of tests on students' academic and personal futures, it seems necessary to explore these processes of grading and scoring in more detail. It would also be important to examine the possible differential functioning of items induced by raters and rubrics (Gómez-Benito, Sireci, Padilla, Hidalgo & Benítez, 2018; Woitschach, Zumbo & Fernández-Alonso, 2019), as well as checking the generalization of the results to evaluations of students with different characteristics (Amor, Verdugo, Calvo, Navas & Aguayo, 2018). Besides the aforementioned analyses, one fundamental aspect is the development of systematic training and practice for evaluators in order to minimize the individual differences related to their grading and scoring strategies. In short, this means complementing training and practice with procedures for analyzing the results, using the most current, sophisticated psychometric and statistical models. One would hope that this pre-post combination would reduce to a minimum the errors in rigorous, objective student evaluation, an unavoidable ethical and deontological requirement

This work is not without its limitations, the results should be taken with caution, and should not be universally generalized without more thought. There are three clear sources of error which may affect the results of this kind of work. Firstly, the type of evaluator, whose personal characteristics, training, background and involvement can vary from one to another. The subject being evaluated may also lead to restrictions that could affect the results. Finally, the design of the rubrics themselves may influence the results of the evaluations, and may for example, minimize or maximize the variance in the resulting scores.

References

Adams, R. & Wu, M. (2010). *The analysis of rater effects*. Recuperado de: https://www.acer.org/files/Conquest-Tutorial-3-RaterEffects.pdf

Amengual, M. (2004). *Análisis de la fiabilidad en las puntuaciones holísticas en ítems abiertos* [Reliability analysis in holistic scores on

Received: 16-11-2018 Accepted: 21-06-2019

constructed response ítems]. Universidad Complutense de Madrid: Tesis doctoral no publicada. Recuperado de: http://biblioteca.ucm.es/tesis/fll/ucm-t26663.pdf

Amengual, M. & Méndez García, M. C. (2012). Implementing the oral English task in the spanish university admission examination: An international perspective of the language. *Revista de Educación*, 357, 105-127.

Amor, A. M., Verdugo, M. A., Calvo, M. I., Navas, P. & Aguayo, V. (2018). Psychoeducational assessment of students with intellectual disability: professional-action framework analysis. *Psicothema, 30*, 39-45. doi: 10.7334/psicothema2017.175

Ato, M. & Vallejo, G. (2007). *Diseños experimentales en psicología* [Experimental designs in psychology]. Madrid: Pirámide.

Baird, J. A., Hayes, M., Johnson, R., Johnson, S. & Lamprianou, I. (2013). *Marker effects and examination reliability. A comparative exploration from the perspectives of generalizability theory, Rasch modelling and multilevel modelling (Ofqual/13/5261)*. Coventry: Office of Qualifications and Examinations Regulation. Recuperado de: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/378059/2013-01-21-marker-effects-and-examination-reliability.pdf

Baird, J. A., Meadows, M., Leckie, G. & Caro, D. (2017). Rater accuracy and training group effects in Expert- and Supervisor-based monitoring systems. *Assessment in Education: Principles, Policy and Practice*, *24*(1), 44-59. doi: 10.1080/0969594X.2015.1108283

Basturk, R. (2008). Applying the many-facet Rasch model to evaluate PowerPoint presentation performance in higher education. *Assessment y Evaluation in Higher Education*, *33*(4), 431- 444. doi: 10.1080/02602930701562775

Bejar, I. I. (2012): Rater cognition: Implications for validity. *Educational Measurement*, *31*(3), 2-9.

Crisp, V. (2012). An investigation of rater cognition in the assessment of projects. *Educational Measurement*, *31*(3), 10-20.

Cochran, W.G. & Cox, G.M. (1974). *Diseños experimentales* [Experimental designs]. México: Trillas. (orig. 1957).

Congdon, P. J. & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, *37*(2), 163-178. doi: 10.1111/j.1745-3984.2000.tb01081.x

Cuxart-Jardí, A. (2000). Modelos estadísticos y evaluación: tres estudios en educación. [Statistical models and assessment: three studies in education]. *Revista de Educación*, 323, pp. 369-394

Cuxart-Jardí, A., Martí-Recober, M. & Ferrer-Juliá, F. (1997). Algunos factores que inciden en el rendimiento y la evaluación en los alumnos de las pruebas de aptitud de acceso a la universidad [Some factors that affect the students' performance in the university entrance exams]. *Revista de Educación*, 314, 63-88.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A multi-faceted Rasch analysis. *Language Assessment Quarterly*, *2*(3), 197-221. doi: 10.1207/s15434311laq0203_2

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Section H). Strasbourg, France: Council of Europe/Language Policy. Recuperado de: https://rm.coe.int/1680667a23#search=eckes

European Commission/EACEA/Eurydice (2009*). National testing of pupils in Europe: Objectives, organization and use of results*, Luxembourg: Publications Office of the European Union. doi: 10.2797/18294

Engelhard, G. (1992). The measurement of writing ability with a multi-faceted Rasch model. *Applied Measurement in Education*, *5*(3), 171-191. doi: 10.1207/s15324818ame0503_1

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a multi-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93-112. doi: 10.1111/j.1745-3984.1994.tb00436.x

Fernández-Alonso, R. & Muñiz, J. (2011). Diseño de cuadernillos para la evaluación de las competencias básicas [Booklet designs for the evaluation of basic skills]. *Aula Abierta*, *39*(2), 3-34.

Frey, A., Hartig, J. & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28*(3), 39-53.

Gómez-Benito, J., Sireci, S., Padilla, J. L., Hidalgo, M. D. & Benítez, I. (2018). Differential item functioning: Beyond validity evidence based on internal structure. *Psicothema, 30*, 104-109. doi: 10.7334/psicothema2017.183

Grau, R. M., Cuxart, A. & Martí-Recober, M. (2002). La calidad en el proceso de corrección de las pruebas de acceso a la universidad: variabilidad y factores [Quality in the scoring process of university entrance exams: variability and factors]. *Revista de Investigación Educativa*, *20*(1), 209-224.

Gyagenda, I. & Engelhard, G. (2009). Using classical and modern measurement theories to explore rater, domain, and gender influences on student writing ability. *Journal of Applied Measurement*, *10*(3), 225-246.

Jonsson, A. & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130-144.

Kuo, S. A. (2007): Which rubric is more suitable for NSS liberal studies? Analytic or holistic? *Educational Research Journal*, *22*(2), 179-199.

Lallmamode, S. P., Mat Daud, N. & Abu Kassim, N. L. (2016). Writing Development and initial argument-based validation of a scoring rubric used in the assessment of L2 writing electronic portfolios. *Assessing Writing, 30*, 44-62. http://dx.doi.org/10.1016/j.asw.2016.06.001

Leckie, G. & Baird, J. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, *48*(4), 399-418. doi: 10.1111/j.1745-3984.2011.00152.x

Ley Orgánica 15/1999 de Protección de Datos de Carácter Personal [Organic Law 15/1999 on Protection of Personal Data]. *Boletín Oficial del Estado de 14 de diciembre de 1999.* Recuperado de: https://www.boe.es/buscar/doc.php?id=BOE-A-1999-23750

Ley Orgánica 8/2013 para la Mejora de Calidad Educativa [Organic Law 8/2013 for the Improvement of Educational Quality]. *Boletín Oficial del Estado de 10 de diciembre de 2013*. Recuperado de: https://www.boe.es/buscar/pdf/2013/BOE-A-2013-12886-consolidado.pdf

Linacre, J. M., Engelhard, G., Tatum, D. S. & Myford, C. M. (1994) Measurement with judges: Many-faceted conjoint measurement. *International Journal of Educational Research*, *21*(6), 569-577. doi: 10.1016/0883-0355(94)90011-6

Lopes Toffoli, S. F., de Andrade, D. F. & Bornia, A. C., (2015): Evaluation of open items using the many-facet Rasch model, *Journal of Applied Statistics*, doi: 10.1080/02664763.2015.1049938

Received: 16-11-2018 Accepted: 21-06-2019

Lunz, M. E. & Stahl, J. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions*, *13*(4), 425-444. doi: 10.1177/016327879001300405

Lunz, M. E., Wright, B. D. & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, *3*(4), 331-345. doi: 10.1207/s15324818ame0304_3

Mahmood, O., Dagnæs, J., Bube, S., Rohrsted, M. & Konge, L. (2017). Nonspecialist raters can provide reliable assessments of procedural skills. *Journal of Surgical*. Recuperado de: http://dx.doi.org/10.1016/j.jsurg.2017.07.003

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman

Ministerio de Educación, Cultura & Deporte (2016). *Pruebas de la evaluación final de Educación Primaria. Curso 2015-2016* [Tests of the final evaluation of Primary Education. Course 2015-2016]. Madrid: Instituto de Evaluación

Myford, C. M. & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*, 386-422

Myford, C. M. & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*, 189-227

Organisation for Economic Co-operation and Development [OECD]. (2014). *PISA 2012 Technical Report*. Paris: OECD Publishing. Recuperado de: https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

Park, T. (2010). *An investigation of an ESL placement test of writing using multi-faceted Rasch measurement*. Teachers College, Columbia University Working Papers in TESOL and Applied Linguistics, 4(1), 1-19

Prieto-Adánez, G. (2011). Evaluación de la ejecución mediante el modelo Many Facet Rasch Measurement [Performance assessment using the Many-Facet Rasch Measurement]. *Psicothema*, *23*, 233-238

Prieto-Adánez, G. (2015). Análisis de un test de desempeño en expresión escrita mediante el modelo de MFRM [Analysis of a writing test with the MFRM model]. *Actualidades en Psicología, 29*(119), 1-17. http://dx.doi.org/10.15517/ap.v29i119.19822

Received: 16-11-2018 Accepted: 21-06-2019

Prieto-Adánez, G. & Nieto, E. (2014). Analysis of rater severity on written expression exam using Many Faceted Rasch Measurement. *Psicológica*, 35, 385-397

Saal, F. E., Downey, R. G. & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*(2), 413-428

Shackleton, C. (2018). Linking the University of Granada CertAcles listening test to the CEFR. *Revista de Educación, 381*, 35-60. doi: 10.4438/1988-592X-RE-2017-381-380

Suárez Álvarez, J., González-Prieto, C., Fernández Alonso, R., Gil, G. & Muñiz, J. (2014). Psychometric assessment of oral expression in English language in the University Entrance Examination. *Revista de Educación*, 364, 93-118. doi: 10.4438/1988-592X-RE-2014-364-256

Sudweeks, R. R., Reeve, S. & Bradshaw, W. S. (2005). A comparison of generalizability theory and many facet measurement in an analysis of college sophomore writing. *Assessing Writing, 9*, 239-261.

Suto, I. (2012). A critical review of some qualitative research methods used to explore rater cognition. *Educational Measurement, 31* (3), 21-30.

Wang, Z. & Yao, L. (2013). The effects of rater severity and rater distribution on examinees' ability estimation for constructed-response items. *ETS Research Report Series*, i-22. doi:10.1002/j.2333-8504.2013.tb02330.x

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, *46*(1), 35-51.

Wolfe, E. W. & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement*, *31* (3), 31-37.

Woitschach, P., Díaz-Pérez, C., Fernández-Argüelles, D., Fernández-Castañón, J., Fernández-Castillo, A., Fernández-Rodríguez, L., ... & Fernández-Alonso, R. (2018). Efectos del corrector en las evaluaciones educativas de alto impacto. [Rater effects in high-impact educational assessments]. *REMA Revista Electrónica de Metodología Aplicada, 23*(1), 12-27.

Woitschach, P., Zumbo, B. D. & Fernández-Alonso, R. (2019). An ecological view of measurement: Focus on multilevel model explanation of differential item functioning. *Psicothema, 31*(2), 194-203. doi:10.7334/psicothema2018.303

Received: 16-11-2018 Accepted: 21-06-2019

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. A. (2007). *ACER ConQuest 2.0: generalised item response modelling software*. Camberwell, Victoria: Australian Council for Educational Research.

Contact address: Rubén Fernández-Alonso. Consejería de Educación y Cultura del Gobierno del Principado de Asturias y Universidad de Oviedo, Facultad de Formación del Profesorado y Educación, Departamento de Ciencias de la Educación. Calle Aniceto Sela, 1, 33005 Oviedo, Asturias. E-mail: fernandezaruben@uniovi.es

Received: 16-11-2018 Accepted: 21-06-2019