Actualidades en Psicología, 29(119), 2015, 1-17
http://revistas.ucr.ac.cr/index.php/actualidades

ISSN 2215-3535
DOI: http://dx.doi.org/10.15517/ap.v29i119.19822

Análisis de un test de desempeño en expresión escrita mediante el modelo de MFRM*

Analysis of a Writing Test with the MFRM Model

Gerardo Prieto-Adánez [1]
Universidad de Salamanca, España

Resumen. Este trabajo muestra la utilidad de un modelo de Rasch (Many-Facet Rasch Measurement, MFRM) para medir la competencia de los examinados, la severidad de los calificadores, la dificultad de las tareas y de las variables puntuadas en las pruebas de respuesta construida que se suelen emplear para evaluar el desempeño. Inicialmente se describe el modelo y sus estadísticos básicos. Finalmente se presenta un ejemplo ilustrativo en el que se analizan, mediante el programa FACETS, las fuentes de la variabilidad de las calificaciones de los estudiantes en un test de expresión escrita. Los resultados muestran que el procedimiento es útil para detectar a los calificadores que presentan valores extremos en la variable severidad/benignidad y para obtener calificaciones objetivas de los examinados (libres de la severidad del calificador).

Palabras clave. Evaluación con calificadores, Evaluación del desempeño, Evaluación de la expresión escrita, Modelo de Rasch, Modelo de Muchas Facetas.

Abstract. This paper describes how a Rasch model (Many-Facet Rasch Measurement) can be applied to performance assessment focusing on analysis of examinee, raters, tasks and variables. The article provides an introduction to MFRM, a description of analysis procedures, and an illustrative example to examine the effects of various sources of variability on students' performance on a writing test by means of the FACETS program. Results highlight the usefulness of the MFRM to detect raters that have extreme values on the continuum of severity/leniency as well as providing objective measurement of examinee (scores free of rater severity).

Keywords. Rater-Mediated Assessment, Performance Assessment, Writing Assessment, Rasch Model, Many-Facet Rasch Measurement.

[1]Gerardo Prieto-Adánez. Facultad de Psicología. Universidad de Salamanca, España. Dirección postal: Avenida de la Merced, 109-131, 37005 Salamanca, España. Email: gprieto@usal.es.

*Agradecimientos. Instituto Cervantes y Cursos Internacionales de la Universidad de Salamanca, S.A.U.

Introducción

Aunque el formato de elección múltiple sigue siendo muy utilizado en la evaluación educativa por razones prácticas (economía en el tiempo de administración, corrección objetiva y automática, amplio muestreo de un dominio, etc.), se han destacado sus limitaciones para evaluar las habilidades de nivel cognitivo superior y para estimar el nivel de destreza de las personas en contextos realistas (Martínez-Arias, 2010). Por ello, en la última década del pasado siglo los procedimientos de evaluación educativa incorporaron a los clásicos tests de elección múltiple los denominados tests de desempeño que se utilizan para evaluar competencias complejas como, por ejemplo, la expresión escrita y oral en el contexto del aprendizaje de lenguas (Hambleton, 2000).

Los tests de desempeño (performance assessment) son procedimientos de evaluación en los que se demanda a los sujetos que lleven a cabo tareas en las que han de demostrar su capacidad para aplicar conocimientos y destrezas en situaciones similares a las de la vida real (Martínez-Arias, 2010). Destaca en este tipo de pruebas su formato abierto: las respuestas, que son estructuradas o construidas por el examinado, han de ser cuantificadas mediante una escala de categorías ordenadas (rating scale) por un calificador. Obviamente la influencia del criterio del calificador en la puntuación otorgada es determinante, por lo que se ha utilizado la denominación de *evaluación mediada por el calificador* (rater mediated assessment) para caracterizar el aspecto central de este tipo de metodología (Engelhard, 2002). Es decir, la magnitud de las calificaciones que reciben los examinados no dependen sólo de su nivel de competencia, sino que hay considerar otras facetas que la influyen notablemente como la dificultad de las tareas y de los atributos a evaluar, la severidad del calificador y su uso de las rubricas o categorías de calificación. Sin duda, el comportamiento de los calificadores ha de ser tomado en consideración si se desea evaluar de forma válida y justa el constructo de interés (Lane & Stone, 2006).

El objetivo de este trabajo es mostrar la utilidad de uno de los modelos de Rasch, denominado Many-Facet Rasch Measurement (Linacre, 1989), para medir en una dimensión común los elementos de las facetas que suelen estar incluidas en el marco de una evaluación del desempeño: los examinados, los calificadores, las tareas y los atributos evaluados.

Como un ejemplo del uso del modelo se presenta el análisis de una prueba de Expresión Escrita integrada en un examen para la obtención del Diploma de Español como Lengua Extranjera (DELE) de Nivel B. Los DELE son títulos oficiales que emite el Instituto Cervantes en nombre del Ministerio de Educación, Política Social y Deporte del Gobierno de España. El Diploma de Español de nivel B1 acredita la capacidad del usuario de la lengua para comprender los puntos principales de textos orales y escritos en variedades normalizadas de la lengua que versen sobre asuntos conocidos, ya sean estos relacionados con el trabajo, el estudio o la vida cotidiana; para desenvolverse en la mayoría de las situaciones y contextos en que se inscriben estos ámbitos de uso, y para producir asimismo textos sencillos y coherentes sobre temas conocidos o que sean de interés personal, tales como la descripción de experiencias, acontecimientos, deseos, planes y aspiraciones o la expresión de opiniones. El modelo Many-Facet Rasch Measurement se ha utilizado extensamente para analizar exámenes de expresión oral y escrita del inglés y del español (Eckes, 2011; Gyagenda & Engelhard, 2010; Kondo-Brown, 2002; Park, 2004; Prieto, 2011; Prieto & Nieto, 2014; Tyndall & Kenyon, 1996; Wolfe, 2009).

El modelo Many-Facet Rasch Measurement (MFRM)

MFRM fue desarrollado por Linacre (1989) para extender el Modelo Dicotómico de Rasch (1960), el Modelo de Escalas de Calificación (Andrich, 1978) y el Modelo de Crédito Parcial (Masters, 1982) a las evaluaciones en las que uno o varios calificadores puntúan el desempeño de una persona en una tarea de formato abierto. El objetivo de los modelos de Rasch, Andrich y Masters es medir conjuntamente en una dimensión los elementos de dos facetas: las personas y los ítems. La propiedad principal de esta familia de modelos, denominada *objetividad específica* por Rasch (1977), es el fundamento de la invarianza de las medidas.

De acuerdo con Engelhard (2013), la medición invariante de personas e ítems se puede lograr si se cumplen cinco requisitos: (a) las medidas de las personas deben ser independientes de los ítems específicos utilizados para medir, (b) una persona con mayor competencia (o aptitud) debe de tener siempre una mayor probabilidad de éxito en un ítem que una persona menos competente (las curvas características de las personas no se cruzan), (c) las medidas de los ítems han de ser independientes de las personas empleadas para la calibración, (d) cualquier persona habrá de tener mayor probabilidad de éxito en un ítem fácil que en un ítem más difícil (las curvas características de los ítems no se cruzan) y (e) los ítems y las personas deben ser localizados simultáneamente en una única dimensión latente (mapa de la variable). El objetivo de MFRM es incorporar más facetas al marco de medición: tal es el caso de los calificadores (evaluadores).

La incorporación de esta faceta implica redefinir las condiciones anteriores para lograr la invarianza de las medidas: (a) la medición de las personas ha de ser independiente de los calificadores específicos que han intervenido en la medición, (b) una persona con mayor competencia tendrá mayor probabilidad de obtener de los evaluadores mayores calificaciones que una persona con menor competencia (las curvas características de las personas no se cruzan), (c) la calibración de los calificadores ha de ser independiente de las personas a las que han evaluado, (d) cualquier persona ha de tener mayor probabilidad de obtener una puntuación alta de los calificadores benévolos que de los calificadores más severos (las curvas características de los calificadores no se cruzan) y (e) los calificadores y las personas deben ser localizados simultáneamente en una única dimensión latente (mapa de la variable).

El cumplimiento de los requisitos enunciados anteriormente en un conjunto de datos puede ser contrastado mediante los estadísticos de ajuste que se describen más adelante. Si los datos se ajustasen aceptablemente al modelo, las medidas pueden recibir algunas interpretaciones especialmente útiles y deseables: las calificaciones de los examinados son independientes de la severidad de los calificadores que les han evaluado y los valores de los calificadores en la dimensión denominada severidad/benevolencia son independientes de la competencia de los alumnos evaluados.

MFRM puede ser expresado de distintas formas en función del número de facetas y de los objetivos del estudio. Una expresión muy común se basa en una extensión del Modelo de Escalas de Calificación a una situación en la que existen más de dos facetas que contribuyen a la variabilidad de las medidas (examinados, calificadores, atributos evaluados, tareas, etc.). En este caso, se trata de un modelo lineal aditivo basado en una transformación logística de los cocientes entre las probabilidades de que una persona, puntuada mediante una escala de categorías numéricas, reciba una puntuación o la inmediatamente inferior.

En concreto,

$$\ln\left(P_{nijlk} / P_{nijl\,(k-1)}\right) = B_n - D_i - R_j - C_l - F_k \qquad (1)$$

siendo,

P_{nijlk} = probabilidad de que el examinado *n* reciba del calificador *j* la puntuación *k* en el atributo *l* de la tarea *i,*

$P_{nijl(k-1)}$ = probabilidad de que el examinado n reciba del calificador j la puntuación inferior *(k-1)* en el atributo *l* de la tarea *i,*

B_n = nivel en el atributo del examinado *n,*

D_i = dificultad de la tarea *i,*

R_j = severidad del calificador *j,*

C_l = dificultad del atributo *l,* y

F_k = dificultad del punto en el que las categorías *k y k-1* son equiprobables.

En la ecuación 1, el logit ($\ln (P_{nijk} / P_{nij(k-1)})$) es la variable dependiente y las diversas facetas (personas, tareas, calificadores, atributos, etc) son las variables independientes. Es decir, el modelo especifica que la probabilidad de que el calificador *j* otorgue a una persona *n* una calificación (*k*) en lugar de la inferior (*k-1*) en atributo *l* depende de los efectos aditivos de la dificultad de la tarea (D_i), de la severidad del calificador (R_j), de la competencia de la persona (B_n), de la dificultad del atributo evaluado (C_l) y del valor del paso entre las categorías *k y k-1* (F_k). El paso o umbral no es considerado una faceta del modelo y en esta formulación se asume que es invariante en los distintos calificadores, tareas y dominios. Cuando no se asume la invarianza de los pasos (suponiendo, por ejemplo, que los calificadores difieren en el uso de las rúbricas) la formulación de MFRM es una extensión del Modelo de Crédito Parcial con el que se desea analizar las peculiaridades de los calificadores al usar las categorías numéricas:

$$\ln \left(P_{nijlk} / P_{nijl(k-1)} \right) = B_n - D_i - R_j - C_l - F_{jk} \qquad (2)$$

En (2) se asume que los pasos (F_{jk}) pueden variar entre los calificadores.

Mediante MFRM, los parámetros de cada faceta pueden ser estimados independientemente del resto de las facetas en una escala común (la escala logit). Las sumas de las puntuaciones directas son los estadísticos suficientes para estimar los parámetros (Linacre & Wright, 2002). La escala logit puede oscilar entre 0 ± ∞. El punto 0 se fija convencionalmente en el nivel medio de los ítems, de los calificadores, de las tareas y de los atributos, permitiendo la variación libre en la escala común de las personas evaluadas.

Para cada elemento de cada faceta, el análisis facilita una medida en logit, un error típico de medida (SE=la precisión del valor estimado) e índices de ajuste entre las respuestas observadas y las predichas por el modelo.

Aunque lo más común es utilizar la escala logit, puede ser útil presentar los valores en una escala denominada *promedio imparcial* (fair average). El promedio imparcial (M_I), el caso de los examinados, es la media de las puntuaciones que otorgaría a una persona un calificador con un nivel promedio de severidad (Eckes, 2011). Para calcular cada puntuación esperada para una persona *n*, se fijan en la media (*M*) los parámetros de todas las facetas excepto la de la persona. Es decir, la ecuación (1) se reformularía de la siguiente forma:

$$\ln \left(P_{nijlk} / P_{nijl(k-1)} \right) = B_n - D_M - R_M - C_M - F_k \qquad (3)$$

Por tanto, P_{nijlk} es la probabilidad de que la persona examinada *n* reciba la puntuación de la categoría *k* en una tarea y un atributo de dificultad media por un calificador con un nivel de severidad promedio.

En este caso, M_I para el sujeto *n* sería igual a:

$$M_I = \sum r p_{nr} \qquad (4)$$

Siendo *r* la puntuación en la escala de categorías recibidas por un examinado *n*. La media debe de contabilizar las puntuaciones de cada persona en las tareas, los atributos y los calificadores. M_I es un valor en la escala de puntuaciones brutas asignadas a las categoría (0-3, por ejemplo) por lo que los valores son de más fácil interpretación que los logit.

De forma similar es posible obtener el promedio imparcial de cada calificador para puntuar su nivel de severidad. En este caso, M_I es la media de las puntuaciones que otorgaría un calificador a un examinado con un nivel medio de competencia. Además de estos estadísticos a nivel individual, es posible obtener estadísticos grupales indicativos del ajuste promedio, la media, la variabilidad y la fiabilidad de las medidas de las personas, los ítems y los calificadores (Myford & Wolfe, 2004a).

Si los datos se ajustan al modelo, las medidas tienen unas características muy deseables: invarianza de las medidas de las personas, los calificadores, las tareas y los atributos, la medición con propiedades de intervalo y la cuantificación de la precisión a nivel local (Prieto & Delgado, 2003). Los análisis

pueden llevarse a cabo con el programa FACETS (Linacre, 2015).

Estadísticos básicos

Índices de ajuste. Indican el grado en el que las calificaciones observadas se diferencian de las esperadas. Una calificación observada es la otorgada por un calificador a un evaluado en un atributo. Una calificación esperada es la predicha por el modelo, dado el nivel del examinado, la severidad del calificador y la dificultad de la tarea. Los índices de ajuste son medias de los cuadrados de las diferencias estandarizadas: *Outfit* es la media no ponderada de estos valores (muy sensible a desajustes extremos) e *Infit*, la media de los valores ponderados con la función de información (Wolfe, 2009).

Ambos estadísticos tienen un valor esperado de 1 y pueden oscilar entre 0 e infinito. Los valores menores que 1 revelan que los residuos (diferencias entre los valores observados y esperados) son menores que los esperados por azar (es decir, se puede interpretar como sobreajuste).

Son los valores superiores a 1 los que manifiestan más desajuste de lo esperado. Convencionalmente, se considera que los valores que oscilan entre 0.5 y 1.5 indican un desajuste muy pequeño y que los superiores a 2 revelan un desajuste severo que degrada las medidas (Linacre, 2010). Sin embargo, otros criterios más estrictos indican que Infit y Outfit deberían ser menores a 1.2 y 1.7 respectivamente para obtener una medición productiva (Smith, Schumaker & Bush, 1998). FACETS aporta valores individuales de ajuste para los evaluados, los calificadores, los ítems y las categorías de calificación.

Contraste Chi-cuadrado de la inexistencia de diferencias en una faceta (χ^2)

Mediante el estadístico Chi-cuadrado se puede contrastar la hipótesis nula de que las estimaciones de los parámetros de los elementos de una faceta no difieren significativamente entre sí. Por ejemplo, en la faceta "calificador", se puede contrastar la hipótesis de que, contabilizando el error de medida, todos los calificadores ejercen el mismo nivel de severidad (Myford & Wolfe, 2004a).

Fiabilidad entre calificadores

Los índices estadísticos que se han utilizado para analizar la fiabilidad entre los calificadores pueden ser clasificados en dos categorías: *índices de consenso* e *índices de consistencia* (Eckes, 2011). Los índices de consenso reflejan el grado en el que los calificadores atribuyen las mismas calificaciones en idénticas circunstancias. En esta categoría se inscribe el Porcentaje de Acuerdo (*% Acuerdo*), aportado por FACETS, que indica el porcentaje de veces que un calificador atribuye las mismas calificaciones que otros calificadores en idénticas circunstancias (examinado, tarea, atributo, etc.). Los índices de consistencia indican el grado de asociación entre las calificaciones otorgadas por distintos calificadores. A esta categoría pertenece la *Correlación calificador-resto de los calificadores* ($R_{c,rc}$) que cuantifica el grado en el que las evaluaciones de cada calificador son consistentes con las del resto de los calificadores. Convencionalmente, los valores inferiores a .30 permiten identificar a los evaluadores muy inconsistentes, en los que la ordenación de las personas difiere notablemente de la del resto de los calificadores.

Fiabilidad de la separación de las medidas (SR: separation reliability). Además de evaluar la precisión individual de las medidas mediante el error estándar (de cada persona, cada calificador o cada ítem), FACETS proporciona evaluaciones de la fiabilidad a nivel de grupo. SR es un índice empleado para evaluar la fiabilidad de las puntuaciones de las distintas facetas (las personas, las tareas, los atributos o los calificadores) que refleja cuál es la proporción de la varianza verdadera respecto de la varianza observada de las medidas.

Las interpretaciones sustantivas de SR difieren entre las facetas (Myford & Wolfe, 2004a). En el caso de las medidas de las personas, PSR (Person separation reliability) es comparable al coeficiente alfa empleado en la Teoría Clásica de

los Tests, indicando qué proporción de la varianza observada de las medidas de las personas es su varianza verdadera:

$$PSR = 1 - ((Media\ (SE^2_{Bn}) / Varianza\ (B_n)) \quad (5)$$

Siendo SE_{Bn}, el error estándar de medida del valor de la persona *n* en la variable.

En este caso, se esperan altos valores de PSR cuando las medidas reflejan fiablemente la variabilidad de las personas en el constructo. Dado que se suele desear que no existan variaciones sustanciales entre los calificadores en el nivel de severidad, los valores bajos de RSR (Rater separation reliability) son los aceptables (las diferencias observadas en la severidad de los calificadores serían atribuibles al error de medida).

Estadísticos de las categorías de evaluación. Para determinar si las categorías numéricas (rúbricas) son funcionales empíricamente (ordenadas y distinguibles) se toman en consideración varios indicadores: orden de los promedios en las categorías de las medidas de las personas, Outfit, y orden de los pasos entre las categorías (Linacre, 2004).

Si las categorías de evaluación funcionan adecuadamente, los promedios de las medidas (logit) de las personas que reciben una calificación deben estar ordenados monotónicamente. Este patrón de resultados revela que cuanto mayor sea la calificación recibida, mayor será el nivel de las personas en el constructo (Park, 2004). Los valores Outfit de las categorías son también un indicador de su funcionalidad.

Para cada categoría de evaluación, FACETS calcula la medida promedio de las personas incluidas en la categoría (la medida observada) y una medida esperada (el promedio esperado si los datos se ajustasen al modelo). Como se indicó con anterioridad, si el valor observado y el esperado son muy semejantes, Outfit adoptará un valor próximo a 1.0.

Los valores de Outfit superiores a 2.0 indican que la categoría de evaluación no ha sido utilizada de manera adecuada. Finalmente, se puede observar si los pasos entre las categorías están ordenados monotónicamente y suficientemente separados. El desorden de los pasos indica que existen categorías que no son las de más probable uso en ningún rango de la variable medida. Esta circunstancia se manifiesta en el aplanamiento de las curvas características de las categorías.

Método

Participantes

Realizaron el examen 948 personas con edades entre 12 y 69 años, y una media de edad de 25 años. El país de nacimiento de los participantes fue muy diverso, siendo los más frecuentes Italia (18.8%), China (13.7%), Grecia (10.5%), Japón (13.9%) y Corea del Sur (13.8%). Los examinados realizaron la prueba en el mes de julio de 2014 en centros propios o adscritos al Instituto Cervantes de más de 30 países, correspondiendo los grupos más numerosos de participantes a centros de Italia (18.4%), China (12.0%), España (11.7%), Japón (10.7%), Grecia (10.7%) y Corea del Sur (10.3%).

Las pruebas fueron evaluadas por un total de 14 calificadores (2 hombres y 12 mujeres), todos ellos profesores de Cursos Internacionales de la Universidad de Salamanca con una amplia experiencia en la enseñanza del español a alumnos extranjeros.

Prueba

La prueba de expresión escrita constaba de dos tareas. En la primera se le pidió al examinado que redactase un texto con una intención comunicativa (carta, correo electrónico, etc.) con una extensión entre 100 y 200 palabras. La segunda tarea consistió en la redacción de un texto, con una extensión entre 130 y 150 palabras, con una finalidad narrativa.

Procedimiento

Los textos de cada examinado fueron evaluados independientemente por dos calificadores, los cuáles puntuaron cada texto en cinco atributos: una calificación *holística* (que representa la impresión global del desempeño del examinado) y cuatro calificaciones analíticas. Las variables analíticas fueron: *Adecuación al género discursivo* (adaptación de la redacción al contexto), *Coherencia* (control de los

recursos necesarios para establecer relaciones entre el discurso y la situación de comunicación), *Corrección* (conocimiento y capacidad de uso de la ortografía, de las categorías gramaticales y de las reglas morfosintácticas) y *Alcance* (equilibrio de los recursos léxicos utilizados con los temas y las situaciones de comunicación). Para emitir las calificaciones se usó un sistema de rúbricas consistentes en cuatro categorías numéricas ordenadas (0-3). El valor 2 es el equivalente a la superación mínima del umbral correspondiente al nivel B1. El valor 3 supone una consecución sobrada del nivel. El valor 1 indica que el desempeño no permite superar el umbral correspondiente a la consecución del nivel. La puntuación 0 se asigna cuando el texto es ilegible o la información es irrelevante y no se ajusta al objetivo. Los calificadores recibieron instrucciones con ejemplos prototípicos de textos clasificables en cada categoría.

Dado que los textos de cada examinado no eran evaluados por todos los calificadores, se implementó un sistema de asignación de las pruebas para garantizar la conectividad que es imprescindible para lograr que todas las medidas estén en la misma escala. Es decir, debe ser posible discernir si la diferencia en las calificaciones recibidas por dos examinados se debe a que uno de ellos es más competente que el otro o a que el calificador que evaluó al primero es más benigno que el que evaluó al segundo. Una estimación óptima de la severidad de los calificadores requeriría que todos los calificadores evaluasen todas las pruebas. Obviamente, este procedimiento no puede ser empleado en muchas situaciones. Por tanto, es necesario aplicar un plan de asignación de las pruebas a los calificadores que, garantizando la conectividad, sea asequible en la práctica. Los requisitos mínimos de estos planes consisten en que al menos dos calificadores evalúen cada prueba y que cada examinado comparta un calificador con otro examinado. Existen varios diseños de asignación de las pruebas a los calificadores para garantizar la conectividad (Eckes, 2011; Wright & Stone, 1979). El diseño más simple consiste en asignar a un calificador un subconjunto pequeño de las pruebas evaluadas por cada uno de los demás calificadores (Eckes, 2011). En este trabajo se utilizó un diseño simple de rotación de los examinados y los calificadores similar al descrito por Tesio et al. (2015).

Resultados

Dada la necesaria brevedad de este artículo, se comentarán principalmente los resultados correspondientes a las facetas de examinados y de calificadores. No obstante, una inspección del mapa de la variable o mapa de Wright (Figura 1) es útil para observar las calibraciones de los elementos de todas las facetas en un único marco de referencia.

En la primera columna del mapa aparece la escala logit en la que se miden los examinados, los calificadores, las tareas, los atributos y los umbrales entre las categorías.

En la columna Examinado de la Figura 1 se representa la distribución de los examinados en la escala. Cada asterisco (*) representa a 10 personas y cada punto a una frecuencia inferior. Los candidatos con mayor nivel en la prueba de expresión escrita se sitúan en la parte superior de la columna y en la parte inferior los de menor puntuación. Se observa que el rendimiento de los examinados es elevado (2.82 logits en promedio) y que hay una gran variabilidad en su competencia (entre 11.26 y -7.19 logits). En la columna Calificador aparecen los valores de severidad de los calificadores, siendo el número 332 el más severo (1.96) y el número 39 el más benigno (-2.29). Los valores en severidad oscilan en torno a 0 (suele situarse el punto cero de la escala en la media en severidad de los calificadores). La variabilidad de los calificadores en severidad es alta (la desviación típica es 1.36 logits) y mayor de lo que sería deseable: idealmente habría de observarse que los calificadores apenas difieren entre sí en la variable severidad, un indicador de que los criterios de asignación de las calificaciones son usados de manera uniforme por los calificadores. En la columna Tarea se muestra el nivel de dificultad de los textos que debían escribir los examinados; se observa que las dos tareas apenas difieren en dificultad. En la columna Atributo aparecen los valores de dificultad relativa de los atributos en los que se han calificado las tareas. Se ha de notar que, aunque las diferencias en dificultad son pequeñas, la variable corrección es la más difícil y la variable adecuación la más fácil. Finalmente, en la columna Categoría se muestran, mediante líneas, la localización en logits de los umbrales entre las

categorías utilizadas para puntuar las respuestas de los candidatos (de 0 hasta 3). En este caso, se utilizó la formulación del modelo expresada en (1). Es decir, se asume que las categorías son utilizadas de manera similar por todos los calificadores.

En adelante, se comenta de forma más detallada las propiedades psicométricas de los valores de las facetas más relevantes: los examinados y los calificadores.

Examinados

En la parte superior de la Tabla 1 se muestran, como ilustración de la salida proporcionada por el programa FACETS, los resultados de tres examinados cuyo número de identificación aparece en la primera columna (ID). En la columna 2 se muestran sus puntuaciones logit en la variable medida. Se observa que el sujeto 1 presenta una alta competencia (3.24). Por el contrario, el nivel en la variable del sujeto 82 es muy bajo (-2.25 logits). La suma de las 20 calificaciones (*r* = 2 calificadores x 2 tareas x 5 atributos) recibidas por cada examinado aparecen en la columna *X* y su media en la columna M_O (Promedio de las calificaciones observadas). Los valores *X* y M_O dependen del grado de severidad de los calificadores que han puntuado a cada examinado. Por ello, es conveniente utilizar como indicador de su competencia la puntuación en logit o el *Promedio imparcial* (M_I), que es la media de las calificaciones que recibiría el examinado de un calificador con un nivel medio de severidad. En

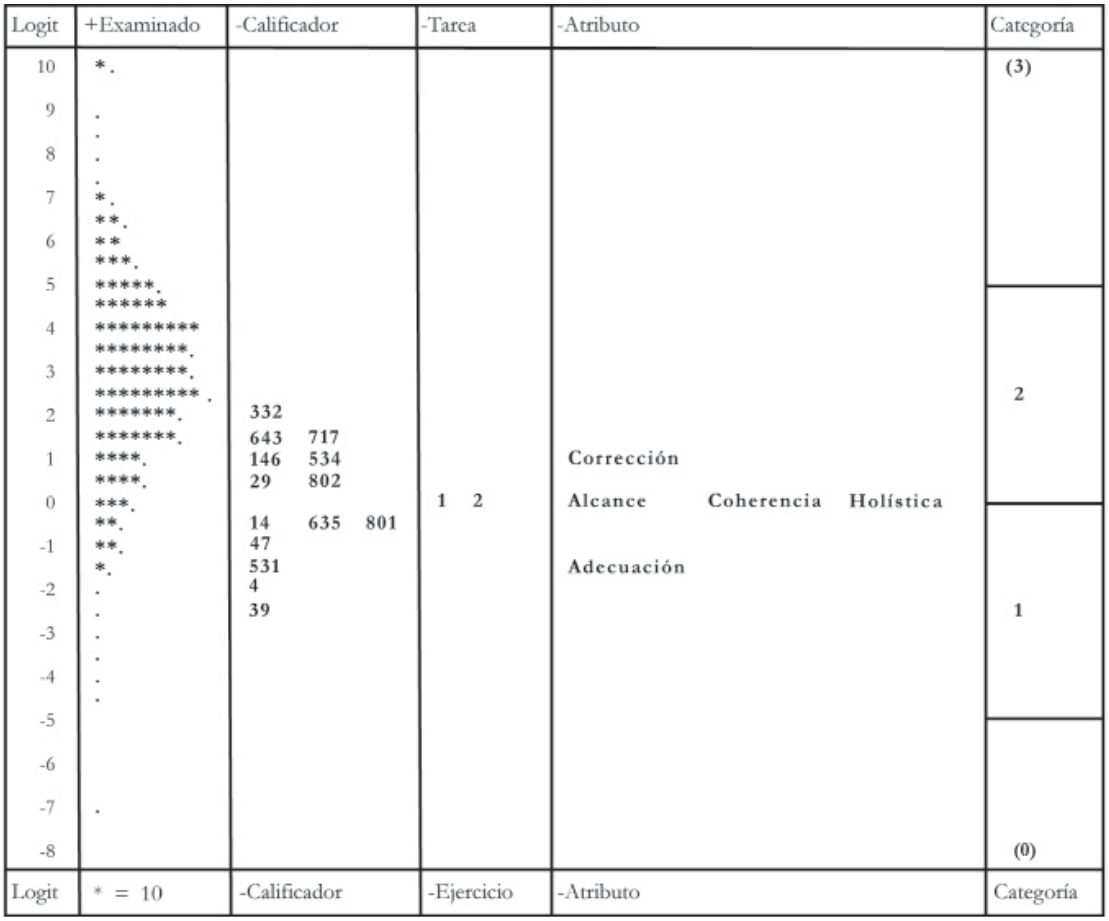

Figura 1. Mapa de la variable derivado del análisis con MFRM. Cada estrella representa a 10 examinados y cada punto a menos de 10. Los calificadores están representados por su número de identificación.

Tabla 1
Puntuaciones y estadísticos descriptivos de los examinados

ID	*Medida*	*SE*	*X*	*r*	M_0	M_I	*Infit*	*Outit*
1	3.24	.49	47.0	20.0	2.35	2.11	.89	.89
10	2.45	.48	34.0	20.0	1.70	1.98	.95	1.11
82	-2.25	.51	25.0	20.0	1.25	1.03	1.34	1.39
Media	2.82	.54	40.6	19.9	2.04	2.04	.99	1.03
DE	3.38	.19	9.60	.7	.47	.45	.58	.85
Max	11.26	1.85	60.0	20.0	3.00	3.00	7.48	9.00
Min	-7.19	.62	6.00	12.0	.50	.11	.09	.11

Nota. Medida: Valor en logits; SE: Errór estándar; X: suma de las calificaciones; r: número de calificaciones; M_O: Promedio observado; M_I: Promedio imparcial; Infit y Outfit: estadísticos de ajuste.

Tabla 2
Puntuaciones y estadísticos descriptivos de los calificadores

ID	*Medida*	*SE*	*X*	*r*	M_O	M_I	*Infit*	*Outfit*	*Rc,rc*	*% Acuerdo*
332	1.96	.06	2240	1360	1.65	1.66	.94	.94	.71	48.4
647	1.57	.06	2457	1396	1.76	1.75	.88	.85	.72	50.9
717	1.37	.06	2431	1360	1.79	1.80	1.22	1.30	.46	48.1
146	1.13	.06	2535	1371	1.85	1.84	.64	.58	.79	56.1
534	.99	.06	2591	1359	1.91	1.87	.90	.87	.78	55.4
29	.56	.06	2603	1345	1.94	1.94	.95	.93	.76	57.0
802	.45	.06	2611	1347	1.94	1.95	.81	.79	.70	55.8
635	-.26	.06	2888	1360	2.12	2.06	.89	.88	.66	55.1
14	-.40	.06	2863	1350	2.12	2.08	.96	.94	.79	57.8
801	-.53	.06	2670	1279	2.09	2.11	1.27	1.33	.80	51.4
47	-1.11	.06	2914	1302	2.24	2.21	1.18	1.21	.73	47.5
531	-1.63	.06	3165	1340	2.36	2.33	.88	.90	.75	52.3
4	-1.81	.06	3159	1340	2.36	2.37	1.39	1.47	.72	47.2
39	-2.29	.06	3400	1380	2.46	2.49	1.01	1.56	.70	43.8
Media	.00	.06	2751.9	1349.2	2.04	2.03	.99	1.04	.72	51.9
DE	1.36	.00	326.7	29.6	.25	.25	.20	.28	.09	4.40

Nota. Medida= Valor en logits; SE= Errór estándar; X= suma de las calificaciones; r = número de calificaciones otorgadas; M_O= Promedio observado de las calificaciones; M_I= Promedio imparcial de las calificaciones; Infit y Outfit: estadísticos de ajuste; $R_{c,rc}$: estadístico de consistencia entre calificadores; % Acuerdo: estadístico de consenso entre calificadores.

las dos últimas columnas de la derecha aparecen los estadísticos de ajuste: se observa que los examinados 1, 10 y 82 presentan valores de un ajuste aceptable (indicativo de que su desempeño ha sido interpretado por los calificadores consistentemente).

En la parte inferior de la Tabla 1 aparecen los principales estadísticos de las puntuaciones de los candidatos (N = 948) que realizaron el examen de expresión escrita. Se aprecia un rendimiento medio muy superior (2.82 logits) a la dificultad media de las tareas (situada en 0). Asimismo, se observa una alta variabilidad (DE = 2.38 logits) entre los examinados que es significativa estadísticamente (χ^2= 16135.1, gl=947, p < .0001). Las medidas de los candidatos oscilaron entre 11.26 logits y -7.19 logits. La fiabilidad de las puntuaciones es muy elevada (PSR=.95), lo cual indica que las puntuaciones en el examen permiten diferenciar fiablemente entre los diferentes niveles de competencia de los examinados. El ajuste al modelo de las calificaciones otorgadas a los examinados es aceptable, dado que las medias de Infit y Outfit apenas difieren de 1.0 y que el porcentaje de los candidatos que presentan un desajuste severo con las predicciones del modelo es bajo (7.28%).

Calificadores

En la Tabla 2 se muestran los promedios (en la escala de 0 a 3) de las evaluaciones de los calificadores: M_O y M_I, las puntuaciones en severidad (en la escala logit), su precisión (SE), los estadísticos de ajuste y los índices de fiabilidad entre calificadores (consenso y consistencia). Se observa que la variabilidad de los calificadores en severidad es elevada (DE = 1.36 logits) y significativa estadísticamente (χ^2= 6062.8, gl=13, p < .0001), este dato no es el deseable. Idealmente las variaciones en severidad habrían de ser bajas y atribuibles al error de medida, por lo que RSR (Rater separation reliability) debería ser bajo. Sin embargo, el índice RSR observado alcanza el valor máximo (1.00): un valor tan alto revela que las diferencias observadas en severidad entre los calificadores son muy fiables. De hecho, la precisión de las estimaciones de la severidad es alta: el error estándar de los estimadores del parámetro de severidad es muy bajo (.06) en todos los casos. Los extremos en la escala de severidad están ocupados por el calificador ID 332 (el más severo: 1.96 logits) y el calificador ID 39 (el más benigno: -2.29 logits).

En la escala de puntuaciones brutas en la que se expresa el *Promedio imparcial* (M_I), indicador del grado de severidad en la escala de 0 hasta 3, se observa igualmente una gran variabilidad que oscila entre 2.49 para el calificador más benigno (ID 39) y 1.65 para el calificador más severo (ID 332). Entre ambos calificadores la diferencia en puntuaciones directas es notable (.84 puntos). Se ha de notar que la relación entre las escalas en M_I y en logits es inversa: las mayores puntuaciones en la primera indican mayor benevolencia, mientras que en la segunda manifiestan mayor severidad.

Los valores de los estadísticos de ajuste indican un ajuste adecuado de los calificadores. Por un lado, las medias de Infit y Outfit difieren escasamente de la unidad con una variabilidad baja. Por otro, se observa que los valores se sitúan en un rango aceptable: Infit entre .64 y 1.39; Outfit entre .58 y 1.56. Estos datos indican que todos los calificadores muestran una adecuada consistencia interna (intra-calificador) en sus evaluaciones.

Los estadísticos relacionados con la fiabilidad entre calificadores aparecen en las dos últimas columnas de la Tabla 2. Las correlaciones de las evaluaciones de cada calificador con los demás calificadores son altas (media= .72 ; rango entre .46 y .80) indicando que hay una consistencia elevada entre los calificadores (la ordenación de los examinados en competencia es muy semejante entre ellos). Los índices de consenso de los calificadores (% Acuerdo: porcentaje de veces que cada calificador atribuye las mismas calificaciones que otros calificadores en idénticas circunstancias) oscilan entre 43.8% y 57.8%, con una media de 51.9%. En consecuencia, se ha de considerar que el grado de consistencia externa es mayor que el grado de consenso. Este dato se corrobora al analizar las propiedades métricas de las categorías de respuesta (rúbricas).

Categorías

En la Tabla 3 aparecen los estadísticos correspondientes a las categorías de respuesta derivadas de la formulación (1), la extensión de MFRM a partir del Modelo de Escalas de Calificación. Puede observarse que han sido empleadas todas las categorías numéricas y que la asignada más veces (55%) fue la 2. Además, las medidas promedio de cada categoría se incrementan monotónicamente desde -1.54 hasta 5.72 logits, el incremento de los promedios indica que cuanto mayor es la categoría, mayor es el nivel en la variable latente. Se observa asimismo que ninguna categoría desajusta severamente (Outfit < 2.0) y que los umbrales (pasos) entre las categorías sucesivas no están desordenados. Este dato implica que todas

Tabla 3
Estadísticos de las categorías

Categoría	*Frecuencia*	*Porcentaje*	*Medida promedio*	*Outfit*	*Paso*	*SE*
0	127	1	-1.54	1.7	--	--
1	3756	20	-.32	1.0	-5.10	.10
2	10247	55	2.55	1.1	.11	.02
3	4419	24	5.72	.90	4.99	.02

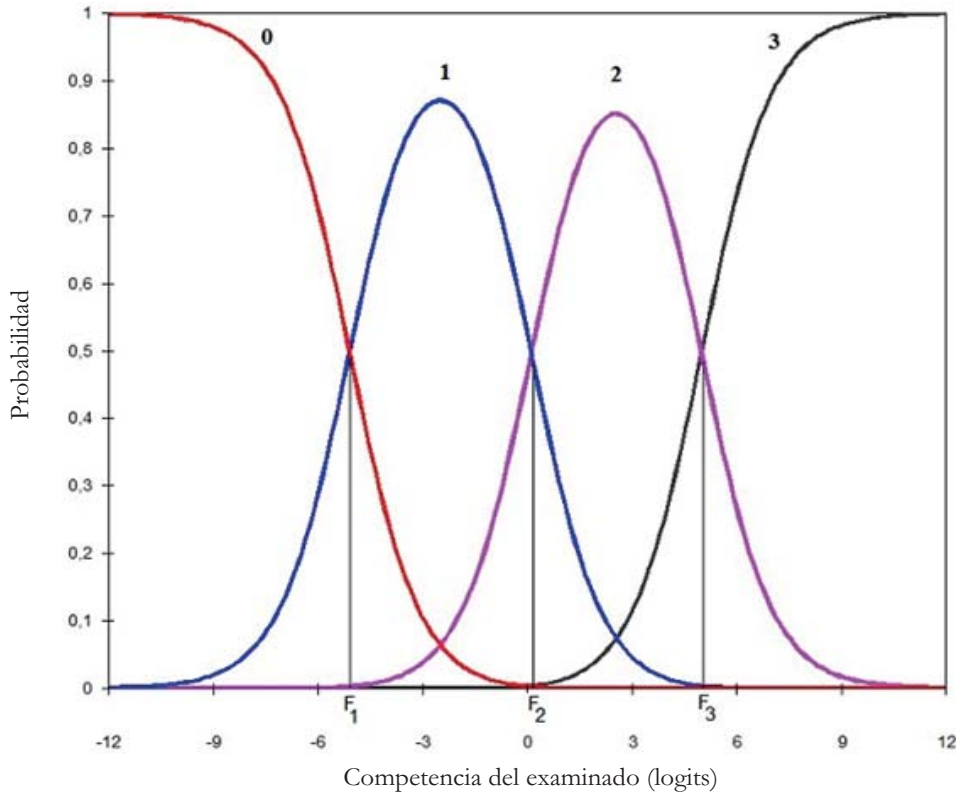

Figura 2. Curvas características de las categorías de respuesta

las categorías son modales (Figura 2): cada una es la de más probable elección en algún intervalo de la variable medida. Además, los incrementos entre los umbrales sucesivos son grandes y permiten distinguir adecuadamente rangos amplios de diferente magnitud en la variable latente.

En consecuencia, del análisis grupal de los calificadores, se puede concluir que las categorías numéricas presentan una funcionalidad óptima para obtener medidas en la variable latente (Linacre, 2004).

Si no se asume que los calificadores utilizan las categorías de manera uniforme, conviene analizar los datos con un modelo híbrido que es una extensión del Modelo de Crédito Parcial (Véase la ecuación (2)). Este enfoque aporta evidencias sobre el uso de las categorías por cada calificador. En la Tabla 4 se muestra el uso de cada categoría por cada calificador y los valores de los umbrales entre las categorías adyacentes de cada uno de ellos. Puede observarse, por ejemplo, que el calificador más severo (ID 332) sólo ha asignado la calificación 3 un 4% de las veces, mientras que el calificador más benigno (ID 39) la asignó el 52% de las veces. En la Tabla 3 se mostró que dichos calificadores presentaban un adecuado ajuste al modelo (alta consistencia interna). Además ambos tenían una alta consistencia con otros calificadores ($R_{c,rc} \geq .70$), lo cual indica que sus calificaciones ordenan a los examinados de forma similar. Sin embargo, los estadísticos de consenso con otros calificadores son moderados: el porcentaje de acuerdo de los dos calificadores que ocupan los extremos del continuo de severidad/benignidad es menor del 50%.

Tareas y Atributos

Ya se observaba en el mapa de Wright (Figura 1) que las dos tareas que debían desarrollar los examinados diferían escasamente en dificultad (tarea 1 = -.18 logits; tarea 2 = .18 logits). Sin embargo, aparecen mayores diferencias en dificultad entre los atributos evaluados, siendo la característica más difícil la Corrección gramatical del texto (1.18 logits) y la más fácil la Adecuación de la redacción al contexto (-1.46). Dado el gran número de calificaciones emitidas, los

Tabla 4.
Calificadores. Porcentaje de asignación de las categorías y Umbrales entre las categorías adyacentes

ID	*Severidad*	*C0*	*C1*	*C2*	*C3*	*F1*	*F2*	*F3*
332	**1.96**	1	40	55	**4**	-6.04	.07	**5.97**
647	1.57	1	32	59	8	-5.91	.23	5.69
717	1.37	--	27	67	5	--	-3.07	3.07
146	1.13	0	25	69	6	-7.74	.41	7.33
534	.99	0	27	55	18	-5.50	.49	5.01
29	.56	0	25	57	17	-5.25	.28	4.97
802	.45	1	16	72	11	-5.01	-1.06	6.07
635	-.26	--	10	68	22	--	-2.77	2.77
14	-.40	0	20	50	29	-5.13	.54	4.58
801	-.53	3	22	41	35	-3.49	.16	3.33
47	-1.11	1	14	48	37	-4.85	.34	4.52
531	-1.63	0	8	49	43	-5.14	.10	5.05
4	-1.81	2	11	38	49	-2.65	-.40	3.04
39	**-2.29**	1	5	43	**52**	-3.56	-.60	**4.16**

Nota. C0 – C1: Porcentaje de veces que se ha asignado cada categoría numérica; F1 – F3: Umbrales en logits entre las categorías adyacentes.

estimadores de la dificultad de los elementos de ambas facetas son muy precisos (Item Separation Reliability, ISR): ISR_{Tarea} = .99 e $ISR_{Atributo}$ = 1.00, y su variabilidad es significativa estadísticamente: Tarea (χ^2 = 115.5, gl = 1, p < .0001) y Atributo (χ^2 = 2358.2, gl = 3, p < .0001).

Discusión

Como se indicó al inicio, los tests de desempeño (performance assessment) son procedimientos de evaluación en los que los examinados realizan tareas de formato abierto que han de ser cuantificadas por un calificador mediante una escala de categorías numéricas. La influencia del criterio del calificador en la puntuación otorgada es determinante. Es decir, la magnitud de las calificaciones que reciben los examinados no dependen sólo de su nivel de competencia, sino que hay que considerar los efectos del calificador en la calificaciones: su grado de severidad o benignidad, su uso de las rubricas y otros efectos idiosincráticos como el de *halo* y el de *tendencia central* (Myford & Wolfe, 2004b). Los efectos del calificador han de ser considerados una amenaza para la validez de las calificaciones de los examinados (Lane & Stone, 2006). Además, en los programas de evaluación a gran escala se suele evaluar el comportamiento de los calificadores con el fin de detectar a los más eficaces y medir el efecto del entrenamiento destinado a mejorar sus prácticas.

Con los procedimientos tradicionales de medición, basados en la Teoría Clásica de los Tests, no es posible discernir si la diferencia en las calificaciones recibidas por dos examinados se debe a que uno de ellos es más competente que el otro o a que el calificador que evaluó al primero es más benigno que el que evaluó al segundo. De forma similar, si el promedio de las calificaciones otorgadas por un evaluador es elevado, no es posible determinar si la magnitud de las calificaciones se debe a que el calificador es muy benigno, o si la muestra de personas que ha puntuado tiene una alta competencia. Para desenredar esta madeja conviene utilizar modelos psicométricos que permitan obtener la separabilidad de los parámetros de las personas y los calificadores (Tesio et al., 2015). Tal es el caso del modelo de Rasch en el que el teorema de la *objetividad específica* demuestra que es posible obtener medidas invariantes (Engelhard, 2013).

Desde esta perspectiva, se ha descrito el modelo MFRM, una extensión del modelo dicotómico de Rasch, que es apto para medir en la misma métrica los elementos de las distintas facetas que pueden influir en la variabilidad de las calificaciones: los examinados, los calificadores, las tareas, los atributos evaluados, etc.

Como un ejemplo del uso del modelo y de la interpretación de sus estadísticos, se ha realizado el análisis de una prueba de Expresión escrita integrada en un examen para la obtención del Diploma de Español como Lengua Extranjera (DELE) de Nivel B1. El examen fue realizado por 948 personas con lenguas maternas muy diversas. En la prueba los examinados escribieron dos textos que fueron evaluados independientemente por dos calificadores, los cuáles puntuaron los textos en cinco características o atributos. Participaron en total 14 calificadores a los que se asignaron las pruebas mediante un procedimiento simple de rotación de los examinados y los calificadores para garantizar la conectividad. El programa Facets no detectó ningún subconjunto de datos desconectado, por lo que fue posible localizar los elementos de todas las facetas en la misma métrica.

El objetivo prioritario de la evaluación era obtener medidas que representasen de forma válida la competencia en expresión escrita de los examinados y su localización en el mapa de la variable.

Tanto en logits como en la escala de puntuaciones brutas (Promedio Imparcial), se observó que rendimiento medio de los examinados es superior a la dificultad media de las tareas y que su variabilidad es alta. Al estar en la misma escala de las categorías usadas en la evaluación, el uso del estadístico Promedio Imparcial es muy recomendable para informar a los usuarios poco familiarizados con la escala logit.

La fiabilidad de las puntuaciones de las personas examinadas fue muy elevada, indicando que las calificaciones otorgadas por los evaluadores diferenciaban fiablemente entre los distintos niveles

de competencia de los examinados. El ajuste al modelo de las calificaciones otorgadas a los examinados era aceptable, dado que las medias de Infit y Outfit apenas difieren de 1.0 y que fue bajo el porcentaje de los candidatos que presentaban un desajuste severo (Outfit y/o Infit > 2.0).

Otro objetivo importante del análisis mediante MFRM fue obtener medidas que representasen de forma válida el grado de severidad/benignidad de los calificadores y su localización en el mapa de la variable. Se observó que las puntuaciones logits en el constructo eran muy fiables y que su variabilidad era elevada y estadísticamente significativa. La variabilidad en severidad constituye una fuente de error que decrementa la validez de las puntuaciones de los examinados. En la escala de puntuaciones brutas en la que se expresa el Promedio imparcial (M_I), se observó igualmente una gran variabilidad.

Un gran número de factores puede contribuir a que el estilo habitual de un calificador sea más severo o más benevolente: la experiencia profesional, los rasgos de personalidad, la carga de trabajo, las consecuencias de la evaluación, etc. En muchas ocasiones, por ejemplo, los calificadores más experimentados suelen ser más severos que los más novatos. Como indica Eckes (2011), no existe aún suficiente investigación acerca de los determinantes personales y situacionales de la severidad de los calificadores. Tampoco abundan los estudios sobre la estabilidad y el cambio en el estilo de los calificadores al puntuar.

En varios programas de evaluación a gran escala se han implementado programas de entrenamiento para eliminar o reducir la variabilidad entre los calificadores en los modos de evaluar. Los programas suelen basarse en la instrucción mediante calificadores expertos acerca de las tareas y los procedimientos de calificación. Se trata de generar en los calificadores un conocimiento compartido sobre aspectos como el constructo que se desea medir, la clasificación en los niveles de desempeño, los descriptores característicos de cada nivel, y el significado de las categorías numéricas o rúbricas (Eckes, 2011).

Pese a los esfuerzos invertidos en el entrenamiento, existen evidencias de que las diferencias entre los calificadores, aunque se reduzcan, perviven tras los programas de intervención (Congdom & McQeen, 2000). Por ello, es imprescindible utilizar un modelo, como el MFRM, que permita medir el desempeño de los examinados de manera objetiva separando la influencia de la severidad de los calificadores (McNamara, 2000).

Un elemento muy importante de evaluación mediada por el calificador es el análisis de la fiabilidad del calificador, que presenta dos vertientes: la consistencia interna (intra-calificador) y la fiabilidad entre calificadores. Los estadísticos de ajuste de los calificadores suelen ser interpretados como el grado de *consistencia interna* del calificador (Eckes, 2011; Engelhard, 2013): indican la consistencia al interpretar las rúbricas y los atributos en los distintos examinados. En este estudio se observó que los valores se situaban en un rango aceptable (Infit entre .64 y 1.39; Outfit entre .58 y 1.56).

Anteriormente se indicó que se usan dos tipos de indicadores para analizar la fiabilidad entre los calificadores: índices de consenso e índices de consistencia (Eckes, 2011). Los índices de consenso reflejan el porcentaje de veces que un calificador atribuye las mismas calificaciones que otros calificadores en idénticas circunstancias (examinado, tarea, atributo, etc.). Los índices de consistencia indican el grado de asociación entre las calificaciones otorgadas por distintos calificadores. Las correlaciones de las evaluaciones de cada calificador con los demás calificadores fueron altas, indicando que hay una consistencia elevada entre los calificadores. Sin embargo, los índices de consenso de los calificadores (% Acuerdo) oscilaron entre 43.8% y 57.8%. En consecuencia, se ha de considerar que, como en otros estudios, el grado de consistencia externa es mayor que el grado de consenso (Eckes, 2011).

Además de las diferencias en severidad, existen otros modos de calificar que, al introducir una variabilidad en las calificaciones que no está asociada al constructo de interés, han sido clasificados tradicionalmente como errores del calificador: el

efecto de tendencia central y el efecto de halo. El primero, un caso particular de la restricción del rango, se produce cuando sistemáticamente un calificador evita usar las categorías extremas. El segundo ocurre cuando el calificador no es capaz de diferenciar entre atributos conceptualmente diferentes del desempeño y tiende a asignar calificaciones similares a todos ellos (el término de halo se aplicó bajo la suposición de que la impresión producida en el calificador por uno de los atributos u otra característica del examinado influye en las demás calificaciones).

Suelen considerarse evidencias del efecto de tendencia central un bajo índice de fiabilidad (Rater Separation Reliability) o valores de Infit y Outfit sobreajustados (Myford & Wolfe, 2004). Asimismo, es una evidencia de este efecto la baja frecuencia de calificaciones en las categorías extremas. En nuestro estudio, el análisis del grupo de calificadores no presentó fuertes indicios de un efecto de tendencia central: los promedios de Infit y Outfit apenas difieren de la unidad, RSR = 1.0 y el 44% de las evaluaciones se distribuyen simétricamente en las categorías 1 y 3.

Por su parte, cuando la mayoría de los calificadores están influidos por el efecto de halo, las medidas de los atributos evaluados apenas difieren entre sí (el índice de fiabilidad de las medidas de los atributos es bajo). En nuestro caso, la fiabilidad alcanza el máximo nivel y el rango de las medidas de los atributos es elevado.

Además de estos efectos, los calificadores pueden diferir en el uso y la interpretación de las rúbricas. El análisis del grupo de calificadores revela que las rúbricas han sido utilizadas de forma adecuada de acuerdo con los criterios de Linacre (2004). Se ha de resaltar que los umbrales (pasos) entre las categorías sucesivas no están desordenados y que los incrementos entre los umbrales sucesivos son grandes y permiten distinguir adecuadamente rangos amplios de diferente magnitud en la variable latente.

No obstante, el análisis del comportamiento individual de los calificadores con un modelo híbrido de MFRM permite describir de manera más pormenorizada el uso de las categorías por parte de los calificadores. En nuestro caso, pudimos observar el modo idiosincrático de puntuar de los calificadores situados en los extremos de la escala de severidad.

En este trabajo, se ha presentado una introducción al MFRM y a su empleo para analizar los tests de desempeño. Por razones de extensión, otras aplicaciones más especializadas no han sido descritas en este artículo, tales como el análisis del funcionamiento diferencial de los calificadores y los procedimientos para analizar su exactitud (accuracy). Los índices de exactitud de los calificadores muestran las discrepancias o distancias entre las calificaciones observadas y un patrón o punto de referencia (benchmark) que comúnmente se define mediante el promedio de las calificaciones otorgadas por un grupo de calificadores muy expertos (Murphy y Cleveland, 1995). Pueden consultarse excelentes exposiciones de estos temas en Eckes (2011), Englehard (2013) y Myford y Wolfe (2004 a y b).

Referencias

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43,* 561-573.

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

Eckes, T. (2011). *Introduction to Many-Facet Rasch Measurement.* Franfurt am Main: Peter Lang.

Engelhard, G. (2002). Monitoring raters in performance assessment. En G. Tindall & T. Haladyna (Eds.), *Large-scale assessment programs for all students: Development, implementation, and analysis.* (pp. 261-287). Mahwah, *NJ:* Erlbaum.

Engelhard, G. (2013). *Invariant Measurement. Using Rasch Models in the Social, Behavioral, and Health Sciences.* New York and London: Routledge.

Gyagenda, I. S., & Engelhard, G. (2010). Using Classical and Modern Measurement Theories to Explore Rater, Domain, and Gender Influences on Student Writing Ability. En M. L. Garner, G. Engelhard,

W. P. Fisher & M. Wilson (Eds). *Advances in Rasch Measurement Volume I* (398-429). Maple Grove, MN: JAM Press.

Hambleton, R. K. (2000). Advances in performance assessment methodology. *Applied Psychological Measurement, 24*, 291-293.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*, 3-31. doi: 10.1191/0265532202lt218oa

Lane, S., & Stone, C.A. (2006). Performance Assessment. En R. L. Brennan (Ed.): *Educational Measurement* (pp 387-431). Wesport, CT: ACE/Praeger.

Linacre, J. M. (1989). *Many-facet Rasch measurement.* Chicago: MESA Press.

Linacre, J. M. & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement, 3*, 484-509.

Linacre, J. M. (2004). Optimizing rating scale category effectiveness. En E. V. Smith & R. M. Smith (Eds.) *Introduction to Rasch Measurement.*(pp. 48-72). Maple Grove, MN: JAM Press.

Linacre, J. M. (2010). *A user's guide to Facets: Rasch model computer programs.* Chicago: Winsteps.com.

Linacre, J. M. (2015). *Facet Rasch Measurement computer program* (Version 3.71.3) (Computer program). Chicago: Winsteps.com.

Martínez-Arias, R. (2010). La evaluación del desempeño. *Papeles del Psicólogo*, *31*, 85-96.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

McNamara, T. F. (2000). *Language testing*. Oxford, UK: Oxford University Press.

Murphy, K. R. & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives.* Thousand Oaks, CA: Sage.

Myford, C. M. & Wolfe, E. W. (2004a) Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part I. En E. V. Smith & R. M. Smith (Eds.) *Introduction to Rasch Measurement* (pp. 460-517). Maple Grove, MN: JAM Press.

Myford, C. M. & Wolfe, E. W. (2004b) Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part II. En E. V. Smith & R. M. Smith (Eds.) *Introduction to Rasch Measurement* (pp. 518-574). Maple Grove, MN: JAM Press.

Park, T. (2004). An Investigation of an ESL Placement Test of Writing Using Many- facet Rasch Measurement, *Papers in TESOL* & *Applied Linguistics, 4,* 1-21.

Prieto, G. (2011). Evaluación de la ejecución mediante el modelo Many-Facet Rasch Measurement. *Psicothema*, *23,* 233-238.

Prieto, G. & Delgado, A. (2003). Análisis de un test mediante el modelo de Rasch. *Psicothema*, *15,* 94-100.

Prieto, G. & Nieto, E. (2014). Analysis of rater severity on written expression exam using Many Faceted Rasch Measurement. *Psicológica*, *35,* 285-397.

Rasch, G. (1960). *Probabilistics models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Rasch, G. (1977). On specific objectivity: An attempt at formalazing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy, 14,* 58-94.

Smith R. M., Shumacker R. E. & Bush M. J. (1998). Using item means squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, *2,* 66-78.

Tesio, L., Simone, A., Grzeda, M. T., Ponzio, M., Dati, G., Zaratin, P., Perucca, L. & Battaglia, M. A. (2015). Funding Medical Research Projects: Taking into Account Referees' Severity and Consistency through Many-Faceted Rasch Modeling of Projects' Scores. *Journal of Applied Measurement, 16,* 129-152.

Tyndall, B. & Kenyon, D. M. (1996) Validation of a new holistic rating scale using Rasch multi- faceted analysis. En A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 39-57). Clevedon: Multilingual Matters.

Wolfe, E.W. (2009). Item and Rater Analysis of Constructed Response Items via the Multi-Faceted Rasch Model. *Journal of Applied Measurement, 10,* 335-347.

Wright, B. D. & Stone, M. H. (1979). *Best test design.* Chicago: MESA Press.

Recibido: 25 de julio de 2015
Aceptado: 18 de setiembre de 2015