https://doi.org/10.21865/RIDEP76.2.02

Uso del Modelo de Rasch de Facetas Múltiples para Analizar Test con Ítems de Respuesta Abierta

Using the Many Facet Rasch Model to Analyse Tests with Open-ended Items

Gerardo Prieto[1]

Resumen

Las respuestas a los test de respuesta abierta han de ser puntuadas por calificadores cuyo correcto proceder es la clave para obtener mediciones fiables y válidas. El comportamiento de los calificadores ha de ser evaluado si se desea medir de forma fiable y válida el constructo de interés. Los estadísticos clásicos de consenso y consistencia entre calificadores no son apropiados porque arrojan resultados contradictorios en función de su grado de severidad. Además, con los procedimientos de puntuación basados en la suma de los valores otorgados a los ítems por varios calificadores no es posible esclarecer si la magnitud de las puntuaciones recibidas se debe al nivel de competencia de los examinados o a los efectos del calificador. Los modelos psicométricos de tipo Rasch permiten obtener la separabilidad de los parámetros de las personas y los calificadores. En este artículo se muestra la utilidad de un modelo de Rasch (Many-Facet Rasch Measurement, MFRM) para obtener medidas invariantes del rendimiento de los examinados, de la severidad de los calificadores, de la dificultad de las tareas y de otras facetas adicionales de las pruebas de respuesta construida. Se ilustra la formulación del modelo y sus estadísticos básicos con un ejemplo 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.

Palabras clave: test de respuesta abierta, evaluación mediante calificadores, efectos del calificador, Modelo de Rasch de Facetas Múltiples

Abstract

Responses to open-ended tests must be scored by raters whose correct behavior is the key to obtaining reliable and valid measurements. Rater behavior must be evaluated if the construct of interest is to be measured reliably and validly. The classical inter-rater consensus and consistency statistics are not appropriate because they yield contradictory results depending on their degree of severity. Moreover, with scoring procedures based on the sum of the values given to the items by several raters, it is not possible to clarify whether the magnitude of the scores received is due to the level of competence of the examinees or to the effects of the rater. Rasch-type psychometric models make it possible to obtain the separability of person and rater parameters. This paper shows the usefulness of a Rasch model (Many-Facet Rasch Measurement, MFRM) for obtaining invariant measures of examinee performance, rater severity, task difficulty, and additional facets of constructed-response tests.The formulation of the model and its basic statistics are illustrated with an example in which the sources of variability in student scores on a written expression test are analyzed using the FACETS program.

Keywords: open-ended test, rater mediated assessment, rater effects, Many-Facet Rasch Measurement

[1]Doctor en Psicología, Profesor Emérito Honorífico. Facultad de Psicología, Universidad de Salamanca, Avenida de la Merced, 109-131, 37005 Salamanca (España).Tel.:+34923294610. Correo: gprieto @usal.es

ISSN: 1135-3848 print /2183-6051online

Introducción

Los test de elección múltiple son muy utilizados en la evaluación psicológica y educativa por su menor costo: requieren un tiempo breve para emitir las respuestas y facilitan la corrección objetiva y automática. Sin embargo, ese formato presenta varias limitaciones. Por un lado, permite la adivinación para elegir la respuesta correcta y el uso de estrategias de repuesta no deseadas (por ejemplo, la eliminación de opciones antes de emitir la respuesta). Por ello, los ítems de elección múltiple presentan menor dificultad, discriminación y fiabilidad que los ítems abiertos (Wu et al.,2016). Por otro lado, su validez ecológica es deficiente: en la vida real no se ha de elegir la más plausible de una lista de opciones para resolver un problema. Por ello, se suele considerar que el formato de elección múltiple no es el más apto para evaluar las competencias de las personas en contextos realistas y las habilidades de nivel cognitivo superior (Bravo & Fernández, 2000).

En los ítems de respuesta abierta el sujeto produce libremente la respuesta a un interrogante o un problema. Los formatos varían en su longitud. Pueden solicitarse respuestas cortas (completar de frases, por ejemplo) o producciones más largas denominadas ensayos, ejecuciones o desempeños. En estos procedimientos la corrección y puntuación de las respuestas ha de ser llevada a cabo por calificadores. Estos agentes, que son la clave del proceso para lograr mediciones válidas y fiables, cobran extraordinaria importancia en los contextos de evaluación de la producción. Por ejemplo, los profesores de los conservatorios evalúan la competencia de los alumnos puntuando su ejecución de obras musicales, los jueces de los campeonatos de gimnasia asignan calificaciones a los deportistas en función de la calidad de los ejercicios realizados, los evaluadores de los exámenes académicos califican las redacciones de los candidatos, etc. Todos estos casos son ejemplos de la metodología denominada evaluación de la ejecución (Hambleton, 2000) o evaluación del desempeño (Martinez-Arias, 2010; Prieto, 2011).

La importancia de los calificadores

La magnitud de las calificaciones que reciben los evaluados no dependen solo de su nivel de competencia, sino que se han considerar otras facetas que influyen en ella notablemente como el criterio del calificador, su severidad, su consistencia y su uso de las rubricas o categorías de calificación. 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 & Wind, 2018). Por tanto, el comportamiento de los calificadores ha de ser evaluado si se desea medir de forma fiable y válida el constructo de interés (Lane & Stone, 2006).

El enfoque estándar para abordar la evaluación mediante calificadores consta de tres aspectos: la formación de los calificadores, el uso de calificaciones independientes del mismo rendimiento por parte de dos o más evaluadores y el análisis de la fiabilidad entre calificadores (Eckes, 2015). El objetivo de la formación de los evaluadores suele ser familiarizarlos con el constructo a medir, el formato de la prueba, las tareas y los criterios de calificación. Con la formación o entrenamiento de los evaluadores se persigue incrementar la convergencia de sus evaluaciones de manera que la denominada fiabilidad entre evaluadores sea lo más alta posible. Sin embargo, los resultados empíricos muestran de manera recurrente que ese objetivo es difícil de alcanzar en la mayoría de las situaciones: la variabilidad entre los calificadores suele ser persistente (Eckes, 2009).

Por ello, se suele cuantificar el grado de fiabilidad de los calificadores con el fin de determinar si las calificaciones otorgadas están poco contaminadas por errores de medida y son científicamente legítimas.

Los procedimientos para evaluar la fiabilidad se basa en las calificaciones independientes emitidas por varios calificadores. En los diseños de recogida de datos es muy importante planificar la independencia de las calificaciones, asegurando que cada calificador no conoce las evaluaciones del resto de loa calificadores. Por tanto, para analizar la fiabilidad se han de evitar las calificaciones emitidas por consenso.

El enfoque clásico para analizar la fiabilidad de los calificadores distingue entre dos grandes clases de indicadores estadísticos: índices de consenso e índices de consistencia (Eckes, 2015). Los índices de consenso reflejan el grado en que evaluadores independientes proporcionan la misma calificación a una producción (correspondencia absoluta de las calificaciones). Los estadísticos más empleados en este enfoque son el índice de acuerdo exacto y el coeficiente kappa de Cohen.

Los índices de consistencia indican el grado en el que evaluadores independientes proporcionan la misma ordenación de las personas u objetos evaluados. Los estadísticos mas empleados son la correlación de Pearson, la correlación ordinal de Spearman, el coeficiente tau-b de Kendall.

Como indica Eckes (2009), estos dos tipos de indicadores pueden arrojar resultados y conclusiones discrepantes. Es posible observar un bajo consenso entre evaluadores y, al mismo tiempo, una alta consistencia entre ellos. Por ejemplo, un evaluador muy estricto puede asignar puntuaciones a los examinados que sean sistemáticamente uno o dos puntos más bajas que las puntuaciones que otro evaluador asigna a los mismos examinados. En consecuencia, la ordenación relativa de los examinados será equivalente en ambos evaluadores (con índices muy altos de consistencia), siendo por el contrario muy bajos los índices de consenso porque las calificaciones asignadas no han sido exactamente las mismas. Sin duda, las diferencias en el grado de severidad/benignidad de los calificadores pueden producir esas discrepancias.

Lamentablemente el procedimiento clásico de puntuación (suma de las calificaciones otorgadas por un calificador a los examinados) no es una estimación adecuada del grado de severidad de un calificador si no ha puntuado a todos los examinados, como suele ocurrir en la mayoría de las aplicaciones. Una puntuación alta podría deberse al alto grado de benignidad del calificador y/o a que el subconjunto de sus evaluados son de elevado nivel en el constructo. Un marco de medida más adecuado consiste en emplear un modelo de medida que permita separar el nivel de los examinados de la severidad de los calificadores. Tal es el caso del denominado Modelo de Rasch de Facetas Múltiples (Linacre, 1989) el cual, como el resto de las variantes del modelo de Rasch, posee la propiedad de *objetividad específica* (Rasch, 1977) permitiendo, si los datos se justan al modelo, mediciones invariantes de los elementos de las facetas incluidas en el marco de medición (personas, ítems, calificadores, etc).

El Modelo de Rasch de Facetas Múltiples (MRFM)

MFRM extiende algunos de los modelos de Rasch más usuales a las evaluaciones en las que los calificadores puntúan la respuesta a un ítem abierto (Eckes, 2018; Engelhard & Wind, 2015; Myford & Wolfe, 2003 y 2004; Prieto, 2011 y 2015; Prieto & Nieto, 2014). Dependiendo del formato de respuesta empleado (dicotómico o politómico), se puede formular MFRM como una extensión del modelo dicotómico de Rasch (1960) o de modelos politómicos como el de Escalas de Calificación (Andrich, 1978) o el de Crédito Parcial (Masters, 1982). El objetivo general de estos modelos de Rasch es medir conjuntamente en una dimensión los elementos de las facetas incluidas en el marco de medición: las personas, los ítems y otras facetas como los evaluadores o calificadores.

La propiedad principal de esta familia de modelos es la invarianza de las medidas. De acuerdo con Engelhard (2013), la medición invariante de las personas, los ítems y los calificadores se logra si se cumplen los siguientes requisitos: (a) las medidas de las personas deben ser independientes de los ítems específicos utilizados para medir, (b) las funciones de respuesta de las personas asociadas a la dificultad de los ítems no se cruzan (una persona con mayor competencia debe de tener siempre una mayor probabilidad de éxito en cualquier ítem que una persona menos competente), (c) las medidas de los ítems han de ser independientes de las personas empleadas para la calibración, (d) las funciones de respuesta de los ítems asociadas al nivel de las personas no se cruzan (cualquier persona habrá de tener mayor probabilidad de éxito en un ítem fácil que en un ítem difícil), (e) la medición de las personas ha de ser independiente de los calificadores que han intervenido en la medición, (f) las curvas características de las personas asociadas a la severidad/benignidad de los calificadores no se cruzan (una persona con mayor competencia tendrá mayor probabilidad de obtener de cualquier evaluador mayores calificaciones que

una persona con menor competencia), (g) la calibración de los calificadores ha de ser independiente de las personas a las que han evaluado, (h) las curvas características de los calificadores no se cruzan (cualquier persona ha de tener mayor probabilidad de obtener una puntuación alta de los calificadores benévolos que de los calificadores más severos) e (i) los elementos de las facetas analizadas (personas, ítems, calificadores, etc) deben ser localizados conjuntamente en una única dimensión latente (mapa de la variable).

El cumplimiento de los requisitos enunciados anteriormente puede ser contrastado mediante índices de ajuste.

Para explicitar la formulación de MRFM y sus estadísticos básicos se describe en adelante un ejemplo de la aplicación del modelo para analizar las calificaciones en una prueba de expresión escrita en español (Prieto, 2015). Un total de 14 profesores calificaron dos textos escritos (tareas) producidos por casi un millar de estudiantes de español como segunda lengua. Los evaluadores puntuaron en cinco atributos la ejecución de los estudiantes con una escala de 0 a 3 (rúbricas o categorías numéricas): impresión holística, adecuación al contexto, coherencia, corrección gramatical y alcance. El modelo propuesto es una extensión del Modelo de Escalas de Calificación a una situación en la que existen cuatro facetas que contribuyen a la variabilidad de las medidas (examinados, calificadores, tareas y atributos evaluados). 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 reciba una puntuación o la inmediatamente inferior.

En concreto,

$$ln(Pnijlk /Pnijl\ (k\text{-}1) = Bn - Di - Rj - Cl - Fk \quad (1)$$

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*, D*i*=dificultad de la tarea *i*, R*j*=severidad del calificador j, C*l*=dificultad del atributo *l*, F*k*=valor en logit del punto en el que las categorías *k* y *k-1* son equiprobables.

En la ecuación 1, el logit *(ln (Pnijk /Pnij(k-1))* es la variable dependiente y las diversas facetas (personas, tareas, calificadores y atributos) 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 puede ser una extensión del Modelo de Crédito Parcial con el que se desea analizar las peculiaridades de los calificadores al usar las categorías numéricas:

$$ln(Pnijlk /Pnijl\ (k\text{-}1) = Bn - Di - Rj - Cl - Fjk \quad (2)$$

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 $0 \pm \infty$. 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 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 los valores observados y los predichos por el modelo.

Para los usuarios que no estén familiarizados

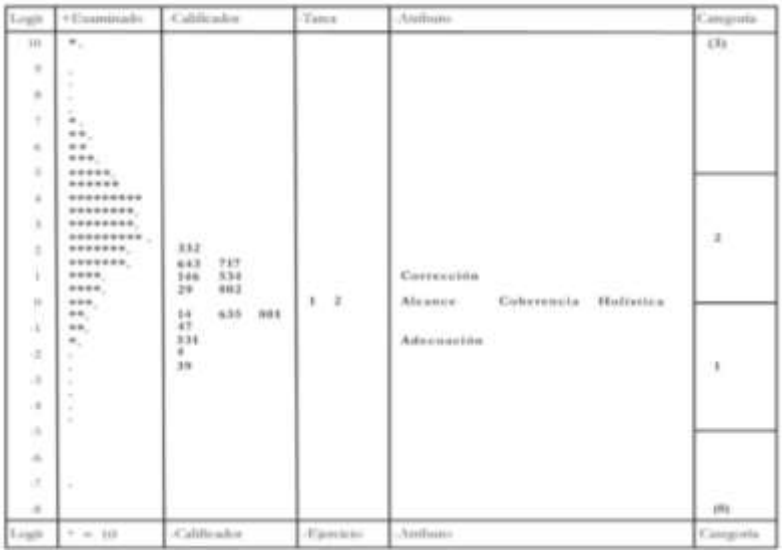

Figura 1. Mapa de Wright. Cada estrella representa a 10 examinados y cada punto a menos de 10. Los calificadores están representados por su número de identificación

con la escala logit, puede ser útil presentar los valores de las personas, los ítems y los calificadores en la misma escala de puntuaciones directas usada para otorgar las calificaciones (en el ejemplo, 0-3). Esta transformación se denomina *promedio justo* o imparcial (fair average). En el caso de las personas evaluadas, el promedio justo es la media de las puntuaciones que otorgaría a un evaluado un calificador con un nivel promedio de severidad (Eckes, 2011; Prieto & Nieto, 2014). De forma similar es posible obtener el promedio justo de cada calificador para puntuar su nivel de severidad. En este caso, es la media de las puntuaciones que otorgaría un calificador a un examinado con un nivel medio de competencia. Además de las puntuaciones para cada uno de los elementos de las distintas facetas, 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, 2003).

El programa FACETS es el software más empleado para llevar a cabo los análisis (Linacre, 2015).

Estadísticos básicos

Ajuste

El ajuste se cuantifica mediante los residuos (diferencias entre los valores observados y los predichos por el modelo). 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 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 .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). FACETS aporta valores individuales de ajuste para los evaluados, los calificadores, los ítems y las categorías de calificación.

Mapa de la variable

En la Figura 1 aparece el mapa de la variable (denominado mapa de Wright) que permite observar la distribución de las medidas de los elementos de todas las facetas en la variable latente.

En la primera columna del mapa aparece la escala logit en la que se puntúan los examinados, los calificadores, las tareas, los atributos y los umbrales entre las categorías numéricas. La magnitud de los valores logit se interpreta como rendimiento o competencia en los examinados, como severidad en los calificadores y como dificultad en las tareas y los atributos.

En la columna “Examinado” se representa la distribución de los examinados en la escala. Cada asterisco (*) representa a 10 personas y cada punto

a una frecuencia inferior. Los examinados 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. Su rendimiento es elevado (Media=2.82 logits) y aparece una gran variabilidad (entre 11.26 y -7.19 logits). En la columna Calificador aparecen las medidas en severidad/benignidad 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 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/benignidad. Una escasa variabilidad revelaría que los criterios de asignación de las puntuaciones son usados de manera uniforme por los calificadores. En la columna Tarea se muestra el nivel de dificultad de los textos que escribieron los examinados; se observa que las dos tareas apenas difieren en dificultad. En la columna Atributo aparecen los valores en dificultad de los atributos en los que se han calificado las tareas. Se ha de notar que, aunque las diferencias en dificultad son pequeñas, los atributos Corrección y Adecuación al contexto son el más difícil y el más fácil respectivamente. Finalmente, en la columna Categoría se muestra, mediante líneas, la localización en logits de los umbrales entre las categorías utilizadas para puntuar las respuestas de los examinados (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.

Fiabilidad de las medidas

Como en el resto de los modelos de tipo Rasch, la precisión de las medidas de cada uno de los elementos de cada faceta (persona, ítem, calificador) se expresa en el error estándar (SE) que es la desviación típica de la distribución muestral de los estimadores del parámetro: la precisión es mayor cuanto menor es SE. Además de evaluar la precisión individual de las medidas, el modelo permite obtener una cuantificación de la precisión o fiabilidad a nivel de grupo. El Índice de Fiabilidad de la Separación (SR) cuantifica la fiabilidad de las medidas de las distintas facetas (las personas, las tareas, los atributos o los calificadores) indicando cuál es la proporción de la varianza verdadera respecto de la varianza observada de las medidas. Otros índices grupales de precisión son el Índice de Separación (G), que es el cociente entre la desviación típica verdadera y el error estándar promedio y el Índice de Estratos (H=(4G +1)/3). El índice H indica el número de distintos niveles de la variable medida que es posible identificar fiablemente (separados por 3 errores estándar de medida). Las interpretaciones sustantivas de SR difieren entre las facetas (Myford & Wolfe, 2003). En las medidas de las personas, PSR (Person separation reliability) es comparable al coeficiente alfa de Cronbach empleado en la Teoría Clásica de los Tests. En este caso, se desean altos valores de PSR indicativos de que la varianza de las medidas de las personas en el constructo tiene un bajo componente de error. Sin embargo, dado que se suele preferir que no existan variaciones sustanciales en la severidad de los calificadores, los valores bajos de RSR (Rater Separation Reliability) serían los más apreciados (las diferencias observadas en la severidad de los calificadores serían atribuibles al error de medida). Un índice de separación de los examinados de 1.5 o un PSR de .7 representan un nivel aceptable de separación y se consideran el valor mínimo necesario para dividir la muestra en dos estratos distintos (en el caso de los examinados, bajo y alto rendimiento). Si la desviación típica verdadera fuese igual al error estándar promedio, ocurriría que G=1, SR=.50 y H=1.67. Por tanto, el error de medida influiría altamente en la variabilidad observada y no sería posible identificar fiablemente rangos de distinto nivel en el constructo. Los valores aceptables para los examinados, tareas y atributos deberían ser: SR>.70, G>1.5 y H>2. En la faceta de los calificadores serían deseables valores menores que indicarían que su estilo de calificación es homogéneo.

Adecuación de las categorías numéricas

Para determinar si las categorías numéricas empleadas por los calificadores (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, ajuste 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 de Outfit de las categorías superiores a 2.0 indican que la categoría de evaluación no ha sido utilizada de manera adecuada. Finalmente, se ha de observar si los pasos entre las categorías están ordenados monotonicamente 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.

Presentación e interpretación de los resultados básicos del ejemplo

Personas evaluadas

Como ejemplo, en la parte superior de la Tabla 1 se muestran los resultados de dos evaluados 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 Mo (Promedio de las calificaciones observadas). Los valores X y Mo 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 justo (Mj), que es la media de las calificaciones que recibiría el examinado de un calificador con un nivel medio de severidad. En las dos últimas columnas de la derecha aparecen os estadísticos de ajuste: se observa que los examinados 1 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 todos los evaluados (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 (DT=2.38 logits) de los evaluados. Las medidas de los candidatos oscilaron entre 11.26 logits y -7.19 logits. La fiabilidad de las puntuaciones es muy elevada (PSR=.95; G=4.4; H=6.2), 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%).

Tabla 1. Puntuaciones y estadisticos descriptivos de los examinados

Nota. Logit: Medida en logits; SE: Error estándar; X: suma de las calificaciones; r: número de calificaciones; MO: Promedio observado; MI: Promedio justo; Infit y Outfit: Estadisticos de ajuste. M: Media; DT: Desviación Típica; Max: Valor Máximo; Min: Valor Mínimo.

ID	Logit	SE	X	r	Mo	Mj	Infit	Outfit
1	3.24	.49	47.0	20.0	2.35	2.11	.89	.89
82	-2.25	.51	25.0	20.0	1.25	1.03	1.34	1.39
M	2.82	.54	40.6	19.9	2.04	2.04	.99	1.03
DT	3.38	.19	9.6	.7	.47	.45	.58	.85
Max	11.26	2.85	60.0	20.0	3.00	3.00	7.48	9.00
Min	-7.19	.62	6.0	12.0	.50	.11	.09	.11

Calificadores

En la Tabla 2 se muestran los promedios (en la escala de 0 a 3) de las evaluaciones de dos calificadores: Mo y Mj, las puntuaciones en severidad (en la escala logit), su precisión (SE), los estadísticos de ajuste y los índices clásicos de fiabilidad entre calificadores (consenso y consistencia). Se observa que la variabilidad de los calificadores en severidad es elevada (DT=1.36 logits). 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), GR y

Tabla 2. Puntuaciones y estadisticos descriptivos de los calificadores

ID	Logit	SE	X	r	Mo	Mj	Infit	Outfit	Rc,rc	% Acuerdo
332	196	.06	2240	1360	1.65	1.66	.94	.94	.71	48.4
39	-2.29	.06	3400	1380	2.46	2.49	1.01	1.56	.70	43.8
M	.00	.06	2751	1349	2.04	2.03	.99	1.04	.72	51.9
DT	1.36	.00	326	30	.25	.25	.20	.28	.09	4.40

Nota: Logit: Medida en logits; SE: Error estándar; X: suma de las calificaciones; r: número de calificaciones otorgadas; MO: Promedio observado de las calificaciones; MI: Promedio justo de las calificaciones; Infit y Outfit: Estadisticos de ajuste. Rc,rc: Estadístco de consistencia entre calificadores; % Acuerdo: Estadístico de consenso entre calificadores.

HR deberían ser bajos. Sin embargo, estos índices son muy elevados (RSR=.99; GR=9.9; HR=13.5): unos valores tan altos revelan que las diferencias observadas en severidad entre los calificadores son muy fiables. 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 justo (Mj), 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). Se ha de notar que la relación entre las escalas en Mj 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 (intracalificador) 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; DT=.09) 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) son moderados (media=51.9%; DT=4.40). En consecuencia, se ha de considerar que el grado de consistencia entre los calificadores es mayor que el grado de consenso.

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). Los estimadores de la dificultad de los elementos de ambas facetas son muy precisos: ISR_{Tarea}=.99; G_{Tarea}=9.9; H_{Tarea}=13.6; $ISR_{Atributo}$=.99; $G_{Atributo}$=9.9; $H_{Atributo}$=13.6.

Categorías de respuesta (rúbricas)

En la Tabla 3 aparecen los estadísticos correspondientes a las categorías de respuesta derivadas de la formulación (1). 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 monotonicamente 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 las categorías son modales: 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, se puede concluir que las categorías numéricas presentan una funcionalidad

Tabla 3. Estadístico de las categorias de respuesta

Categoria	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

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 el Modelo de Crédito Parcial (Véase la ecuación (2)). Este enfoque aporta evidencias sobre el uso de las categorías por cada calificador.

Efectos del calificador

La heterogeneidad de los calificadores al asignar las puntuaciones es considerada una fuente de error que decrementa la fiabilidad y validez de las calificaciones. Por ello, se han acuñado términos como sesgo, errores o efectos del calificador, existiendo una abundante literatura especializada en la que se categorizan, se definen estos efectos y se proponen procedimientos para detectarlos (Hung et al., 2012; Myford & Wolfe, 2003 y 2004; Engerhard & Wind, 2018). Del amplio catálogo de efectos del calificador, los investigadores han destacado la importancia de cuatro de ellos: la severidad/benignidad, tendencia central, halo y efecto de aleatoriedad. Se presenta a continuación la definición de cada uno de ellos y algunos de los procedimientos psicométricos más sencillos para detectarlos.

Severidad/Benignidad

La Severidad se define como la tendencia de un evaluador a asignar calificaciones que son más bajas que las que asignan otros evaluadores a las mismas personas. Por el contrario, la Benignidad es la tendencia a otorgar calificaciones más altas que las asignadas por otros evaluadores. La puntuación en logit o la Media justa correspondiente al calificador permiten identificar a los calificadores más severos y más benignos. En comparación con otros calificadores, las puntuaciones logit muy altas o muy bajas revelan el efecto de severidad y benignidad respectivamente. Por el contrario, en el caso de la media justa, las puntuaciones muy altas indican benignidad y las muy bajas severidad. Asimismo, el mapa de Wright permite detectar fácilmente la magnitud de la variabilidad de los calificadores en el continuo de Severidad/Benignidad y cuales de ellos ocupan los extremos de la variable.

Tendencia central

Tendencia sistemática a otorgar las calificaciones usando las categorías numéricas centrales. La formalización de MFRM mediante el Modelo de Crédito Parcial (fórmula 2) permite observar la distribución de las categorías usadas por cada calificador y detectar a los que tienden a concentrar sus calificaciones en las categorías centrales. Los bajos índices de fiabilidad de los examinados (PSR, GP, HP) y los valores bajos de Infit y Outfit (sobreajuste) de los calificadores manifiestan la presencia de este efecto. Myford & Wolfe (2004) apuntan a que la presencia de este efecto en la mayoría de los calificadores podría deberse a problemas en el sistema de categorías. En esos casos, es necesario reducir el número de categorías para lograr una medición eficiente.

Efecto de Halo

Tendencia de un evaluador a asignar a sus evaluados puntuaciones similares en atributos conceptualmente distintos. Un evaluador que presenta el efecto halo no puede distinguir fácilmente entre distintos tributos y, por lo tanto, asigna a un evaluado puntuaciones similares en todos ellos. Los bajos índices de fiabilidad de los atributos ($ISR_{Atributo}$, $G_{Atributo}$, $H_{Atributo}$) manifiestan la presencia del efecto de halo.

Efecto de Aleatoriedad

Los calificadores asignan las calificaciones de forma descuidada o aleatoria, de manera que las puntuaciones no están relacionadas con el rendimiento o capacidad de los participantes. El evaluador que asigna de manera aleatoria las puntuaciones es demasiado inconsistente en el uso de las escalas, mostrando más variabilidad aleatoria de la esperada. El orden de los examinados de un evaluador "aleatorio" será diferente al de los demás evaluadores (Myford &

Wolfe, 2004). En consecuencia, la correlación de las evaluaciones de ese calificador con los demás calificadores (Rc,rc) será baja. Asimismo, los elevados valores de desajuste del calificador pueden revelar el efecto de aleatoriedad.

Conectividad

En muchos estudios no es posible que todos los calificadores evalúen a todos los examinados en todas las tareas. En esos casos es necesario emplear diseños de bloques incompletos garantizando la *conectividad*, de forma que las estimaciones de los parámetros estén en la misma escala. Los requisitos mínimos de estos diseños de evaluación 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, 2015). El diseño más simple consiste en asignar a un calificador un subconjunto pequeño de las pruebas evaluadas por cada uno de los demás calificadores. En el estudio que se ha presentado como ejemplo se utilizó un diseño simple de rotación de los examinados y los calificadores similar al descrito por Tesio et al. (2015).

Conclusión

Los tests que incluyen ítems o tareas de respuesta abierta pueden tener mayor validez ecológica que las pruebas de elección múltiple. Sin embargo, las respuestas han de ser puntuadas por calificadores cuyo correcto proceder es la clave para obtener mediciones fiables y válidas.

La influencia del criterio del calificador en la puntuación otorgada es determinante. 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 idiosincraticos como el de halo y el de tendencia central (Myford & Wolfe, 2004). Los efectos del calificador han de ser considerados una amenaza para la validez de las puntuaciones 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 objetivo es identificar a los calificadores que han de ser informados sobre un posible sesgo en su evaluación y que han de ser entrenados para homogeneizar su proceder en evaluaciones futuras.

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 cuya propiedad denominada *objetividad específica* garantiza la obtención de medidas invariantes si los datos se ajustan a los supuestos (Engelhard, 2013).

Desde esta perspectiva, se ha descrito en este artículo el modelo MFRM, un modelo de tipo Rasch, que permite 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. Además de utilizar la escala logit como métrica común, se ha propuesto un procedimiento para expresar los valores invariantes de los elementos de las facetas en la escala numérica usada inicialmente para puntuar las respuesta (escala denominada media justa o imparcial).

Se han expuesto los estadísticos fundamentales que aporta el modelo para cuantificar el ajuste, la precisión de las medidas y la adecuación de las categorías numéricas. Además, se describe el mapa de Wright como un recurso muy útil para visualizar la variabilidad y localización en la escala de los examinados, los calificadores, las tareas y los atributos evaluados.

Asimismo se han descrito los principales efectos del calificador, frecuentemente considerados como fuentes de errores de medida: severidad/benignidad, halo, tendencia central y aleatoriedad.

Se ha ilustrado la formulación del modelo y el uso e interpretación de los estadísticos con 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 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 uso de MFRM se ha extendido en diversos contextos de la evaluación psicológica y educativa. Es paradigmática su utilización en los programas de evaluación a gran escala de la adquisición de segundas lenguas como el alemán (Eckes, 2005), el inglés (Farroki et al., 2012; Knoch et al., 2020; Wind & Schumaker, 2017), el portugués (Toffoli et al., 2016) y el español (Mendoza, 2018; Prieto, 2011 y 2015; Prieto & Nieto, 2014). Son muy interesantes las aplicaciones para medir la Creatividad (Hung et al., 2012; Primi et al., 2019), la ejecución musical (Wesolowsky et al., 2015 y 2016), el liderazgo (Barney, 2016), la calidad de las codificaciones en investigaciones observacionales (Gordon et al., 2021), la selección de estudiantes de arquitectura (Hernández-Ureña & Montero-Rojas, 2023) y la calidad docente (Börkan, 2017).

Esta versatilidad del modelo MFRM garantiza su expansión en el futuro.

Referencias

Barney, M. (2016). Calibrating Charisma: The many-facet Rasch model for leader measurement and automated coaching. Journal of Physics: Conference Series 772 012051. https://doi.org/10.1088/1742-6596/772/1/012051.

Bravo, A., & Fernández, J. (2000). La evaluación convencional frente a los nuevos modelos de evaluación auténtica. *Psicothema, 12*, 95-99.

Börkan, B. (2017). Exploring Variability Sources in Student Evaluation of Teaching via Many-Facet Rasch Model. *Journal of Measurement and Evaluation in Education and Psychology, 8*(1),15-33.

Eckes, T. (2005) Examining Rater Effects in TestDaF Writing and Speaking Performance Assessments: A Many-Facet Rasch Analysis. *Language Assessment Quarterly: An International Journal, 2*(3), 197-221.

Eckes, T. (2009). Many Facet Rasch Measurement. En 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 Division.

Eckes, T. (2015). Introduction to many-facet rasch measurement: Analyzing and evaluating rater-mediated assessments. Peter Lang

Engelhard, G., & Wind, S. A. (2018). Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments. Routledge.

Farrokhi, F., Esfandiari, R. & Schaefer, E. (2012). A many-facet rasch measurement of differential rater severity/leniency in three types of assessment. *JALT Journal*, 34.1, 79-102.

Gordon, R. A., Peng, F., Curby, T.W., Zinsser, K. M. (2021) An introduction to the many-facet Rasch model as a method to improve observational quality measures with an application to measuring the teaching of emotion skills. *Early Childhood Research Quarterly*, 55, 149-164.

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

Hernández-Ureña, O., & Montero-Rojas, E. (2023). Validation of a standardized performance test for selection of Architecture students with the Many-Facet Rasch Measurement Model. *Revista de Arquitectura (Bogotá), 25*(1), 3-11.

Hung, S., Chen, P. & Chen, H. (2012): Improving Creativity Performance Assessment: A Rater Effect Examination with Many Facet Rasch Model, *Creativity Research Journal, 24* (4), 345-357.

Knoch, U., Zhang, B. Y, Elder, C., Flynn, E., Huisman, A., Woodward-Kron, R., Manias, E.,

& McNamara, T. (2020) 'I will go to my grave fighting for grammar': Exploring the ability of language-trained raters to implement a professionally-relevant rating scale for writing. *Assessing Writing, 46,* 10488.

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

Linacre, J.M. (1989). Many-facet Rasch Measurement. Mesa Press.

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

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

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

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

Mendoza, A. (2018). El uso de Many-Facet Rasch Measurement para examinar la calidad del proceso de corrección de pruebas de desempeño. *Revista Mexicana de Investigación Educativa, 23* (77), 1-19.

Myford, C. M. & Wolfe, E. W. (2003) Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part I. *Journal of Applied Measurement, 4*(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*(2), 189-227.

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. y Nieto, E. (2014). Analysis of rater severity on written expression exam using Many Faceted Rasch Measurement. *Psicológica, 35,* 363-375.

Prieto, G. (2015). Análisis de un test de desempeño en expresión escrita mediante el modelo de MFRM. *Actualidades en Psicología, 29* (119), 1-17.

Primi, R., Silvia, P.J., Jauk, E, & Benedek, M. (2019). Applying Many-Facet Rasch Modeling in the Assessment of Creativity. *Psychology of Aesthetics, Creativity, and the Arts, 13*, (2), 176–186.

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.

Toffoli, S., Bornia, A. C., & De Andrade, D. F. (2016). Evaluation of open items using the many-facet Rasch model. *Journal of Applied Statistics, 43*, 2, 299-316.

Wesolowsky, B. C., Wind, S. A. & Engelhard, G. (2015) Rater fairness in music performance assessment: Evaluating model-data fit and differential rater functioning. *Musicae Scientiae, 19*(2), 147 –170.

Wesolowsky, B. C., Wind, S. A. & Engelhard, G. (2016). Examining rater precision in music performance assessment: an analysis of rating scale structure using the Multifaceted Rasch Partial Credit Model. *Music Perception, 33* (5), 662–678.

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.

Wu, M., Tam, H. P. & Jen, T. H. (2016). Educational Measurement for Applied Researchers. Theory into Practice. Springer.