Psicothema 2011. Vol. 23, nº 2, pp. 233-238
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

Evaluación de la ejecución mediante el modelo Many-Facet Rasch Measurement

Gerardo Prieto Adánez
Universidad de Salamanca

Este trabajo describe cómo se puede aplicar el modelo Many-Facet Rasch Measurement (MFRM) para medir los elementos de las facetas (examinados, ítems y calificadores) involucradas en las pruebas para evaluar la ejecución. Con este fin se incluye una introducción al MFRM, una descripción de los procedimientos de análisis y un ejemplo ilustrativo de sus potencialidades en el que se analizan, mediante el programa FACETS, las fuentes de variabilidad de la ejecución de los estudiantes en un test de expresión escrita del DELE (Diplomas de Español como Lengua Extranjera). Los resultados muestran que el procedimiento es especialmente útil para detectar a los calificadores que presentaron diferencias en el grado de severidad, debido a su diferente empleo de los criterios de asignación de las calificaciones. La medición de los elementos de cada faceta en una métrica común facilita la comprensión de los distintos aspectos que influyen en las evaluaciones y permite obtener medidas objetivas de los elementos de una faceta.

Performance assessment using the Many-Facet Rasch Measurement. This paper describes how the Many-Facet Rasch Measurement (MFRM) can be applied to constructed-response items and rater analysis. The article provides an introduction to MFRM, a description of facet analysis procedures, and an illustrative example to examine the effects of various sources of variability on students' performance on a DELE (Diplomas in Spanish as a Foreign Language) 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. MFRM facilitates comprehension of the assessment process as well as providing objective measurement of facet elements.

Los profesores de los conservatorios evalúan la competencia de los alumnos puntuando la ejecución de obras musicales. Los jueces de los campeonatos de gimnasia o de saltos de trampolín asignan calificaciones a los deportistas en función de la calidad de los ejercicios realizados. Los evaluadores de las pruebas de selectividad o de los exámenes que acreditan la competencia para expresarse por escrito en una lengua extranjera califican las redacciones de los candidatos. Todos estos casos son ejemplos de la metodología denominada *evaluación de la ejecución* (performance assessment), cuyo estatus ha escalado puestos en el ranking de los métodos de evaluación por el empuje de los partidarios de la *evaluación auténtica* (Bravo y Fernández, 2000; O'Malley y Pierce, 1996). Las calificaciones otorgadas en los ejemplos antes mencionados no dependen solo del nivel de los examinados en el constructo de interés, sino de la dificultad de las tareas, la severidad del calificador y el uso de las categorías de calificación. La dificultad de las tareas y el sistema de categorías pueden ser definidos por los diseñadores de las pruebas. Sin embargo, el comportamiento de los calificadores escapa en buena medida a su control, constituyendo una fuente importante de la varianza de las calificaciones que puede ser irrelevante para evaluar de forma válida el constructo de interés (Lane y Stone, 2006). Las diferencias entre los calificadores en la interpretación de las tareas y de las categorías de evaluación, en su severidad, el efecto de halo, el sesgo al aplicar las calificaciones a grupos de diferente género y cultura, etc., contribuyen al error de medida, a la validez y a la justicia de las evaluaciones. El objetivo de este trabajo es mostrar la utilidad de uno de estos modelos, denominado *Many-Facet Rasch Measurement* (Linacre, 1989). El modelo MFRM es una extensión de uno de los modelos de Rasch más conocidos: el Modelo de Crédito Parcial (Wright y Masters, 1982). Su propiedad más característica, denominada *objetividad específica*, implica que las medidas de las personas no dependen de las muestras de ítems y de calificadores. El modelo MFRM permite obtener estimaciones en una escala común de los parámetros de los elementos de las facetas implicadas en la evaluación (los evaluados, las tareas, los calificadores y las categorías de puntuación). En este trabajo se muestra la utilidad del MFRM utilizando como ejemplo el análisis de un examen de expresión escrita de nivel elemental (A1) del sistema de las pruebas para la obtención de los Diplomas de Español como Lengua Extranjera (DELE). Distintas versiones de MFRM ya han sido utilizadas para medir las facetas en exámenes de expresión oral y escrita del inglés como segunda lengua (Kondo-Brown, 2002; Park, 2004; Tyndall y Kenyon, 1996).

Fecha recepción: 19-5-10 • Fecha aceptación: 27-10-10
Correspondencia: Gerardo Prieto Adánez
Facultad de Psicología
Universidad de Salamanca
37005 Salamanca (Spain)
e-mail: gprieto@usal.es

El modelo MFRM

MFRM es una extensión del Modelo de Crédito Parcial para ítems politómicos en los que la ejecución de la persona es calificada mediante un conjunto de categorías ordenadas. El modelo es aplicable a los casos en los que existen diversas facetas de la medición (personas, ítems, calificadores, categorías ordenadas de evaluación, etc.) que pueden contribuir al error de medida. Este modelo permite representar, controlando el error de medida, la contribución aditiva de cada faceta al *logit* o logaritmo del cociente entre la probabilidad de que una persona reciba una calificación en una tarea (por ejemplo, 3) y la probabilidad de que reciba la calificación inmediatamente inferior (2).

En concreto,

$$\log (P_{nijk} / P_{nij(k-1)}) = B_n - D_i - C_j - F_{jk} \quad (1)$$

Siendo,

P_{nijk}= la probabilidad de que una persona *n* reciba la calificación *k* en el ítem *i* por el calificador *j*.

$P_{nij(k-1)}$= la probabilidad de que una persona *n* reciba la calificación inferior (*k-1*) en el ítem *i* por el calificador *j*.

B_n= nivel en la variable latente de la persona *n*.

D_i= dificultad del ítem *i*.

C_j= severidad del calificador *j*, y

F_{jk}= localización en la variable del *paso* entre las categorías adyacentes *k* y *k*-1 en el calificador *j*.

En la ecuación 1, el logit ($\log (P_{nijk} / P_{nij(k-1)})$) es la variable dependiente y las diversas facetas (personas, ítems, calificadores, 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 el ítem *i* depende de los efectos aditivos de la dificultad del atributo (D_i) , de la severidad del calificador (C_j), del nivel de ejecución de la persona (B_n) y del valor del paso entre las categorías k y k-1 (F_{jk}). En esta formulación del MFRM 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. Las sumas de los pasos son los estadísticos suficientes para estimar los parámetros (Linacre y Wright, 2002). La escala logit puede oscilar entre $0 \pm \infty$. El punto 0 se fija convencionalmente en el nivel medio de los ítems, los calificadores y las categorías, 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. 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 y Wolfe, 2004).

Los análisis con el modelo MFRM pueden ser realizados con el programa *FACETS* (Linacre, 2009). Este programa estima los parámetros mediante el método de estimación conjunta por máxima verosimilitud (JML). Aunque Fischer y Molenaar (1995) criticaron este procedimiento de estimación porque produce cierto sesgo en las estimaciones, el método presenta la ventaja de facilitar la estimación en condiciones tales como la presencia de datos incompletos (missing), anclaje de parámetros y categorías no empleadas en formatos tipo Likert.

Las propiedades y recursos de MFRM son las de los modelos de Rasch: medición conjunta, estadísticos suficientes, nivel intervalar, objetividad específica, cuantificación de la precisión a nivel local y análisis del ajuste al modelo de las personas, los ítems, los calificadores y las categorías de evaluación (Prieto y Delgado, 2003).

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, denominadas *Infit* y *Outfit*. *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 superiores a 2 revelan un desajuste severo que degrada las medidas (Linacre, 2009). FACETS aporta valores individuales de ajuste para los evaluados, los calificadores, los ítems y las categorías de calificación.

Correlación calificador-resto de los calificadores ($R_{c,rc}$). 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 0,30 permiten identificar a los evaluadores inconsistentes, en los que la ordenación de las personas difiere 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 (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 (personas, los ítems 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 y Wolfe, 2004). 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 cuál es la proporción de la varianza verdadera respecto de la varianza observada de las personas evaluadas:

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

Siendo SE_{Bn} el error típico 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.

Finalmente, ISR (*Item separation reliability*) se refiere a la fiabilidad de las diferencias entre las estimaciones de la dificultad de los ítems con los que se mide un atributo. Cuando se pretende incluir ítems de distinta dificultad para garantizar un muestreo adecuado de los distintos niveles del constructo evaluado, se desean altos valores de ISR.

Estadísticos de las categorías de evaluación. Para determinar si las categorías 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, 2002). 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 ha 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.

Método

Participantes

Conforman la muestra 1.428 candidatos que realizaron en el mes de mayo de 2009 el examen para la obtención del DELE Nivel A1 (usuario básico-acceso). Este diploma acredita la competencia lingüística suficiente para comprender y utilizar expresiones cotidianas de uso muy frecuente en cualquier lugar del mundo hispanohablante, encaminadas a satisfacer necesidades inmediatas de comunicación. La procedencia geográfica de los candidatos fue: Italia (60,5%), Marruecos (12,3%), Corea del Sur (8,5%), Grecia (7,4%), Turquía (5,8%) y Japón (5,5%).

Instrumento

El examen de expresión escrita estaba integrado por dos tareas. En la primera, el candidato hubo de cumplimentar un formulario de registro en un hotel en el que se incluían apartados para el nombre y apellidos, el lugar de nacimiento, la dirección, la forma de pago, los motivos del viaje, los intereses por actividades de ocio, etc. En la segunda tarea se pedía al candidato que escribiese un correo electrónico de un máximo de 30 palabras solicitando la participación en un concurso de televisión. En el texto se debían incluir fórmulas de saludo y despedida, descripción de la apariencia física y de los gustos personales.

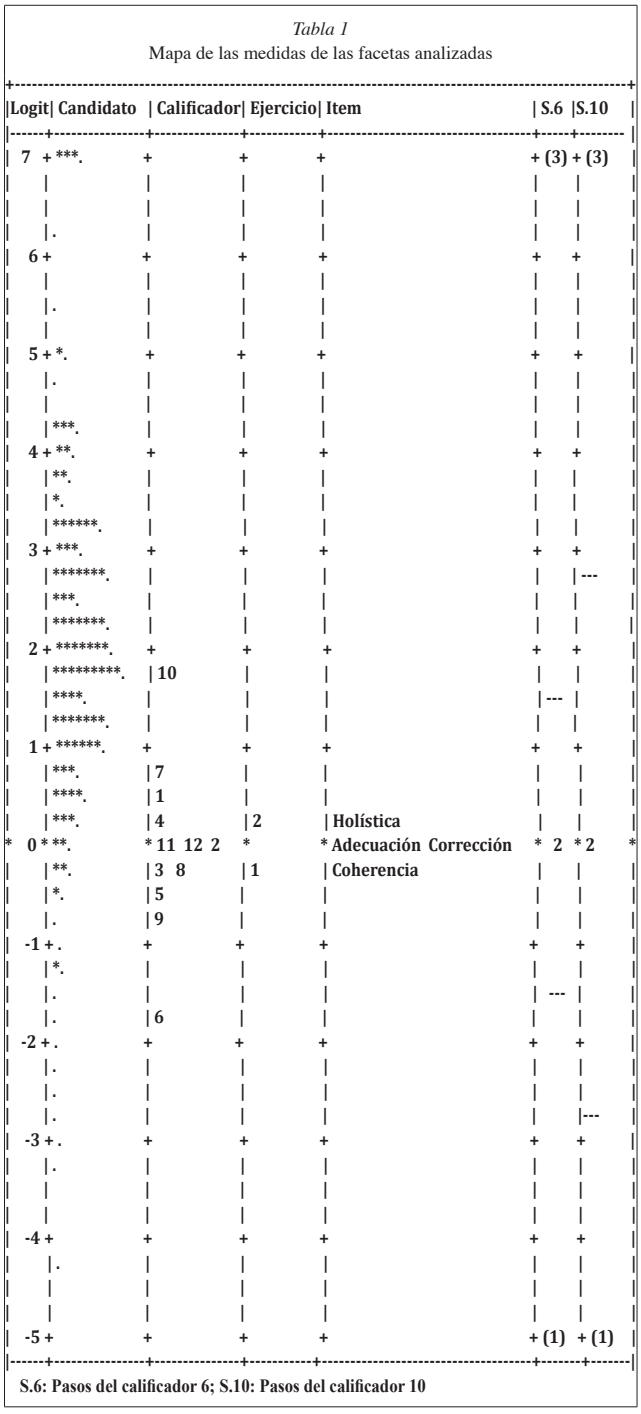
Procedimiento

En la evaluación de los exámenes participaron 12 calificadores. Los textos de cada candidato fueron evaluados independientemente por dos calificadores, los cuales otorgaron sus evaluaciones en cuatro variables o ítems: evaluación *holística* (una calificación global que refleja la eficacia comunicativa en la ejecución de cada tarea), *adecuación* (adaptación del texto al contexto, a los interlocutores y a las intenciones comunicativas), *corrección* (conocimiento y capacidad de uso de las categorías gramaticales y de las reglas morfosintácticas) y *coherencia* (control de los recursos necesarios para establecer relaciones entre el discurso y la situación de comunicación: participantes, circunstancias espacio-temporales, etc.). El rendimiento en cada variable fue puntuado en una escala de 0 a 3. Los calificadores fueron informados de los criterios de asignación de las puntuaciones (*scoring rubrics*). La puntuación directa de un candidato en el examen es la suma de los valores asignados a los ítems en las dos tareas. El hecho de que todos los calificadores no puntuasen a todos los candidatos, un procedimiento inasumible por razones prácticas, supone una limitación para el escalamiento conjunto, especialmente de las diferencias en severidad de los calificadores. Por ello, se ha puesto énfasis en las comparaciones de los calificadores que puntuaron a los mismos grupos de candidatos.

Resultados

En la tabla 1 se muestra el mapa de la variable, un recurso muy útil para visualizar la medición conjunta de la competencia de los candidatos, la severidad de los calificadores y la dificultad de los ejercicios, de los ítems y de los pasos entre las categorías de evaluación. La calibración de los elementos de todas las facetas en la misma escala de intervalos (logit) permite interpretar los resultados en el mismo marco de referencia. En la columna *Candidato* de la tabla 1 se representa la distribución de los candidatos en la escala. Cada asterisco (*) representa a 13 personas y cada punto a una frecuencia inferior. Los candidatos con mayor puntuación se sitúan en la parte superior de la columna y en la parte inferior los de menor puntuación. Se observa que hay una gran variabilidad en la competencia de los candidatos (entre -4,34 y 7,56). En la columna *Calificador* aparecen los valores de severidad de los calificadores, siendo el número 10 el más severo (1,64) y el número 6 el más benigno (-1,63). La variabilidad de los calificadores en severidad es moderadamente alta, mayor de la deseable. En la columna *Ejercicio* se muestra el nivel de dificultad de los ejercicios que han integrado el examen. Se observa que es escasa la diferencia en dificultad de ambas tareas. En la columna *Ítem* aparecen los valores de dificultad relativa de las variables con las que se han calificado los ejercicios. Se ha de notar que las diferencias en dificultad son pequeñas. Finalmente, en la columna *Categoría* se muestran, mediante líneas, la situación de los valores de los *pasos* entre las categorías utilizadas (de 1 hasta 3) para puntuar las respuestas de los candidatos (ningún candidato fue asignado a la categoría 0). Se destacan los pasos de los calificadores que presentan la mayor diferencia en severidad. Se puede observar que los pasos de ambos calificadores difieren notablemente. Esto indica que los criterios de asignación de las calificaciones (*scoring rubrics*) no han sido utilizados por ambos de manera uniforme. En adelante, se comenta de forma más detallada las propiedades psicométricas de los valores de las facetas analizadas.

Tabla 1
Mapa de las medidas de las facetas analizadas



Candidatos

En la tabla 2 aparecen los principales estadísticos de las puntuaciones de los candidatos que realizaron el examen de expresión escrita. Se aprecia un alto rendimiento medio (1,99) y una gran variabilidad (DT= 1,61). Las medidas de los candidatos oscilaron entre 7,56 y -4,34. La fiabilidad de las puntuaciones es elevada (PSR= 0,83). Este estadístico puede interpretarse como el grado en que las puntuaciones en el examen permiten diferenciar fiablemente entre los diferentes niveles de competencia de los examinados. El porcentaje de los candidatos que presentan un desajuste severo con las predicciones del modelo es muy bajo (3,9).

Calificadores

En la tabla 3 se muestran los promedios de las evaluaciones de los calificadores (en la escala de 1 a 3), las puntuaciones en severidad (logit), su precisión (error estándar) y los estadísticos de ajuste. Se observa que la variabilidad de los calificadores en severidad es elevada. Este dato no es deseable. Idealmente las variaciones en severidad habrían de ser despreciables y atribuibles al error de medida, por lo que se espera que RSR (*Rater Separation Reliability*) presente un valor bajo. En este caso, el índice RSR (Rater Separation Reliability) es muy elevado (0,98): 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 (los errores estándar oscilan entre 0,04 y 0,22). Los estadísticos de ajuste (entre 0,81 y 1,37) indican que todos los calificadores muestran una alta consistencia intra-calificador en sus evaluaciones. Las correlaciones de cada calificador con el resto ($R_{c,rc}$) oscilaron entre 0,41 y 0,71 (valor promedio= 0,57) indicando

Tabla 2
Resumen de los estadísticos de los candidatos (N=1428)

Media	1,99
Desviación típica	1,61
Varianza de error (Media (SE^2_{Bn}):	,38
Fiabilidad (PSR)	,83
Media de Infit	,99
Desviación típica de Infit	,41
Media de Outfit	1,01
Desviación típica de Outfit	,67
Porcentaje con desajuste severo *	3,9%

* Infit y/o Outfit > 2

Tabla 3
Valores y estadísticos de los calificadores (N= 12)

Calificador	Promedio	Severidad	Error estándar	Infit	Outfit	$R_{c,rc}$
1	2,2	,58	,22	,99	,98	,64
2	2,4	,07	,05	1,01	1,02	,58
3	2,6	-,20	,04	,92	,81	,53
4	2,6	,27	,04	,97	,94	,53
5	2,5	-,55	,05	,99	,99	,56
6	2,7	-1,63	,05	1,07	1,25	,50
7	2,2	,67	,07	,89	,86	,70
8	2,6	-,20	,04	1,12	1,20	,41
9	2,5	-,66	,04	,98	,97	,52
10	2,0	1,64	,05	,88	,84	,71
11	2,4	-,05	,07	1,25	1,37	,54
12	2,3	,07	,21	1,06	1,08	,63

RSR (Rater Separation Reliability)= ,98

una suficiente consistencia entre los calificadores en la ordenación de los examinados en competencia.

Se observa que los calificadores 6 y 10 difieren sustancialmente del resto en su grado de severidad: el calificador 6 es el más benigno (-1,63) y el calificador 10 el más severo (1,64). Los promedios de las calificaciones directas de ambos calificadores difieren en 0,70 puntos, una diferencia notable si se toma en consideración el estrecho rango de la escala (1-3). A pesar de haber evaluado a los mismos candidatos, ambos calificadores difieren notablemente en los porcentajes de asignaciones en cada categoría y en los valores de los pasos de las curvas características de las categorías (tabla 4). Este aspecto manifiesta que los criterios de asignación de las calificaciones (*scoring rubrics*) no han sido utilizados por ambos de manera uniforme.

Tabla 4
Estadísticos de las categorías usadas por los calificadores con los valores extremos en severidad

Calificador	Categoría	% Evaluaciones	Paso
6	1	2	—
	2	22	-1,47
	3	76	1,47
10	1	15	—
	2	69	-2,6
	3	16	2,6

% evaluaciones: porcentaje de evaluaciones de cada calificador en cada categoría.
Paso: valor del paso entre las categorías sucesivas

Ejercicios e ítems

Son escasas las diferencias en dificultad de los dos ejercicios utilizados en el examen (0,40 logits). Ambos ejercicios se ajustan al modelo (los valores *Outfit* de los ejercicios 1 y 2 son 0,96 y 1,03, respectivamente). La precisión de las estimaciones de la dificultad es muy alta (el error típico de medida es de 0,02 en ambos casos).

Como en el caso de los ejercicios, se observa que son pequeñas las diferencias en dificultad entre los ítems empleados para puntuar los ejercicios (la dificultad oscila entre 0,25 y -0,29 logit). El ajuste de los ítems es adecuado (sus valores *Outfit* oscilan entre 0,85 y 1,26). Las elevadas correlaciones ítem-escala (en un rango entre 0,55 y 0,63) manifiestan que existe un patrón semejante de competencia en los dominios evaluados, por lo que es adecuado combinarlos en una única puntuación para reflejar el rendimiento de los candidatos.

Discusión

En el examen de expresión escrita analizado se ha observado que los examinados obtuvieron un alto nivel de rendimiento (1,99 logits en promedio) y alta variabilidad (DT= 1,61). La precisión de las estimaciones de los candidatos fue adecuada (PSR= 0,83). Solo un 3,9% presentaron un severo desajuste con el modelo. Los ejercicios utilizados y los ítems han diferido poco en dificultad. Las elevadas correlaciones ítem-escala revelaron que existe un patrón semejante de competencia en las variables evaluadas, por lo que es adecuado combinarlas en una única puntuación para reflejar el rendimiento de los candidatos. Aunque dos calificadores (6 y 10), que puntuaron los exámenes de los mismos alumnos, difirieron netamente en su grado de severidad, el resto presentó diferencias menores. Los estadísticos de ajuste manifestaron que todos los calificadores muestran una alta consistencia intra-calificador en sus evaluaciones. La concordancia entre calificadores fue aceptable, dado que las correlaciones de cada calificador con el resto ($R_{c,rc}$) manifiestan una consistencia suficiente entre los calificadores en la ordenación de los examinados en competencia. La diferencia entre los calificadores con valores extremos en el grado de severidad pudo ser debida a que los criterios de asignación de las calificaciones (*scoring rubrics*) no han sido utilizados por ambos de manera uniforme.

El ejemplo analizado ilustra las potencialidades de MFRM para obtener medidas objetivas de las facetas involucradas en la evaluación de la ejecución: examinados, calificadores, tareas, ítems y categorías. La medición de los elementos de cada faceta en una métrica común facilita la comprensión de los distintos aspectos que influyen en las evaluaciones y permite obtener medidas de los elementos de una faceta que son independientes del resto, corrigiendo sus influencias idiosincráticas (Park, 2004). En particular, los resultados de este estudio permiten recomendar que los programas de evaluación estables entrenen a los calificadores para que utilicen de manera uniforme los criterios de asignación de las puntuaciones.

Referencias

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

Cronbach, L.J., Gleser, G.C., Nanda, H., y Rajaratnam, N. (1972). *The dependability of behavioural measurements*. New York: Wiley.

Fischer, G.H., y Molenaar, I.W. (1995). *Rasch models: Foundations, recent developments and applications*. New York: Springer-Verlag.

Kondo-Brown, K. (2002). An analysis of rater bias with FACETS in measuring Japanese L2 writing performance. *Language Testing, 19*, 1-29.

Lane, S., y 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. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3,* 85-106.

Linacre, J.M. (2009). *FACETS (*Computer program, version 3.66.1*)*. Chicago: MESA Press.

Linacre, J.M., y Wright, B.D. (2002). Construction of measures from Many-Facet Data. *Journal of Applied Measurement, 3*, 484-509.

Myford, C.M., y Wolfe, E.W. (2004). Detecting and measuring rater effects using Many-Facet Rasch Measurement: Part I. En E.V. Smith y R.M. Smith (Eds.*), Introduction to Rasch Measurement* (pp. 460-515). Maple Grove, MN: JAM Press.

O'Malley, J.M., y Pierce, L.V. (1996). *Authentic assessment for English Language Learners: Practical approaches for teachers*. New York: Addison-Wesley.

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

Tyndall, B., y Kenyon, D.M. (1996). Validation of a new holistic rating scale using Rasch multi-faceted analysis. En A. Cumming y 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., y Masters, G.N. (1982). *Rating scale analysis*. Chicago: MESA Press.