RMIE, 2018, VOL. 23, NÚM. 77, PP. 597-625 (ISSN: 14056666)

Investigación

EL USO DE MANY-FACET RASCH MEASUREMENT PARA EXAMINAR LA CALIDAD DEL PROCESO DE CORRECCIÓN DE PRUEBAS DE DESEMPEÑO

ARTURO MENDOZA RAMOS

Resumen:
Las pruebas de desempeño son criticadas por la supuesta falta de consistencia en los resultados que otorgan los evaluadores. Sin embargo, herramientas estadísticas como *Many-Facet Rasch Measurement* (MFRM) son útiles para examinar la calidad del proceso de evaluación en pruebas con múltiples facetas de variabilidad. El objetivo de este artículo es dar a conocer el funcionamiento y aportaciones de MFRM en pruebas de desempeño. El estudio se realizó con estudiantes universitarios no hispanohablantes que sustentaron una prueba escrita de Español con fines académicos. Los resultados mostraron niveles de severidad e indulgencia adecuados por parte de los evaluadores, y de dificultad en las tareas y en la rúbrica analítica empleada para evaluar los textos. Del estudio se concluye la utilidad de MFRM para examinar el proceso de corrección de pruebas de desempeño.

Abstract:
Performance assessments are criticized for the assumed lack of consistency in the results delivered by evaluators. However, statistical tools such as the Many-Facet Rasch Measurement (MFRM) are useful for examining the quality of the assessment process in tests with multiple facets of variability. The objective of this article is to present the functioning and contributions of MFRM on performance assessments. The study was carried out with non-Spanish-speaking university students who completed a written test of Spanish with academic ends. The results showed adequate levels of severity and indulgence among the evaluators, as well as in task difficulty and the analytical rubric employed to evaluate the texts. The conclusion is that MFRM is useful for examining the process of evaluating performance assessments.

Palabras clave: análisis estadístico, evaluación cuantitativa, evaluación académica, exámenes y certificación.
Keywords: statistical analysis, quantitative evaluation, academic evaluation, examinations and certification.

Arturo Mendoza Ramos: investigador de la Universidad Nacional Autónoma de México, Escuela Nacional de Lenguas, Lingüística y Traducción, Departamento de Lingüística Aplicada. Circuito interior s/n, Ciudad Universitaria, 04510, Ciudad de México, México. CE: a.mendoza@enallt.unam.mx

Introducción

En los exámenes o pruebas en los que se evalúa el desempeño o la ejecución se diseñan tareas que requieren de la valoración de un juez o evaluador para que determine la habilidad del examinado. Estas pruebas son recurrentes en diversos ámbitos educativos (Bond y Fox, 2007; Linacre, 2013). En el caso de los exámenes de lenguas para evaluar la expresión escrita y oral, se emplean pruebas de respuesta construida; es decir, se requiere de la producción a partir de un *input* o tarea para que el sustentante pueda ser evaluado en sus habilidades orales o escritas. Posteriormente, el desempeño es observado y valorado por un corrector, quien mediante una rúbrica emite puntajes parciales –en este caso analítica– o globales –a través de una holística– para el candidato. En el caso de los exámenes de certificación de alto impacto, es decir, aquellos que conllevan consecuencias importantes en la vida de los examinados, con el fin de garantizar la equidad y justicia en la evaluación es común que intervengan dos evaluadores (Knoch, 2007) y en caso de discrepancia, generalmente interviene un tercero.

Sin embargo, una de las dificultades dentro del proceso de evaluación se debe a la falta de consistencia de los resultados que otorgan los jueces (East, 2009; Eckes, 2009; Hamp-Lyons, 2007), lo cual pone en tela de juicio la calidad de los resultados otorgados y, por ende, del proceso evaluatorio (Attali, Lewis y Steier, 2012; Rezaei y Lovorn, 2010). Pese a que el diseño de la escala de valoración, el proceso de capacitación y de monitoreo de los jueces son indispensables para armonizar los criterios de los correctores, los resultados no siempre son consistentes (Attali, Lewis y Steier, 2012; Knoch, 2007; Huang y Foote, 2010).

Por esta razón, existen análisis estadísticos como *Many-Facet Rasch Measurement* (MFRM) que permiten determinar la calidad del proceso de evaluación cuando intervienen distintas facetas de variabilidad (Eckes, 2009). En el caso de las pruebas de respuesta construida de expresión escrita, las facetas que se presentan son: el nivel del examinado, el evaluador, la tarea y las categorías de la rúbrica analítica. Autores como Knoch (2007, 2009), Eckes (2009), Prieto (2011), Prieto y Nieto (2014) y Wind y Engelhard (2013) han empleado el análisis de MFRM para evidenciar el funcionamiento de las variables de las pruebas escritas (de ahora en adelante, PE) de los exámenes de lenguas administrados a gran escala. Este tipo de estudios es de especial relevancia puesto que permiten determinar

la dificultad de una tarea escrita, el funcionamiento de una rúbrica y la dificultad de sus categorías así como la severidad o indulgencia de los evaluadores al momento de corregir un texto.

El programa Facets

Linacre (2015), siguiendo el modelo de Rasch, diseñó el programa Facets para ejecutar el análisis de facetas múltiples que incluyen reactivos politómicos (véase Linacre, 2012a, 2012b, 2012c, 2012d, 2013 para información detallada del funcionamiento del programa). El modelo permite examinar hasta qué punto las variables (consideradas como facetas) contribuyen al error de medida; una característica del mismo consiste en que los puntajes de los sustentantes no dependen del evaluador ni de la tarea: objetividad específica (Bond y Fox, 2007; Eckes, 2009; Engelhard, 2008; Prieto, 2011). Eckes (2009:4) menciona que esta objetividad específica se refiere a que las mediciones del examinado son independientes del examen y del proceso de corrección; es decir, que cuando las observaciones en un modelo de Rasch son lo suficientemente precisas, la medición de los examinados es invariable a las tareas o evaluadores. Para que el programa funcione adecuadamente no se requiere, entonces, que los evaluadores presenten puntajes exactos sino, más bien, consistentes. El programa calibra las distintas facetas de manera simultánea y otorga un resultado equivalente en una escala de intervalo medida en lógitos (*log-odd units*).

El lógito es el logaritmo natural que se obtiene de la división entre dos probabilidades. En el caso de ítems politómicos (como es el de la rúbrica analítica aquí empleada con valores de 0, 1, 2 y 3), este logaritmo representa la probabilidad de obtener un puntaje y el puntaje inmediatamente inferior. La información se reporta en una escala conocida como mapa de variabilidad, en la cual se pueden comparar las distintas facetas en términos de su dificultad o severidad.

El programa Facets presenta múltiples ventajas de aplicación. En primera instancia, permite ajustar los valores de los sustentantes, los evaluadores, las tareas y las categorías de la rúbrica. Además, Eckes (2009), Knoch (2007), Myford y Wolfe (2003, 2004) y Prieto (2011) mencionan que otro de los grandes potenciales de MFRM es que nos permite identificar diversos efectos del evaluador: grado de severidad, efecto halo y el efecto de la tendencia central. Asimismo, otra de las ventajas del modelo es que permite

determinar el grado en que las calificaciones otorgadas por un evaluador correlacionan o son consistentes con las de los demás (correlaciones inter-evaluador). Finalmente, una de las ventajas más importantes de Facets es que se puede trabajar con datos incompletos (Eckes, 2009; Esfandiari y Myford, 2013; Linacre, 2012b, Linacre, 2013). Esto significa que no es requisito indispensable que todos los correctores corrijan a todos los sustentantes en todas las tareas sino, más bien, que existan interacciones entre todos los correctores. Esto quiere decir que es necesario que el examinado sea observado al menos por dos correctores en cada una de las pruebas y los correctores deberán ser pareados siempre con un corrector distinto.

Descripción del modelo MFRM

El modelo de Rasch para multifacetas de Linacre (2015) considera: el examinado, el juez o evaluador, la tarea y las categorías de la rúbrica analítica; por ejemplo, en el caso de pruebas escritas. Este modelo es esencialmente aditivo y se basa en una transformación logística de las puntuaciones observadas en lógitos. En este caso, el lógito representa la variable dependiente, mientras que las diversas facetas (la habilidad de la persona, la severidad del evaluador, la dificultad de la tarea y la dificultad del ítem) corresponden a la variable independiente. La escala del lógito puede oscilar entre 0 y ± ∞ (Linacre, 2012b, 2013). El cero representa el punto medio de cada una de las facetas, aunque cada una puede variar libremente de forma independiente dependiendo de la dificultad de la tarea, la severidad del evaluador y de los criterios de la rúbrica. Los estadísticos obtenidos pueden ser a nivel faceta o también grupal (Myford y Wolfe, 2004). Entonces, en este modelo, el lógito corresponde a la probabilidad de que una persona reciba una calificación específica en una tarea, en una categoría y por un evaluador, entre la probabilidad de que reciba una calificación inmediatamente inferior (Pnmijk / Pnmij (k–1)). La ecuación que se desprende es la siguiente:

$$\log (P_{nmijk} / P_{nmij\,(k-1)}) = B_n - A_m - D_i - C_j - F_k$$

Donde:

P_{nmijk} = la probabilidad de que una persona en la tarea *m*, sea evaluada en el ítem *i* por el evaluador *j* y reciba la calificación *k*

P_{nmij} ($_{k-1}$)) = la probabilidad de que una persona en la tarea *m*, sea evaluada en el ítem *i* por el evaluador *j* y reciba la calificación inmediatamente inferior (*k–1*)
B_n = la habilidad de la persona (p.ej., alumno 1)
A_m = el reto de la tarea (p.ej., ensayo)
D_i = la dificultad de la categoría de la rúbrica (p.ej., ortografía)
C_j = el nivel de severidad del evaluador (p.ej., evaluador 1)
F_k = el límite entre recibir el puntaje *k* y el puntaje *k–1* (p.ej., 2 y 1)
(Ecuación tomada de Linacre (2013:13) con ejemplos adaptados a tareas de producción escrita).

Entonces, el lógito se refiere a la probabilidad (Pnmijk / Pnmij (*k–1*)) de que un examinado con una cierta habilidad, *Bn*, al ser observado tras desempeñar una tarea con un reto o dificultad específica, *Am*, sea evaluado en un ítem con una cierta dificultad, *Di*, por un evaluador con determinado grado de severidad, *Cj*, reciba una puntuación *k* (Linacre 2013: 13). Para ejemplificar cómo se leería esta ecuación en una PE, se podría plantear de la siguiente manera: la probabilidad de que el examinado Jorge (*n*, persona) sea observado en el escrito de un ensayo (*m*, tarea), y sea evaluado en ortografía (*i,* categoría de la rúbrica) por el evaluador Andrés (*j*, juez o evaluador) con un puntaje de 2 (*k*, escala de la rúbrica en rangos de 0 a 4 puntos).

Reporte informativo de Facets

Facets ofrece información útil para cada una de las distintas facetas de variabilidad. La primera tabla de utilidad que arroja se conoce como mapa de variabilidad o mapa de Wright que nos muestra las distintas facetas que se consideraron en el estudio: el examinado, el evaluador, la tarea y las categorías de la rúbrica. En la primera columna aparece la medida, misma que se encuentra determinada en lógitos. Esta medida especial que genera el programa Facets establece un cero convencional como el punto medio de las facetas en cuestión. Linacre (2012b:4) menciona que típicamente estas medidas en lógitos se encuentran en un rango de –5 a 5 lógitos. Esta medida parte de un cero establecido de manera convencional –punto medio de habilidad, dificultad o severidad– y puede incrementar de forma positiva o negativa. Generalmente solo la faceta de los examinados es positiva (convención adoptada para la educación, Linacre, 2012:26),

y se espera una distribución de estos a lo largo de toda la escala, pues la habilidad de cada uno es distinta. Entre mayor sea la medida del lógito en la escala positiva, significa que la habilidad del examinado aumenta. Contrariamente, entre mayor sea la medida del lógito, pero en la escala negativa, menor será su habilidad (consúltese Linacre 2012a:15 para ver los antecedentes de la definición de medición en las ciencias sociales). En el caso de los evaluadores, las tareas y las categorías de la rúbrica, las facetas son negativas, lo que significa que entre mayor sea el lógito en la escala positiva, la severidad del evaluador será mayor, y la dificultad de la tarea y de las categorías de la rúbrica aumentará. En el caso opuesto, entre mayor sea el valor del lógito en la escala positiva, mayor será la indulgencia del evaluador, y la dificultad de la tarea y de las categorías de la rúbrica disminuirá.

Ahora bien, pese a que no existe un parámetro definido para determinar qué tan aceptable podría ser el rango de severidad de los evaluadores, Eckes (2012) ejemplifica que este se considera sustancial cuando el rango de severidad –en lógitos– excede la cuarta parte del rango de la habilidad de los examinados –en lógitos– (Eckes 2012:280). En el caso de las tareas y de las categorías de la rúbrica, dependiendo del tipo de examen (por ejemplo, colocación, diagnóstico o de certificación), se esperará un grado de variabilidad distinto para estas dos facetas.

En cada una de las facetas se encuentran los elementos que hacen referencia a un examinado en particular, un corrector que lo evalúa, una tarea específica o una categoría de la rúbrica que recibe un puntaje. Entonces, idealmente lo que se busca es que cualesquiera de las facetas negativas (evaluador, tarea y categorías de la rúbrica) se encuentren cerca del cero (de lo contrario significa que están produciendo variabilidad indeseada en los resultados de los examinados). Finalmente, en la última columna aparecen los valores con los cuales se puntuó cada categoría. La línea punteada entre estos rangos denota el nivel umbral en el cual un candidato "x" habría tenido la misma posibilidad de haber caído en un rango o en otro.

Si bien es cierto que el mapa de variabilidad es de gran utilidad para forjarse una rápida imagen de qué tan fiable es el proceso de evaluación, no proporciona un reporte detallado de la calidad de cada una de las facetas. Por ello, el programa Facets arroja tablas adicionales con un reporte detallado de la calidad de cada una de las facetas de variabilidad.

Aquí es donde se encuentran los datos más importantes del análisis de calidad de las tareas, de la rúbrica y de los jueces. El número de tablas que arroja el programa depende de las facetas con que se cuenta. En cada tabla se ofrece la medición de los elementos, en lógitos. En el caso de los examinados, la medición nos permite saber cuál es su habilidad, mientras que en las demás facetas, la medición establece qué tan severo/indulgente es un evaluador o que tan difícil/fácil es una tarea o la categoría de una rúbrica. El error estándar (SE) del modelo nos muestra la precisión de la medición, entre más cercano al cero, más precisa será la estimación de la medición. Naturalmente, entre más observaciones tengamos por parte de un evaluador, mayor precisión tendremos en la manera mediante la cual evalúa.

Los índices de ajuste son los valores más importantes para estimar la calidad de la medición. Estos índices son los que nos sirven para examinar la calidad de cada una de las facetas en cuestión (el examinado, la tarea, el evaluador y las categorías de la rúbrica). Estos índices de ajuste se denominan *Infit MnSq y Outfit MnSq*. El valor *MnSq* (*mean-square*) hace referencia al estadístico de ajuste de la Chi-cuadrada entre sus grados de libertad (para mayor información, consúltese Linacre, 2012b:39). El índice de ajuste muestra si los puntajes observados se parecen o no a los predichos por el modelo. Estas predicciones se llevan a cabo con base en el nivel del sustentante, la severidad del evaluador y la dificultad de la tarea. A la diferencia entre los puntajes observados de los esperados se le conoce como residuales estandarizados, e idealmente deberían ser iguales a 0. Cuando estos residuales se elevan al cuadrado y se suman a lo largo de las diferentes facetas y elementos de cada una, se obtienen los índices de ajuste para el modelo (Eckes, 2009, Linacre, 2012b). El índice de ajuste *Infit* (*inlier fit*) es la media ponderada de los cuadrados de las diferencias estandarizadas y el *Outfit* (*outlier fit*) es la media no ponderada de los cuadrados de las diferencias estandarizadas (Linacre, 2013:191). El índice *Outfit* es particularmente sensible a los resultados atípicos, mientras que el *Infit* es sensible a los patrones de respuestas específicas. Dicho en otras palabras, el índice *Outfit* es sensible a las respuestas con valores extremos (que un estudiante conteste correctamente ítems difíciles y no los intermedios), mientras que el *Infit* es sensible al patrón de respuestas (que un estudiante presente inconsistencias en su patrón de respuestas). Los valores del *Outfit* e *Infit* pueden oscilar en un rango de 0 a ∞, aunque el valor

esperado es 1. Si bien no existe un consenso sobre los rangos aceptables del *Outfit* e *Infit*, Linacre (2012b) ofrece estos valores de interpretación. La tabla 1 muestra dichos valores.

TABLA 1
Interpretación del estadístico de ajuste de las medias cuadradas

> 2.0	Distorsión o degradación del sistema de medición
1.5 – 2.0	No productivo para la medición, pero no degrada la medición
0.5 – 1.5	Productivo para la medición
< 0.5	Poco productivo para la medición, pero no la degrada

Tomado de Linacre (2012b:15) y traducido por el investigador.

Los valores mayores a 1 indican un desajuste mayor a lo esperado, y si este valor llega a 2, se considera como un severo desajuste (las observaciones serían demasiado impredecibles), lo cual significa que las observaciones resultan aleatorias y poco sistemáticas (Linacre, 2012b). Finalmente, los valores menores a 1 se pueden interpretar como sobreajustes (las observaciones son demasiado predecibles), y si son menores a 0.5 son poco productivas para la medición. No obstante, otros autores sugieren un rango más restringido: 0.70 como límite inferior y 1.30 como límite superior (Bond y Fox, 2007; McNamara, 1996; citados en Eckes, 2009:18). De hecho, para las evaluaciones de alto impacto; es decir, en aquellas que conllevan consecuencias importantes en la vida de los examinados, Linacre (2012b) sugiere valores entre 0.8 y 1.2 (p.16). Wright y Linacre (1994) afirman que la media cuadrada de 1.2 indica que hay un 20% de aleatoriedad en los datos, y que una media cuadrada de 0.7 indica una deficiencia de 30% en la aleatoriedad predicha por el modelo de Rasch (en Wright y Linacre, 1994:370).

Objetivo de la investigación

El objetivo de este estudio consistió en examinar la calidad del proceso de evaluación de pruebas escritas de un examen de español como lengua extranjera mediante el programa MFRM.

La pregunta de investigación que se formuló fue: *¿Constituye MFRM una herramienta útil para examinar la calidad del proceso de evaluación de pruebas de ejecución o desempeño?*

Métodos e instrumentos

La muestra

Participaron 100 estudiantes universitarios no hispanohablantes. La muestra se recolectó entre los semestres académicos de 2014 y 2015.

Las tareas de la PE

En la PE del examen se les pidió a los sustentantes que escribieran dos textos: descripción y contraste de gráficos, y un ensayo argumentativo (para mayor información sobre las tareas de la PE, consúltese Mendoza, 2015). La prueba fue resuelta sin apoyo de diccionarios, traductores o correctores automáticos de texto.

La rúbrica analítica

Después de un detallado proceso de diseño de la rúbrica (consúltese Mendoza y Knoch, 2018 para mayor información sobre el proceso de diseño y validación de la rúbrica), se elaboró una versión final con la cual fueron evaluados los examinados. En esta rúbrica se incluyeron tres componentes de la competencia comunicativa; a saber: competencia sociolingüística, pragmática y lingüística. Se incluyeron siete categorías: logro de la tarea, desarrollo de ideas, coherencia, cohesión, léxico, gramática y vocabulario (ver anexo 1).

Los evaluadores

Seis correctores evaluaron los textos producidos por los 100 estudiantes. Los evaluadores 1 y 2 (la responsable del examen y el investigador, respectivamente) eran los únicos experimentados en el uso y diseño de rúbricas. Los otros cuatro eran profesores de español como lengua extranjera, pero sin experiencia.

El proceso de evaluación

De forma inicial, los evaluadores 1 y 2 corrigieron los textos de los 100 estudiantes en ambas tareas. Una vez que los valoraron, se discutieron las discrepancias en los resultados y las dificultades que se presentaron con algunos textos. Posteriormente, cada uno de los otros cuatro jueces

también evaluó a los 100 estudiantes, aunque solamente en una de las dos tareas. Para lograr la mayor cantidad de interacciones posibles, se buscó la alternancia entre evaluadores y tareas.

La capacitación

El responsable del examen y el investigador organizaron la sesión de capacitación para los cuatro evaluadores novatos. A los nuevos evaluadores se les dio la versión afinada de la rúbrica junto con los dos textos elaborados por cuatro examinados, mismos que habían sido previamente seleccionados como ejemplos prototípicos de los distintos niveles de desempeño. Se les pidió a los evaluadores que estudiaran la rúbrica de forma independiente y, en caso de dudas, se explicaron y aclararon. Una vez que estaban familiarizados con la rúbrica, se les pidió que evaluaran los 8 textos (gráfica y ensayo argumentativo de cada sustentante) también de forma independiente. Posteriormente, mostraron sus resultados para cotejarlos con los puntajes previamente establecidos. En este momento, de manera individual, se les proporcionó retroalimentación y se clarificaron las dudas con respecto a la asignación de las distintas categorías de la rúbrica. Estas sesiones de retroalimentación se llevaron a cabo de forma presencial en un lapso de, aproximadamente, una o dos horas para cada evaluador. Concluida la sesión de retroalimentación, se les otorgaron los textos distribuidos en tres paquetes para que los evaluaran a lo largo de un mes. Cada vez que concluían con la evaluación de un paquete se conducían los análisis estadísticos y se les proporcionaba retroalimentación en caso de ser necesario.

El análisis estadístico

Para dar cuenta de la calidad del proceso de evaluación en las facetas mencionadas, se empleó el análisis estadístico MFRM en el programa Facets versión 3.71.4 (Linacre, 2015). Dado que Facets reporta numerosas tablas informativas, se seleccionaron exclusivamente las siguientes: mapa de variabilidad, tablas de la calidad de las facetas en cuestión (examinado, evaluador, tarea y rúbrica) y las tablas de sesgos entre los evaluadores y los examinados, tareas y la rúbrica.

Resultados

Debido a que el análisis estadístico de MFRM presenta diversas tablas informativas, primeramente, se presentará el mapa de variabilidad, el que

aporta información global de la evaluación. Posteriormente, se detallarán cada una de las cuatro facetas de variabilidad identificadas en el presente estudio (los examinados, las tareas, los evaluadores y las categorías de la rúbrica).

Mapa de variabilidad

La figura 1 presenta el mapa de variabilidad, que muestra un panorama general del proceso de evaluación, así como de la dispersión entre los distintos elementos de las facetas.

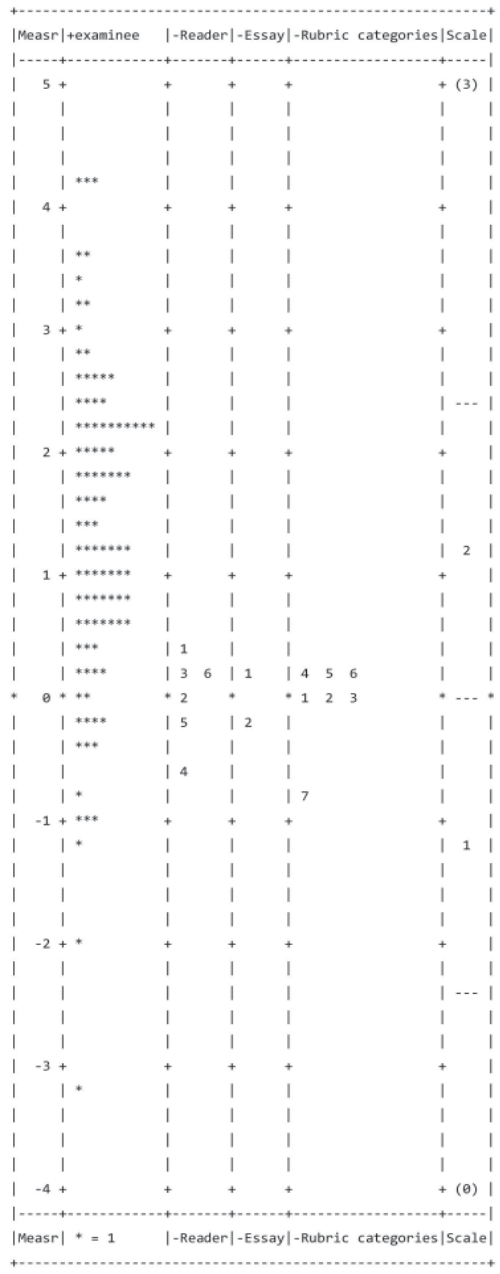
En la primera columna se presenta la escala de medición en lógitos. Esta medida parte de un cero establecido de manera convencional –punto medio de dificultad o severidad– y puede incrementar de forma positiva o negativa. La segunda columna muestra a los examinados. La mayoría se encuentra arriba del cero, lo cual demuestra que cuentan con una habilidad para escribir superior a la media. Esto se debe a que las muestras corresponden a estudiantes ya inscritos en alguna licenciatura o posgrado –la gran mayoría contaba con certificados probatorios de su nivel de español igual o superior al B2 (nivel intermedio alto)–. En la tercera columna se aprecian los evaluadores. Se puede observar cómo el evaluador 1 es el más severo y el 4 el más indulgente. En la cuarta columna se observa que las tareas 1 (descripción y contraste de gráficas) y 2 (redacción de un ensayo de tipo argumentativo) representan prácticamente la misma dificultad. En la quinta columna se observan las categorías de la rúbrica. Como se puede ver, todas las categorías presentan el mismo grado de dificultad, excepto la 7 (ortografía). En la última columna se muestran las bandas de cada una de las categorías de la rúbrica.

Faceta 1: los examinados

No obstante, el mapa de variabilidad no nos permite observar la calidad de cada una de las facetas y de sus elementos. Esta información, sin embargo, se encuentra contenida en las tablas que a continuación se reportan. La primera tabla informativa del proceso de calidad que arroja Facets es la de los examinados (tabla 2) y evidencia qué tan bien permite el examen discriminar la habilidad de los examinados. La media de la medición (M), en lógitos, fue de 1.30 y con una desviación estándar (SD) de 1.32. El valor de M=1.30 demuestra que las tareas fueron en general fáciles para los examinados.

FIGURA 1

Mapa de variabilidad de la prueba escrita del examen de español con fines académicos (EXELEAA)



Nomenclatura

Primera columna
Measr= medición, dada en lógitos.

Segunda columna
Examinee= examinados
* indica un candidato

Tercera columna
Reader= evaluador

Cuarta columna
Essay= texto
1 – descripción de gráfica
2 – redacción de ensayo argumentativo

Quinta columna
Rubric categories = categorías de la rúbrica
1 – Adecuación
2 – Desarrollo de ideas
3 – Cohesión
4 – Coherencia
5 – Léxico
6 – Gramática
7 – Ortografía

Sexta columna
Niveles de la rúbrica
(0 a 3 puntos)

Lo anterior es comprensible dado que la muestra se tomó de los estudiantes que ya se encontraban cursando estudios universitarios. El examinado con menor habilidad es el 47 con -3.25 lógitos y el de mayor habilidad es el 9 con 4.27 lógitos. El SE indica el error estándar; es decir, la precisión del estimado de la medición. Como se puede observar, SE=0.23 en el caso de la media del estimado de la habilidad de los examinados y SE=0.04 para su desviación estándar. La media del *Outfi*t de los examinados fue de 1.02, valor ligeramente mayor al esperado (1), y la desviación estándar fue de 0.31. Cuando el examinado presenta patrones de desajuste (*misfit*), esto puede significar dos cosas: o sus habilidades de escritura son disímiles, o los evaluadores presentan desacuerdo en la calidad del desempeño del examinado (Barkaoui, 2014:1306). En los estadísticos de separación se observa que el examen nos permite distinguir a los 100 estudiantes entre ocho niveles estadísticamente distintos en términos de sus habilidades de escritura. El índice alto de confiabilidad de separación (0.97) indica la probabilidad de que los examinados fueran ordenados de la misma manera si tomaran otro examen que evaluara el mismo constructo. El estadístico de la Chi-cuadrada reporta que el estimado de la habilidad de los examinados es estadísticamente significativo a p< .001.

TABLA 2

Resumen de estadísticos de la faceta de los examinados

Estimado de la habilidad de los examinados (n=100)	
M (modelo SE)	1.30 (.23)
SD (modelo SE)	1.32 (.04)
Min.	-3.25
Max.	4.27
Outfit	
M	1.02
SD	.31
Estadísticos de separación	
Estratos	7.7
Confiabilidad de separación	.97
Estadístico fijo de chi-cuadrada (grados de libertad)	2723.7 (99), p < .001

Faceta 2: Los evaluadores

La tabla 3 es la que proporciona información sobre los evaluadores, esta muestra el resumen de la calidad del proceso de corrección de los evaluadores. Sin embargo, la tabla 3 no solo permite evidenciar la severidad e indulgencia y consistencia (inter-evaluador e intra-evaluador), sino también arroja información sobre efectos indeseables en el proceso de evaluación; a saber: sesgos, efecto halo, tendencia central y restricción de rango (Barkaoui, 2014; Myford y Wolfe, 2004). Según Eckes (2005, 2009), el efecto halo y de tendencia central se pueden identificar cuando los evaluadores presentan valores de sobreajuste en *Infit* y *Outfit*.

TABLA 3

Reporte de medición de los evaluadores

Estimado de medición de los evaluadores					
	Medición	Modelo SE	Outfit	Correlación PtMea	Observaciones exactas %
Evaluador 1	.43	.05	.94	.69	48.4
Evaluador 2	.03	.05	.82	.71	49.6
Evaluador 3	.11	.06	.90	.71	50.0
Evaluador 4	-.61	.06	1.64	.37	38.1
Evaluador 5	-.20	.06	.99	.67	47.3
Evaluador 6	.24	.06	1.00	.59	45.4
Outfit					
M				1.05	
SD				.30	
Estadísticos de separación					
Estratos				8.85	
Confiabilidad de separación (no inter-evaluador)				.98	
Estadístico fijo de chi-cuadrada (grados de libertad)				198.9 (5), p < .001	
Acuerdo entre evaluadores					
Oportunidades de acuerdo entre evaluadores				8890	
Acuerdos exactos				4164 = 46.8%	
Acuerdos esperados				4023.8 = 45.3%	

Nuevamente, en el estimado de medición figuran los valores en lógitos. El evaluador más indulgente es el 4 con -0.61 lógitos, y el más severo es el 1 con .43 lógitos. La suma absoluta de la severidad de estos evaluadores es 1.04 lógitos. Si utilizamos el ejemplo de Eckes (2009: 280) para identificar si el rango de severidad es aceptable, observamos que la suma absoluta del evaluador más severo y el más indulgente no excede la cuarta parte del rango total de la habilidad de los examinados. El examinado con menor habilidad obtuvo un puntaje de -3.25 lógitos, y el de mayor habilidad, 4.27 lógitos. La suma absoluta da un total de 7.52 lógitos, cifra que si dividimos entre 4 nos da un total de 1.88 lógitos de rango aceptable entre los elementos de una faceta. Dado que la suma absoluta de la medición del evaluador más severo y el más indulgente es 1.04, este rango de severidad de los evaluadores se encuentra dentro de los límites aceptables < 1.88 lógitos. Dado que cada evaluador observó a los 100 examinados en una o en las dos tareas, la precisión de la estimación de los resultados obtenidos es bastante confiable (SE entre 0.05 y 0.06).

En cuanto al *Outfit*, el evaluador 4 se encuentra fuera de estos parámetros de calidad: *Outfit* = 1.64. Esto significa que el evaluador 4 es asistemático en sus evaluaciones y presenta dificultades para utilizar de forma consistente la rúbrica analítica. Los rangos de los demás evaluadores se encuentran dentro de los parámetros aceptables. En el ejemplo aquí mostrado, se observa que de los seis evaluadores, ninguno mostró efectos de tendencia central ni efecto halo, pues sus valores de *Infit* y *Outfit* son superiores a 0.8. También el evaluador 4 presenta correlaciones muy por debajo de las esperadas PtMea = 0.37. Esto significa que los puntajes asignados por este evaluador distan considerablemente de los demás correctores. Recordemos que el modelo parte del supuesto que los evaluadores son "evaluadores independientes" y no máquinas, razón por la cual no se esperan correlaciones perfectas, sino más bien consistencia en aquellas predichas por el modelo. En el porcentaje de observaciones exactas, nuevamente el valor más bajo lo presenta el evaluador 4 con 38.1%.

Como podemos observar en la tabla 4, el hecho de que un evaluador sea más severo o indulgente no significa que sea un mal evaluador. Por ejemplo, observamos que el juez 1 es el más severo de todos, dentro de un rango aceptable, pero que sus evaluaciones son consistentes, sistemáticas y su correlación con los demás evaluadores también es adecuada. En contraste, el 4 es el más indulgente, pero sus evaluaciones son deficientes,

poco confiables y con correlaciones bajas. La media del *Outfit* es de 1.05 y la desviación estándar SD = 0.30. En los estadísticos de separación, se observan nueve niveles distinguibles entre los evaluadores. La confiabilidad de esta separación es de 0.97, y el estadístico fijo de la Chi-cuadrada es significativo al $p < 0.001$. En el resumen del acuerdo entre evaluadores observamos que el promedio de los acuerdos observados (46.8%) superó a los esperados (45.3%).

Faceta 3: las tareas

La siguiente faceta de variabilidad es la tarea. En este caso, el tipo de texto: descripción y contraste de gráficos (tarea 1) y redacción de ensayo argumentativo (tarea 2) (véase Mendoza (2015) para una descripción detallada de las tareas). La tabla 4 muestra el reporte de medición de las dos tareas incluidas en la PE. Como se puede apreciar, la dificultad de ambas es muy similar (-.13 y .13). También hay un alto grado de precisión en la estimación dado que todos los examinados respondieron a ambas tareas (SE = 0.03). El índice de calidad *Outfit* también es adecuado para ambas tareas (.96 y 1.06, respectivamente) y la correlación entre ambas es adecuada (PtMea = 0.65).

TABLA 4

Reporte de medición de las tareas

Estimado de medición de las tareas	Medición	Modelo SE	Outfit	Correlación PtMea
Tarea 1	-.13	.03	.96	.65
Tarea 2	.13	.03	1.06	.65
Outfit				
M				1.05
SD				.07
Estadísticos de separación				
Estratos				8.01
Confiabilidad de separación				.97
Estadístico fijo de Chi-cuadrada (grados de libertad)				34.1 (1), $p < .001$

Faceta 4: la rúbrica analítica

La última faceta de interés es la rúbrica. Esta información se encuentra en el reporte de medición de la tabla 5. En relación con la dificultad de cada una de las categorías de la rúbrica, la más fácil fue la 7 (ortografía con -0.72 lógitos) y la más difícil fue la 6 (gramática con 0.24 lógitos). Entonces, el rango de medición absoluto es de .96 lógitos, lo cual si bien es aceptable según el ejemplo provisto por Eckes (2009), la distancia entre la ortografía y la adecuación es de 0.71 lógitos, lo que separa demasiado a la ortografía de las demás categorías. Este distanciamiento se puede apreciar mejor en el mapa de variabilidad (figura 1). La precisión de la medición es aceptable SE = 0.06, y el índice de calidad de la evaluación, *Outfit*, demuestra un ligero desajuste de la categoría 7: ortografía (*Oufit* = 1.29). Este desajuste también se puede apreciar ligeramente en las correlaciones o puntos biseriales (PtMea = .51). Sin embargo, todas las demás categorías presentan un óptimo índice de calidad y también de correlación entre categorías. La media del *Outfit* es M = 1.02 y la SD = 0.18. El valor de los estratos indica que se pueden identificar 8 niveles de dificultad (7.78) y la confiabilidad de la separación es de 0.97. El estadístico fijo de la chi-cuadrada 174.1 (6) es significativo al $p < 0.001$.

TABLA 5

Reporte de medición de la rúbrica analítica

Estimado de medición de las categorías de la rúbrica	Medición	Modelo SE	Outfit	Correlación PtMea
1 Adecuación	-.02	.06	.99	.67
2 Desarrollo de ideas	.02	.06	1.11	.65
3 Coherencia	.08	.06	1.08	.62
4 Cohesión	.23	.06	1.03	.65
5 Léxico	.16	.06	.72	.71
6 Gramática	.24	.06	.91	.67
7 Ortografía	-.72	.06	1.29	.51
Outfit				
M				1.02
SD				.18
Estadísticos de separación				
Estratos				7.72
Confiabilidad de separación				.97
Estadístico fijo de chi-cuadrada (grados de libertad)				174.1 (6), p < .001

La tabla 6 se relaciona con su funcionamiento de la rúbrica y muestra sus estadísticos. En la primera columna se muestra el puntaje de las cuatro bandas (0 a 3) incluidas en la rúbrica. La segunda se refiere a la frecuencia de uso y en la tercera al porcentaje de uso. Como se puede observar, la primera banda (puntaje 0) presenta una restricción del rango de uso, pues fue la menos utilizada (3%). Nuevamente, lo anterior se debe a que la muestra fue tomada de los estudiantes de grado y posgrado que ya se encontraban cursando estudios académicos. Esto se corrobora con la tercera banda (puntaje 2), puesto que fue la más utilizada (43%) en contraste con la segunda y cuarta banda, con un uso de 23% y de 34%, respectivamente; lo cual demuestra que la mayoría de los estudiantes contaban con un nivel de escritura aceptable para el ámbito académico.

TABLA 6
Funcionamiento de la rúbrica

Estadísticos de las categorías Datos			Control de calidad			Umbrales de Rasch-Andrich	
Puntaje	*Uso de la categoría*	*% de uso de la cat.*	*Medición promedio*	*Medición esperada*	*Outfit*	*Medición*	*SE*
0	180	3	-1.33	-1.12	.08		
1	1179	23	.27	.24	1.1	-2.27	.09
2	2262	43	1.30	1.28	1.0	.11	.04
3	1608	31	2.35	2.38	1.0	2.16	.04

En la cuarta columna, la medición promedio de la habilidad de los examinados se encuentra asociada con cada una de las cuatro bandas de la rúbrica (0 a 3 puntos). En esta columna también se aprecia un incremento monotónico de la medición promedio, lo cual significa que los estudiantes menos proficientes recibieron calificaciones más bajas, mientras quienes cuentan con una habilidad mayor recibieron puntuaciones más altas. De acuerdo con Bond y Fox (2007) y Linacre (2011), es imprescindible que exista un incremento monotónico en la medición promedio. Asimismo, se puede observar que esta medición es muy similar a la predicha por el modelo (quinta columna). De igual manera, en la sexta columna se reportan los valores del *Outfit,* mismos que son adecuados y cercanos a lo esperado

(1). La séptima columna presenta los niveles umbrales de Rasch-Andrich. Estos valores de medición son las dificultades estimadas de escoger una categoría de respuesta en relación con otra; es decir, la dificultad de elegir un 3 sobre un 2 (Barkaoui, 2014:1312). Los umbrales de Rasch-Andrich también presentan un incremento monotónico, lo cual demuestra que la rúbrica funciona adecuadamente para discriminar entre estudiantes menos proficientes de aquellos con mayor habilidad.

Reporte de sesgos

El programa Facets permite identificar los sesgos que se presentan entre las distintas facetas. En este caso, se ofrece el reporte de los sesgos identificados entre los evaluadores y las otras tres facetas (los examinados, las tareas y las categorías de las rúbricas). A diferencia de los índices de ajuste, la tabla de sesgos indica que los datos se ajustan al modelo de manera útil (Eckes, 2012:278). La tabla 7 refiere la interacción entre los examinados y los evaluadores. Solamente se muestra un ejemplo con los 10 sesgos negativos de mayor relevancia (significa que el evaluador fue más severo de lo esperado). En las primeras dos columnas encontramos el contraste entre el puntaje observado y el esperado. En la tercera vemos el tamaño del sesgo –este es el dato más importante y mediante el cual se jerarquizan los datos–. En la cuarta columna, figura la precisión de la estimación del modelo (SE del modelo). En la quinta tenemos la prueba estadística ***t*** de confirmación de la hipótesis. La ***t*** representa la hipótesis de que el sesgo sea una cuestión azarosa o no. Entre más se aleja del 0, más contundente es la probabilidad de que el sesgo no sea mera casualidad. De acuerdo con Linacre (2012b), cuando el valor de ***t*** es ≥ ± 2.0, el sesgo es estadísticamente significativo, y cuando el valor de ***t*** es ≥ ± 2.6, el sesgo es altamente significativo (Linacre, 2012b:13). En las últimas dos columnas figuran datos sobre el examinado y el evaluador que está presentando valoraciones asistemáticas no azarosas. Como se puede apreciar en la última columna el evaluador que presenta los sesgos de mayor tamaño es el 4.

La tabla 8 muestra la interacción entre los evaluadores y las tareas. Como se puede observar el único evaluador que presentó sesgos estadísticamente significativos fue el 6. Esto significa que puede tener una interpretación distinta de la dificultad de las dos tareas propuestas (ensayo argumentativo y descripción de gráficas). No obstante, el tamaño del sesgo es bajo en ambos casos (-0.30 y 0.21).

TABLA 7
Sesgos entre evaluadores y examinados

Reporte de interacciones y sesgos entre evaluadores y examinados

Puntaje observado	*Puntaje esperado*	*Tamaño del sesgo*	*Modelo SE*	*t*	*Examinado*	*Evaluador*
0	8.17	-4.59	1.75	-2.62	23	6
1	10.75	-4.27	1.07	-4.00	27	6
7	18.48	-4.03	.60	-6.76	8	4
9	18.78	-3.51	.58	-5.82	11	4
10	19.00	-3.31	.57	-4.75	2	4
4	12.97	-3.11	.65	-5.16	40	5
9	17.72	-2.97	.58	-4.55	20	4
10	17.60	-2.59	.57	-4.57	22	4
11	18.26	-2.58	.56	-4.42	7	4
12	18.68	-2.49	.56	-4.36	1	4

TABLA 8
Sesgos entre evaluadores y tareas

Reporte de interacciones y sesgos entre evaluadores y tareas

Puntaje observado	*Puntaje esperado*	*Tamaño del sesgo*	*Modelo SE*	*t*	*Tarea*	*Evaluador*
540	575.28	-0.30	.09	-3.29	1	6
868	832.71	0.21	.08	2.69	2	6

La tabla 9 muestra la interacción entre los evaluadores y las categorías de la rúbrica. En esta tabla se presentan únicamente los sesgos negativos que resultaron representativos, es decir, con valores de $t \geq -2.0$. Como se puede apreciar, los mayores sesgos se presentan con la categoría 7 (ortografía).

TABLA 9
Sesgos entre evaluadores y categorías de la rúbrica

Reporte de interacciones y sesgos entre evaluadores y categorías de la rúbrica

Puntaje observado	*Puntaje esperado*	*Tamaño del sesgo*	*Modelo SE*	*t*	*Cat. de la Rúbrica*	*Evaluador*
214	257.52	-1.25	.16	-7.83	7	4
205	229.70	-0.63	.16	-4.05	7	6
226	244.73	-0.53	.16	-3.22	7	5
188	207.42	-0.46	.16	-3.02	2	3
247	273.30	-0.42	.13	-3.33	5	1
191	208.71	-0.42	.15	-2.76	1	3
199	215.36	-0.40	.16	-2.60	3	5
333	361.39	-0.40	.12	-3.41	1	2
183	198.59	-0.37	.15	-2.40	4	3
252	268.86	-0.27	.13	-2.13	4	1

Discusiones

Como se pudo observar en los ejemplos, el análisis de MFRM es útil para examinar y evaluar el proceso de corrección de pruebas de desempeño. Este proceso se observa en el mapa de variabilidad, en las tablas de estimación de medición de las distintas facetas y también en los reportes de sesgos. El mapa de variabilidad, provee una visión panorámica de todo el proceso de evaluación, pues muestra las distintas facetas involucradas. Esta primera aproximación permite, a su vez, comparar la medición de cada uno de los elementos de las diversas fuentes de variabilidad. No obstante, esta visión preliminar de los datos no arroja información adicional sobre la calidad y la precisión de medición de cada una de las facetas. Por esta razón las tablas de estimación y de sesgos son indispensables para escudriñar el proceso de evaluación.

En relación con la faceta de los examinados, se pudo apreciar tanto en el mapa de variabilidad como en la tabla del estimado de su habilidad que la mayoría de los estudiantes contaba con una habilidad de escritura

superior a la media. Esto debido a que la muestra se tomó de estudiantes de grado y posgrado, por lo cual se esperaba que la mayoría escribiera adecuadamente en el contexto académico. Eventualmente, en estudios posteriores con muestras tomadas de estudiantes prospectos para cursar estudios universitarios, se esperaría que este patrón de comportamiento cambiara de acuerdo con el nivel de escritura de quienes aún no han sido aceptados para cursar dichos estudios.

Una de las facetas más importantes es la de los evaluadores. Dado que se encuentran en el centro del proceso de valoración es indispensable indagar respecto de la calidad de las evaluaciones que emiten, así como de su sistematicidad. Como se puede apreciar en el mapa de variabilidad, la medición de la severidad de cada uno de los evaluadores se encuentra dentro de un rango aceptable (menor a ± 1 lógito). Cabe mencionar que había cuatro novatos, por lo que se esperaba que hubiera ciertas discrepancias en sus valoraciones. No obstante, al examinar detalladamente la calidad de las evaluaciones emitidas, se observó que si bien la mayoría de los evaluadores fueron sistemáticos en los puntajes que otorgaron, el evaluador 4 demostró un grado de inconsistencia inadmisible (*Outfit* = 1.64), así como bajas correlaciones con el resto de los evaluadores. Linacre (2012b) sugiere que en estos casos se elimine a quienes resulten asistemáticos y se corra nuevamente el programa porque el modelo de Rasch depende de los evaluadores, mismo que se distorsiona cuando existen algunos inconsistentes. De ahí que el análisis de MFRM sea útil en dos sentidos. En primer lugar, ayuda a detectar evaluadores con severas inconsistencias, mismos que deberán ser remplazados por otros cuando, pese a la retroalimentación y formación, no mejoren en la calidad de sus evaluaciones. En segundo lugar, el análisis permite identificar si algún evaluador, aunque sistemático, presenta problemas puntuales al valorar a un examinado en particular. Esto permite ofrecer una retroalimentación a los evaluadores con respecto a su desempeño, pero también sirve para identificar textos que resultan problemáticos. Por ejemplo, en el reporte de sesgos entre los evaluadores y examinados (tabla 7), se aprecia que el número 6 tuvo dificultades para evaluar al examinado 23. En este sentido, se puede proveer retroalimentación puntual a cada uno con aquellos textos cuyas evaluaciones discrepan considerablemente con las del resto de los evaluadores.

La faceta de las tareas también ayuda a identificar qué tan difíciles son las que se les presentan a los examinados. En el caso del ejemplo, dado que cada una de las tareas aportaban un puntaje equivalente a la valoración global de la prueba escrita, se esperaba que la dificultad fuera muy similar. Contrariamente, en el caso de exámenes progresivos o diagnósticos, se espera que las tareas de escritura sean de dificultad disímil. Sin embargo, uno de los posibles problemas al diseñar tareas es que estas pueden ser percibidas consciente o inconscientemente como distintas en su grado de complejidad y dificultad. Esto se observa en la tabla 8, de sesgos entre evaluadores y tareas, en la cual se aprecia que el evaluador 6 fue sistemáticamente más severo al evaluar la tarea 2. La idiosincrasia de los jueces juega un papel importante al momento en que evalúan, razón por la cual las sesiones de estandarización y capacitación son indispensables para armonizar los criterios entre los correctores.

Asimismo, el análisis de MFRM permite observar el funcionamiento de una rúbrica y de cada una de las categorías incluidas en ella. En cuanto a la rúbrica analítica, la única categoría que se encontró ligeramente fuera de rango fue la 7 (ortografía); de hecho, su medición de severidad fue de -0.72 lógitos, seguida de la adecuación con -0.02 lógitos. Esto significa que la categoría no representa el mismo grado de dificultad que las demás. Esta discrepancia en el uso de la categoría 7 se puede apreciar también en la tabla de sesgos donde los más importantes se presentaron entre los evaluadores 4, 5 y 6 con la categoría 7, por lo cual se presupone que hace falta mayor detalle en la redacción de cada una de las bandas. Lo anterior se podría explicar de varias maneras. En primer lugar, pudiera ser que la descripción de cada una de las bandas no fuera lo suficientemente precisa. Otra posible interpretación podría ser que la categoría de ortografía estuviera evaluando, por lo menos en español, otra dimensión de la escritura. Adicional a la indulgencia con que se evaluó la ortografía, esta fue la única categoría con un ligero desajuste (*Outfit* = 1.24) y la única que desplegó una correlación ligeramente menor al predicha por el modelo (*PtMea* = 0.51). En suma, se puede apreciar que esta categoría aún requiere refinamiento y no ayudó en gran medida a discriminar entre los estudiantes menos proficientes de aquellos más aptos.

Si bien la información que arroja el análisis estadístico de MFRM es invaluable, este no permite comprender la razón por la cual un evaluador

es asistemático o presenta problemas de interpretación al valorar un texto o al emplear una rúbrica. Más aún, tampoco podemos saber qué hacen estos jueces para subsanar las dificultades al momento de asignar un puntaje. Por este motivo es importante llevar a cabo estudios cualitativos que permitan identificar el perfil de los evaluadores y también el proceso cognitivo que siguen al momento de evaluar. Esta es la razón por la cual el uso de las herramientas estadísticas y psicométricas como la que aquí se presenta debe ir de la mano con sesiones de capacitación, monitoreo y retroalimentación, fundamentales para contar con un proceso de evaluación de calidad.

Conclusiones

La validez interna de una prueba de desempeño generalmente se lleva a cabo examinando la calidad del proceso de evaluación. En el presente estudio, se consideraron cuatro facetas: los examinados, las tareas, los evaluadores y la rúbrica. En cuanto a la calidad del proceso de evaluación, los análisis estadísticos mediante MFRM demostraron que el examen discrimina entre los distintos niveles de habilidad de los examinados, la diferencia de la dificultad de las tareas era mínima, la segunda versión de la rúbrica demostró ser de mejor calidad que la primera –aunque todavía la categoría de la ortografía requería refinamiento– y los evaluadores también desplegaron un alto grado de consistencia y confiabilidad en sus resultados (excepto por uno que presentó inconsistencias). MFRM representa una herramienta de gran utilidad para conducir procesos de validación en pruebas de desempeño o ejecución en la cual intervienen evaluadores humanos.

Anexo 1

Extracto de rúbrica para evaluar las dos tareas que conforman la prueba de expresión escrita del examen de español con fines académicos

Competencia sociolingüística	/6
Competencia pragmático-discursiva	/18
Competencia lingüística	/18
Total	**/42**

COMPETENCIA SOCIOLINGÜÍSTICA

Adecuación: enfoque en el tema y uso de las convenciones formales		
Marca	Ptos	Descripción
EXCEPCIONAL	3	El discurso logra plenamente su **propósito comunicativo** Sin fallas en el uso de las **convenciones** formales **académicas**: registro formal, tratamiento impersonal, objetividad (evitar anécdotas y frases como **yo creo que...**) Se adecua al número de palabras solicitado (en un rango de 10% mayor o menor)
HÁBIL	2	El discurso **logra** su **propósito comunicativo**, pero presenta **alguna falla** en cuanto a: Inserción de anécdotas personales y frases como **yo creo que**…; cita textual de las palabras utilizadas en la consigna No se ajusta al número de palabras solicitado (en un rango de 20% mayor o menor)
EN DESARROLLO	1	El discurso **NO logra su propósito comunicativo** **Varias fallas** de adecuación: Inserción de anécdotas personales y frases como ***yo creo que***…; cita textual de las palabras utilizadas en la consigna No se ajusta al número de palabras (en un rango menor a 50%)
	0	El discurso **NO logra su propósito comunicativo** Texto **fuera de las convenciones formales:** registro informal (uso de coloquialismos, barbarismos, palabras truncadas, etc.), inconsistencia en el tratamiento (cambio de persona gramatical) Texto demasiado corto para ser evaluado

COMPETENCIA PRAGMÁTICO-DISCURSIVA

Desarrollo de ideas (tarea 1). Esquemas que caracterizan los tipos de texto y que conducen su interpretación y producción (superestructura)		
Marca	Ptos	Descripción
EXCEPCIONAL	3	La redacción **se centra en el tema** propuesto. Se desarrollan completamente **todos los núcleos temáticos** Describe los componentes de **la(s) gráfica(s)** (sin retomar literalmente las palabras de la consigna) **Contrasta** de manera general la información relevante Sintetiza la información **del resto de la tabla** Presenta una **conclusión** que resume y establece relación entre los datos

(CONTINÚA)

ANEXO 1/ CONTINUACIÓN

HÁBIL	2	Se desarrollan en su **mayoría los núcleos temáticos** El texto **carece de alguno** de los siguientes elementos: Descripción de los componentes de la tabla Contraste de información relevante Síntesis del resto de la tabla Conclusión
EN DESARROLLO	1	Se desarrollan **parcialmente algunos de los núcleos temáticos** El texto **carece de dos** de siguientes elementos: Descripción de los componentes de la tabla Contraste de información relevante Síntesis del resto de la tabla Conclusión Presenta alguna de las siguientes fallas: Falta de abstracción del significado principal por descripción demasiado atomizada de los elementos Inclusión de información que no viene en las gráficas Interpretación de los datos
	0	Se desarrolla aisladamente **solo alguno de los núcleos** temáticos o aborda una temática diferente a la solicitada Texto fallido por la acumulación de los siguientes elementos: Falta de abstracción del significado principal por descripción demasiado atomizada de los elementos Inclusión de información que no viene en las gráficas Interpretación de los datos

Desarrollo de ideas (tarea 2). Esquemas que caracterizan los tipos de texto y que conducen su interpretación y producción (superestructura)		
Marca	Ptos	Descripción
EXCEPCIONAL	3	La redacción **se centra en el tema** propuesto. Se desarrollan completamente **todos los núcleos temáticos** Hay una **introducción** que presente el marco temático Enuncia claramente la **tesis** Presenta al menos dos **argumentos** fundamentados Reconoce al menos un **contrargumento** (o menciona que no los hay, según su punto de vista) Hay una **conclusión** que sintetiza el planteamiento o reelabora la tesis Incluye alguna operación retórica: -Secuencia progresiva -Ejemplificación -Definición -Comparación o contraste -Cita de autoridades -Pregunta retórica, etc.

HÁBIL	2	Se desarrollan en su **mayoría los núcleos temáticos** El texto **carece de alguno** de los siguientes elementos: -Introducción -Tesis -Un argumento -Contrargumento -Conclusión -Operaciones
EN DESARROLLO	1	Se desarrollan **parcialmente algunos de los núcleos temáticos** El texto carece de **dos** de los siguientes elementos: -Introducción -Tesis -Un argumento -Contrargumento -Conclusión
	0	Se desarrolla aisladamente **solo alguno de los núcleos** temáticos o aborda una temática diferente a la solicitada El texto carece de estructura argumentativa

Referencias

Attali, Yigal; Lewis, Will y Steier, Michael (2012). "Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring", *Language Testing*, vol. 30, núm. 1, pp. 125-141.

Barkaoui, Khaled (2014). "Multifaceted Rasch Analysis for Test Evaluation", en A. Kunnan (ed.), *The Companion to Language assessment*, vol. 3, cap. 77, Oxford, Reino Unido: Wiley-Blackwell, pp. 1301-1322.

Bond, Trevor y Fox, Christine M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences,* 2ª ed., Mahwah, NJ: Erlbaum.

East, Martin (2009). "Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing", *Assessing Writing*, vol. 14, núm.2, pp. 88-115.

Eckes, Thomas (2005). "Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis", *Language Assessment Quarterly*, vol. 2, pp. 197-221.

Eckes, Thomas (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: Council of Europe/Language Policy Division. Disponible en: www.coe.int/t/dg4/linguistic/manuel1_EN.asp?#P19_2121

Eckes, Thomas (2012). "Operational rater types in writing assessment: Linking rater cognition to rater behavior", *Language Assessment Quarterly*, vol. 9, núm. 3, pp. 270-292.

Engelhard, George (2008). "Standard errors for performance standards based on bookmark judgments", *Rasch Measurement Transactions*, vol. 21, pp. 1132-1133.

Esfandiari, Rajab y Myford, Carol M. (2013). "Severity differences among self-assessors, peer-assessors, and teacher assessors rating EFL essays", *Assessing Writing*, vol. 18, núm. 2, pp. 111-131.

Hamp-Lyons, Liz (2007). "Worrying about rating", *Assessing Writing*, vol. 12, pp. 1-9.

Huang, Jinyan y Foote, Chandra J. (2010). "Grading between the lines: What really impacts professors' holistic evaluation of ESL graduate student writing?", *Language Assessment Quarterly*, vol. 7, núm. 1, pp. 37-41.

Knoch, Ute (2007). "Do empirically developed rating scales function differently to conventional rating scales for academic writing?", *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, vol. 5, pp. 1-36.

Knoch, Ute (2009). "Diagnostic assessment of writing: A comparison of two rating scales", *Language Testing*, vol. 26, núm. 2, pp. 275-304.

Linacre, John M. (2011). *A user's guide to facets: Rasch-Model Computer Programs*, Chicago: Winsteps.com.

Linacre, John M. (2012a). *Facets Tutorial 1. 1-32*. Disponible http://www.winsteps.com/a/ftutorial1.pdf

Linacre, John M. (2012b). *Facets Tutorial 2. 1-40*. Disponible en http://www.winsteps.com/a/ftutorial2.pdf

Linacre, John M. (2012c). *Facets Tutorial 3. 1-29*. Disponible en: http://www.winsteps.com/a/ftutorial3.pdf

Linacre, John M. (2012d). *Facets Tutorial 3. 1-18*. Disponible en: http://www.winsteps.com/a/ftutorial4.pdf

Linacre, John M. (2013). "A user's guide to facets. Rasch-Model Computer Programs", *Program Manual 3.71.0*. Disponible en: www.winsteps.com

Linacre, John. M. (2015). *Facets computer program for Many-facet Rasch Measurement, versión 3.71.4*, Beaverton: Winsteps.com

McNamara, Tim F. (1996). *Measuring second language performance*, Londres, Reino Unido: Longman.

Mendoza, Arturo (2015). "La selección de las tareas de escritura en los exámenes de lengua extranjera destinados al ámbito académico", *Revista Nebrija de Lingüística Aplicada a la Enseñanza de Lenguas*, núm. 18, pp. 106-123.

Mendoza, Arturo y Knoch, Ute (2018). "Examining the validity of the analytic rating scale for a Spanish test for academic purposes using the argument-based approach to validation", *Assessing Writing*, núm. 35, pp. 41-55.

Myford, Carol M. y Wolfe, Edward W. (2003). "Detecting and measuring rater effects using Many-Facet Rasch Measurement: Part I", *Journal of Applied Measurement*, vol. 4, pp. 386-422.

Myford, Carol M. y Wolfe, Edward W. (2004). "Detecting and measuring rater effects using Many-Facet Rasch Measurement: Part II", *Journal of Applied Measurement*, vol. 5, pp. 189-227.

Prieto, Gerardo (2011). "Evaluación de la ejecución mediante el modelo Many-Facet Rasch Measurement", *Psicothema*, vol. 23, núm. 2, pp. 233-238.

Prieto, Gerardo y Nieto, Eloísa (2014). "Analysis of rater severity on written expression exam using Many-Faceted Rasch Measurement", *Psicológica*, vol. 35, pp. 285-397.

Rasch, George (1960). *Probabilistic models for some intelligence and attainment tests,* Chicago: Mesa Press.

Rezaei, Ali R. y Lovorn, Michael (2010). "Reliability and validity of rubrics for assessment through writing", *Assessing Writing*, vol. 15, núm. 1, pp. 18-39.

Wind, Stefanie A. y Engelhard, George (2013). "How invariant and accurate are domain ratings in writing assessment?", *Assessing Writing*, vol. 18, núm. 4, pp. 278-299.

Wright, Benjamin D. y Linacre, John M. (1994). "Reasonable mean-square fit values", *Rasch Measurement Transactions*, vol. 8, núm. 3, p. 370. Disponible: https://www.rasch.org/rmt/rmt83b.htm

Artículo recibido: 22 de junio de 2017
Dictaminado: 19 de enero de 2018
Segunda versión: 20 de febrero de 2018
Aceptado: 22 de febrero de 2018