2018, Vol.23 No 1, pp. 12-27
http://www.unioviedo.es/reunido/index.php/Rema

Efectos del corrector en las evaluaciones educativas de alto impacto.
Rater effects in high-impact educational assessments.

Pamela Woitschach[(1)], Carlota Díaz-Pérez[(2)], Daniel Fernández-Argüelles[(2)], Jaime Fernández-Castañón[(2)], Alba Fernández-Castillo[(2)], Lara Fernández-Rodríguez[(2)], María Cristina González-Canal [(2)], Iris López-Marqués[(2)], David Martín-Espinosa [(2)], Rubén Navarro-Cabrero [(2)], Lara Osendi-Cadenas [(2)], Diego Riesgo-Fernández [(2)], Zara Suárez-García [(2)] y Rubén Fernández-Alonso [(2) (3)]

[(1)] Universidad Complutense de Madrid; [(2)] Universidad de Oviedo; [(3)] Consejería de Educación y Cultura del Principado de Asturias.

RESUMEN

Antecedentes: Los ítems de ejecución que son calificados por diferentes jueces mediante rúbricas es uno de los mayores desafíos en la evaluación educativa a gran escala y de alto impacto. Es conocido que los efectos del corrector afectan a los resultados de la evaluación. En este contexto, el presente estudio analiza la fiabilidad entre-correctores en la evaluación de la expresión escrita. **Método:** un grupo de 13 correctores calificaron 375 escritos de estudiantes de 6º curso, siguiendo una rúbrica analítica compuesta de 8 criterios de corrección. Los correctores se asignaron a 13 tribunales siguiendo un diseño de bloques incompletos balanceado. Los análisis realizados buscaron, en primer lugar, confirmar la estructura unidimensional de la rúbrica. A continuación se emplearon diferentes métodos clásicos para estudiar los efectos del corrector, la consistencia intra-juez y el acuerdo entre jueces. **Resultados**: se encontraron efectos diferenciales entre los correctores. Estas diferencias son importantes cuando se compara el grado de severidad de los jueces. También se encuentran diferencias en la consistencia interna de cada juez y en el acuerdo entre correctores. Este último efecto es especialmente significativo en algunos tribunales. **Conclusiones**: las diferencias entre correctores pueden tener diferentes fuentes, como son la experiencia, familiaridad con la tarea y grado de entrenamiento con las rúbricas; la naturaleza de la tarea a evaluar o el propio diseño de la rúbrica empleada.

Palabras clave: efectos del corrector, fiabilidad entre jueces, rúbricas de evaluación; evaluación de la escritura, evaluar ejecuciones

ABSTRACT

Antecedents: the constructed response test items that are qualified by different correctors with rubrics are one of the biggest challenges for Rater effects in high-impact educational assessments and are applied to large sample groups. It is known that rater bias affects the results of the evaluation. In this context, the present study analyzes the raters' effects of the corrector on written expression. **Method**: a group of 13 raters rated 375 written productions of 6th-grade students, following an analytical rubric composed of 8 correction criteria. The raters were assigned to 13 groups of correction following a balanced incomplete block design. The first step of the analysis carried out was to confirm the one-dimensional structure of the rubric. The next and final step used different classical methods to study the raters' effects, the intra-rater consistency and the agreement between judges. **Results:** differential effects were found among the raters. These differences are important when the raters' severity is compared. There are also differences in the internal consistency of each judge and in the agreement between correctors. This last effect is especially significant in some raters' groups. **Discussion:** differences between raters may have different sources, such as experience and familiarity with the task; the degree of training with the rubrics; the nature of the test; and the design of the rubric used.

Keywords: rater effects; inter-rater reliability; scoring rubrics, assessing writing, performance ratings.

Contacto

Rubén Fernández-Alonso
Facultad de Educación y Formación del Profesorado
Universidad de Oviedo
33007 Oviedo (Spain)
e-mail: fernandezaruben@uniovi.es

Agradecimientos

Los autores expresan su mayor gratitud a la Consejería de Educación y Cultura del Gobierno del Principado de Asturias, sin cuya colaboración esta investigación no hubiese sido posible.

1.- Introducción

Los sistemas educativos prevén momentos de transición o finalización de estudios jalonados por exámenes de titulación, certificación, acceso, admisión y concesión de premios. Además, los resultados de estas pruebas o estudios similares aplicados sobre grandes muestras sirven para valorar la calidad de la oferta educativa de los centros y ubicar sus resultados en el conjunto del sistema educativo (European Commission/EACEA/Eurydice, 2009, 2014, 2016). En ambos casos se trata de evaluaciones que impactan, bien en el futuro académico y personal del alumnado, bien en el juicio sobre el funcionamiento de los centros.

Por otra parte, en el último cuarto del siglo XX surge una corriente educativa que demanda, que los exámenes de alto impacto sean evaluaciones auténticas y que empleen formatos de prueba al margen del ítem cerrado: ensayos escritos, exposiciones orales, ejecuciones artísticas o físicas, resolución de casos, elaboración de informes, presentaciones, portafolios o proyectos (Bravo-Arteaga, & Fernández del Valle, 2000). Pese a su sustantividad, estos formatos alternativos no están exentos de la exigencia de equidad y justicia inherente al proceso de evaluación, por lo que calificar estas ejecuciones es uno de los mayores retos a los que se enfrentan las evaluaciones educativas de impacto académico y social. Es evidente que las puntuaciones otorgadas a los estudiantes deben reflejar su verdadera destreza en el dominio evaluado, sin embargo, existe abundante evidencia que señala que en las pruebas de ejecución no siempre se cumple el axioma básico de la evaluación objetiva y justa: la mejor calificación, para el mejor preparado (Engelhard, 1992, 1994, 1996; Gyagenda & Engelhard, 2009; Leckie & Baird, 2011; Lunz & Stahl, 1990; Lunz, Wright, & Linacre, 1990; Park, 2010; Wang, & Yao, 2013; Wolfe, 2004), lo que lleva a concluir a Congdon y McQueen (2000) que las limitaciones en términos de medición objetiva de este tipo de pruebas pueden convertir una evaluación de gran impacto en una lotería.

Los desafíos y problemas mencionados han hecho que, en las últimas décadas, la investigación sobre la fiabilidad y efectos del corrector se haya desarrollado fuertemente. Su estudio ha sido abordado desde diferentes aproximaciones metodológicas, como la Teoría Clásica de los Tests (Barrett, 2001; Gyagenda, & Engelhard, 2009; Saal, Downey & Lahey, 1980; Stemler, 2004), la Teoría de Respuesta al Ítem (Adams & Wu, 2010; Eckes, 2009; Lunz et al., 1990; Prieto, 2011; Wolfe & McVay, 2012), la Teoría de la Generalizabilidad (Sudweeks, Reeve, & Bradshaw, 2005) o los modelos Jerárquico-Lineales (Leckie, & Baird, 2011).

La presente investigación se desarrolla en el marco del análisis clásico. Saal et al. (1980), en una de las primeras revisiones sobre la calidad del proceso de corrección, señalaron las posibles fuentes de error de los correctores: severidad-permisividad, efecto halo, tendencia central y restricción del rango, ofreciendo hasta 11 definiciones operativas dentro del enfoque clásico para evaluar la calidad de la corrección: descriptivos básicos de posición, dispersión y distribución de frecuencias, índices de correlación y análisis de varianza.

Por su parte, Jonsson y Svingby (2007), en su estudio de revisión, indican que la fiabilidad de las calificaciones de los correctores se analizan desde una doble perspectiva: la fiabilidad del corrector (*intra-rater reliability*), que sería la estabilidad del corrector individualmente considerado a la hora de calificar diferentes producciones, y la fiabilidad entre correctores (*inter-rater reliability*), que sería el grado de acuerdo entre los correctores que califican la misma producción. Los métodos empleados tradicionalmente para explorar la consistencia interna de cada corrector son, por frecuencia de uso, el alfa de Cronbach y los índices de correlación Spearman y Pearson

(Barrett, 2001; Gwet, 2014; Jonsson & Svingby, 2007). Por su parte, los métodos basados en el acuerdo entre jueces más empleados serían los porcentajes de consenso y discrepancia entre jueces, la kappa de Cohen, los índices de consistencia y correlación ya mencionados, los coeficientes de correlación intraclase derivados de algún modelo del análisis de varianza o los diseños de medidas repetidas (Suárez-Álvarez, González-Prieto, Fernández-Alonso, Gil & Muñiz, 2014). Los índices de acuerdo entre jueces tratan de demostrar la existencia de la concordancia entre correctores como paso previo a resumir en un único promedio las calificaciones de jueces independientes (Barrett, 2001; Gwet, 2014; Hallgren, 2012; OECD, 2014; Stemler, 2004). Por su parte, los procedimientos de análisis de consistencia intra-juez tienen asunciones más débiles que los índices de acuerdo entre-jueces, ya que un determinado juez puede mostrar un uso consistente de la rúbrica, sin que ello garantice una corrección exacta, pudiendo encontrarse alta consistencia intra-juez sin que por ello deba existir acuerdo entre-jueces (Eckes, 2009; Stemler, 2004).

Por otro lado, en la última década en España, al amparo de la Ley Orgánica 2/2006 de Educación, se han generalizado las evaluaciones educativas sobre grandes muestras que incluyen pruebas destinadas a evaluar la expresión escrita del alumnado. En base a estos antecedentes el presente trabajo se plantea los siguientes objetivos:

- Estudiar la dimensionalidad de la rúbrica empleada para calificar las expresiones escritas en la *Evaluación Final de Educación Primaria*.
- Estudiar las fuentes de error y consistencia interna de los correctores con el fin de identificar posibles efectos del corrector en las puntuaciones obtenidas por el alumnado.
- Analizar la fiabilidad entre jueces, es decir, el grado de acuerdo entre las puntuaciones de los correctores de un mismo tribunal.

2.- Método

2.1.- Participantes

Las expresiones escritas fueron seleccionadas entre las producciones realizadas en la *Evaluación Final de Educación Primaria del Principado de Asturias,* un estudio censal para el alumnado matriculado en 6º de educación primaria en el curso 2015/16 donde se recogieron 6653 ejercicios provenientes de 403 grupos aula. Para elegir los ejercicios que formarían parte del estudio corrección se determinó que las unidades primarias de muestra fueran los grupos aula, es decir, se seleccionaron grupos-aula y se revisaron todas las producciones realizadas por el alumnado del grupo elegido. Esta selección estuvo presidida por los siguientes condicionantes: (a) se disponía de 13 correctores; (b) cada corrector revisaría unas 100 expresiones escritas; (c) cada ejercicio sería calificado por 4 correctores; (d) para no trabajar con grupos aula muy pequeños se eliminaron del marco muestral las aulas con menos de 10 estudiantes; (e) la distribución de los ejercicios escritos a los correctores debía ajustarse a un diseño matricial que permitiera controlar experimentalmente el efecto de asignación de los ejercicios a los correctores.

Con estos condicionantes el marco muestral quedó conformado finalmente por 5739 estudiantes escolarizados en 301 grupos-aula, y se estimó que sería necesario seleccionar 18 grupos-aula que, en conjunto, aportarían cerca de 400 expresiones escritas. La selección de las aulas se realizó siguiendo un muestreo sistemático, probabilístico y aleatorio donde los aulas fueron seleccionadas con una probabilidad proporcional a su tamaño (LaRoche, Joncas, & Foy, 2016). Finalmente, las aulas

seleccionadas aportaron un total de 375 expresiones escritas que pasaron a formar parte del estudio.

2.2.- Procedimiento

Inicialmente se gestionó el permiso para que la Consejería de Educación y Cultura del Principado de Asturias hiciera una cesión parcial, con fines de estudio e investigación, del fichero que contenía las producciones escritas de las aulas seleccionadas de acuerdo con lo establecido en la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal, lo que garantiza la privacidad y anonimato de los datos manejados. Las producciones escritas, que originalmente se encontraban en formato lápiz y papel, fueron escaneadas; anonimizadas mediante un código alfanumérico; editadas con el programa Adobe Acrobat Prof DC® para eliminar cualquier señal de la corrección original; y guardadas en un fichero nombrado con el código de identificación. Adicionalmente se prepararon 13 plantillas para la corrección (una por corrector), que contenían la relación de las expresiones escritas asignadas a cada corrector ordenadas en el orden en que debían ser corregidas y los criterios de corrección a emplear.

Por otra parte se diseñó una plataforma *online* donde se cargaron los 375 ficheros con las expresiones escritas, y en la que cada corrector solo tenía acceso a los ficheros de las expresiones escritas asignadas. Los correctores calificaron y codificaron en la plataforma los resultados de su corrección. Una vez descargadas las correcciones se realizó un chequeo para verificar que los códigos introducidos en la plataforma electrónica correspondían con las calificaciones contenidas en las plantillas para la corrección y, de ese modo, descartar errores en la codificación de las respuestas en la plataforma.

Los correctores, estudiantes voluntarios del Máster de Investigación e Innovación de Educación Infantil y Primaria de la Universidad de Oviedo, recibieron seis horas de formación: dos horas de introducción a la corrección mediante rúbricas y cuatro de trabajo con el instrumental de la prueba, que incluían entrenamientos de corrección con producciones similares que no se emplearían en el estudio pero que sirvieron para familiarizarse con el material y el procedimiento de corrección.

2.3.- Constitución de tribunales: distribución de los correctores a las producciones escritas

Para garantizar el control experimental a la hora de asignar los jueces a las producciones escritas, los correctores fueron organizados de acuerdo a un diseño matricial denominado cuadrado Youden de 13 bloques, un cuadrado latino incompleto y balanceado que desarrolla las 4 primeras réplicas del cuadrado latino 13 x 13 (Cochran & Cox, 1974). Para lograr el arreglo Youden de 13 bloques, las expresiones escritas de las 18 aulas se agruparon en 13 tribunales de calificación (o bloques en términos de diseño experimental). La tabla 1 muestra los 13 tribunales y el número de expresiones escritas asignadas a cada uno. De promedio a cada tribunal le correspondieron 28 expresiones escritas, si bien en los tribunales TB09 a TB13 se corrigieron más ejercicios ya que incluían las producciones de dos grupos-aula.

Tribunales (Bloques)	Aula(s) asignadas al tribunal	Nº expresiones por tribunal	Correctores que conformarán cada tribunal			
TB01	Aula 01	28	C01	C09	C12	C10
TB02	Aula 02	26	C02	C01	C08	C11
TB03	Aula 03	25	C03	C02	C09	C05
TB04	Aula 04	25	C04	C03	C01	C06
TB05	Aula 05	24	C05	C11	C04	C12
TB06	Aula 06	23	C06	C05	C10	C08
TB07	Aula 07	23	C07	C13	C05	C01
TB08	Aula 08	22	C08	C12	C13	C03
TB09	Aulas 09 y 18	34	C09	C08	C07	C04
TB10	Aulas 10 y 17	36	C10	C04	C02	C13
TB11	Aulas 11 y 16	36	C11	C10	C03	C07
TB12	Aulas 12 y 15	36	C12	C07	C06	C02
TB13	Aulas 13 y 14	37	C13	C06	C11	C09

Tabla 1. Composición de los 13 tribunales: aulas asignadas, número de expresiones escritas a corregir y correctores que conforman cada tribunal

Las cuatro últimas columnas de la tabla recogen la identificación del cuarteto de correctores asignados a cada tribunal y permiten comprobar la consistencia, balanceo y eficiencia del arreglo Youden. La consistencia del diseño hace que cada corrector sea asignado a cuatro tribunales distintos y que cada expresión escrita será corregida por cuatro correctores diferentes. El diseño está completamente balanceado ya que cada corrector coincidirá una vez, y sólo una vez, con el resto de los correctores a la hora de calificar un bloque o grupo de expresiones escritas. La eficiencia del diseño radica en que, de promedio, cada corrector revisará poco más de 100 expresiones escritas y no las 375 que serían necesarias si se hubiera empleado el cuadrado latino completo. En el conjunto del estudio se realizarán 1500 correcciones y no las casi 5000 que obligaría el diseño completo, ahorrando aproximadamente el 60% de las correcciones posibles sin que las conclusiones del análisis pierdan validez. En definitiva, esta distribución de los correctores a las expresiones escritas asegura el doble control propio del cuadrado latino (Fernández-Alonso & Muñiz, 2011) de tal modo que la posible severidad o benevolencia de los jueces no puede ser imputada al hecho de que le fuera asignado un grupo de alta o baja competencia en la expresión escrita.

2.4.- Instrumentos

El estímulo empleado era una lectura que contenía información turística de tres ciudades. El alumnado debía elegir una de ellas y redactar un texto argumentativo para convencer al resto de compañeros de que dicha ciudad era el mejor destino para un hipotético viaje de estudios[1]. La rúbrica de corrección evaluaba tres procesos cognitivos, que a su vez se dividen en ocho estándares de aprendizaje recogidos en el currículo establecido por la normativa que desarrolla la Ley Orgánica 8/2013 de Mejora de la Calidad Educativa. La tabla 2 muestra la organización de los procesos y estándares evaluados, señalando los estándares 1 a 5 se puntuaron en una escala de tres niveles (0 = No logrado; 1 = Parcialmente logrado; 2 = Logrado) y los estándares 6 a 8 se calificaron binariamente. La puntuación total de cada corrector se calculó como la suma de las puntuaciones en los ocho estándares.

[1] La prueba junto con sus especificaciones y rúbricas de corrección están disponibles en Ministerio de Educación, Cultura y Deporte (2016): *Pruebas de la evaluación final de Educación Primaria. Curso 2015-2016*. Madrid: Instituto de Evaluación (p. 97 -117), y se puede consultar desde el siguiente enlace: https://sede.educacion.gob.es/publiventa/descarga.action?f_codigo_agc=18314

Procesos cognitivos	Estándares
Coherencia	1. Organiza las ideas con claridad y progresión temática [(p)] 2. Expresa opiniones, reflexiones y valoraciones con coherencia [(p)]
Cohesión	3. Usa conectores básicos para dar cohesión al texto [(p)] 4. Usa signos de puntuación [(p)] 5. Usa sustituciones pronominales y sinónimos para evitar reiteraciones [(p)]
Adecuación y presentación	6. Respeta las normas gramaticales y ortográficas [(d)] 7. Usa un registro adecuado al interlocutor y asunto tratado [(d)] 8. Limpieza, claridad, precisión y orden del escrito [(d)]

[(p)] Estándar de aprendizaje puntuado politómicamente: Códigos 0, 1 y 2
[(d)] Estándar de aprendizaje puntuado dicotómicamente: Códigos 0 y 1

Tabla 2. Procesos cognitivos evaluados y estándares de la rúbrica de corrección

2.5.- Análisis de datos

Para el primer objetivo, estudiar la dimensionalidad de la rúbrica, se empleó la matriz donde los estándares hacen de variables (N = 1500). La muestra total, se dividió aleatoriamente en tres submuestras (N_1 = 516; N_2 = 477; N_3 = 507). Con la primera submuestra se realizó un análisis de factorial exploratorio (AFE), empleando como método de extracción el de máxima verosimilitud robusta. Para determinar el número óptimo de factores a retener se utilizó la implementación óptima del análisis paralelo de Horn propuesta por Timmerman & Lorenzo-Seva (2011) con 10.000 remuestreos aleatorios para el programa FACTOR 10.5.03. Con la segunda submuestra se realizó un análisis factorial confirmatorio (AFC) y para mejorar el ajuste se tuvieron en cuenta los índices de modificación, manteniéndolos constantes en la tercera de las muestras (Byrne, 2001; Abad, Olea, Ponsoda & García, 2011). Con la tercera submuestra se realizó otro AFC, aunque sin modificar el modelo propuesto en la primera con la intención de hacer una validación cruzada (Pérez-Gil, Chacón Moscoso y Moreno Rodríguez, 2000). Puesto que no se pudo probar la normalidad multivariada de la matriz de datos, y para mantener coherencia con el AFE, el método de extracción empleado en los dos AFC fue el de máxima verosimilitud robusta. La evaluación de la bondad de ajuste se determinó mediante el índice de ajuste comparativo (CFI), la razón entre χ^2 y los grados de libertad del modelo (χ^2 /g.l.) y la media cuadrática estandarizada de los residuales (SRMR). Los AFC se realizaron con el programa MPlus 7.

Para dar cuenta del segundo objetivo, identificación de los sesgos del corrector, también se usó la matriz de estándares como variables. En función del efecto a estudiar se realizaron diferentes análisis. El grado de severidad o permisividad se examinó mediante los estadísticos de posición y la comparación de medias por juez. Para identificar los efectos de restricción del rango y centralidad se estudiaron los estadísticos de dispersión y la distribución de frecuencias por corrector (Saal et al., 1980). En la comprobación del efecto halo se revisaron las correlaciones entre los estándares de corrección (Feeley, 2002). Finalmente, la consistencia intra-juez se estimó mediante dos índices: el alfa de Cronbach y el coeficiente de correlación intraclase (CCI) del modelo ANOVA de dos factores de efectos mixtos. La elección del CCI se realiza de acuerdo con el protocolo establecido por Koo y Li (2016) asumiendo que, mientras los sujetos se asignaron aleatoriamente a los tribunales, los estándares no son elementos aleatorios, sino que fueron seleccionados precisamente por su sustantividad para evaluar la expresión escrita.

Para estimar concordancia entre jueces se empleó la matriz donde los correctores se organizan como variables (N = 375), es decir, la matriz contiene cuatro columnas que se corresponden con las puntuaciones totales que cada estudiante recibió de su cuarteto de correctores. El grado de consenso entre correctores se analizó mediante las correlaciones de las puntuaciones totales. Como indicador de la fiabilidad interjueces y

de la relevancia de las correlaciones anteriormente obtenidas (esto es, el tamaño del efecto) se obtuvo el CCI por tribunal según un modelo ANOVA de dos efectos aleatorios, empleando como tipo de análisis de homogeneidad el acuerdo absoluto entre múltiples correctores (Koo & Li, 2016; McGraw & Wong, 1996; Shrout & Fleiss, 1979). Las razones de utilizar el CCI en vez del coeficiente kappa son dos: su mejor adecuación al nivel métrico de las variables empleadas y que, en cualquier caso, se trata de índices comparables (Abad et al., 2011). Por otro lado, para estudiar las diferencias de las puntuaciones totales dentro de cada tribunal se realizó un ANOVA de medidas repetidas –un factor intrasujeto– ya que cada alumno dispone de cuatro puntuaciones correspondientes a su cuarteto de evaluadores.

3.- Resultados

3.1.- Dimensionalidad del constructo

La matriz de datos es adecuada para ser sometida a una reducción factorial: el valor del test KMO fue .83 y la prueba de esfericidad de Bartlett fue estadísticamente significativa (χ^2 = 836.5; 28 gl; p < .001). El AFE mostró que la estructura óptima era unidimensional, explicando el factor el 38.2% de la varianza total. Por su parte, los dos AFC verifican el ajuste de los datos al modelo propuesto. La tabla 3 recoge los valores de ajuste de los tres análisis.

	χ^2	g.l	χ^2 /g.l.	RMSEA	CFI	N
AFE	95.45	20	4,77	0.07	0.98	516
AFC_1	54.41	19	2,86	0.06	0.95	477
AFC_2	94.75	19	4.99	0.08	0.91	507

Tabla 3. Valores de ajuste de los análisis exploratorio y confirmatorios de la unidimensionalidad de la rúbrica de evaluación

3.2.- Efectos y consistencia de los correctores

La tabla 4 resume los estadísticos fundamentales para analizar los efectos del corrector. En general la prueba resultó fácil (Media Total = 8.8 puntos, máximo 13), si bien existen importantes diferencias entre correctores: C07 parece mucho más severo que el resto, siendo el único juez que presenta una asimetría positiva. Por otra parte sólo un tercio de los correctores han cubierto el 100% del rango total de puntuaciones, mientras que C05 y C08 han compactado todas las puntuaciones en poco menos del 60% del rango. Como no podría ser otro modo existe una correlación alta y positiva (en torno a 0.70) entre la varianza y el porcentaje de rango cubierto por cada corrector. Por su parte el promedio de las correlaciones entre los criterios de corrección no indican efecto halo en los correctores, si bien el promedio en de las correlaciones en el C11 duplica el total general. Los valores ofrecidos por el alfa de Cronbach señalan un corrector con un nivel de consistencia excelente (α > .90); cinco en el rango de alta consistencia (α entre .75 y .90), cinco en los límites de la consistencia moderada (α entre .50 y .75) y dos casos de consistencia débil (α < 50). Los índices del CCI por corrector son sensiblemente inferiores y corrigen a la baja el diagnóstico que se acaba de realizar: ningún corrector alcanza la marca de excelencia y la gran mayoría se mueve en los márgenes de la consistencia moderada.

Los gráficos profundizan sobre los efectos de severidad y restricción del rango de los correctores. El gráfico 1 muestra la media y su intervalo de confianza por corrector (IC = 95%). La línea horizontal representa la media general y su anchura el rango de la verdadera media. Hay siete correctores centrados sobre la media, tres que

otorgan puntuaciones por debajo de la media (C07, C03 y C04) y, en el extremo contrario, otra terna (C10, C08 y C05) que se mostró más permisiva que el promedio.

El gráfico 2 representa la distribución de puntuaciones por corrector. En general se observan rangos de puntuación estrechos. En todos los casos, salvo C11, la distancia intercuartílica es de 3 o 4 puntos y la mayoría de los jueces ubican el 90% de las puntuaciones en un rango de entre 5 y 7 puntos. En el caso de los jueces más benévolos podría pensarse que la estrechez del rango se debe a que las mejores puntuaciones chocan contra el límite superior de la escala, sin embargo la restricción del rango también se observa en correctores más severos.

Juez	Media	DT	% Rango cubierto	Asimetría	Curtosis	Halo ($r_{Cx,Cy}$)	α	$CCI_{(3,k)}$
C01[(a)]	9.8	3.2	100%	-1.6	2.3	.40	.82	.77
C02	8.1	2.8	93%	-0.2	-0.5	.17	.65	.59
C03	7.8	2.0	86%	-0.1	-0.1	.10	.49	.41
C04	7.4	2.4	93%	-1.1	1.5	.33	.78	.74
C05	10.9	1.9	57%	-0.7	-0.6	.22	.69	.52
C06	8.8	2.6	86%	-0.3	-0.3	.16	.60	.57
C07	4.5	2.9	100%	0.8	0.1	.31	.80	.77
C08	10.4	1.7	57%	-0.4	-0.4	.10	.47	.37
C09	9.1	2.5	100%	-0.4	0.2	.22	.69	.64
C10[(a)]	10.5	2.2	93%	-1.1	1.7	.29	.75	.65
C11	9.5	3.8	100%	-1.0	0.2	.61	.92	.86
C12	8.9	2.6	79%	-0.5	-0.5	.26	.76	.70
C13	9.3	2.6	86%	-0.5	-0.1	.23	.65	.53
Total[(a)]	8.8	3.1	100%	-0.7	-0.0	.29	.77	.62

[(a)] La significación del Test de Kolmogorov-Smirnov (p > ,05) señala que la distribución de las puntuaciones de estos jueces no se ajusta a normalidad. En estos casos el CCI debe interpretarse cuidadosamente
DT: Desviación típica; $r_{Cx,Cy}$ promedio de las correlaciones entre cada para de estándar; α: Alfa de Cronbach; $CCI_{(3,k)}$: Coeficiente de Correlación Intraclase, ANOVA de dos factores de efectos mixtos

Tabla 4. Efectos de corrector

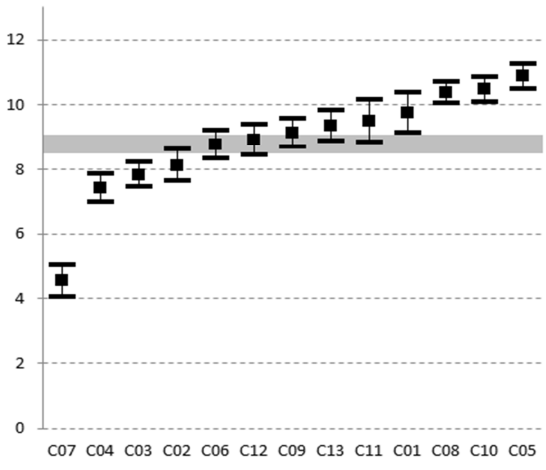

Gráfico 1. Barras de error por corrector: media y error típico de la media

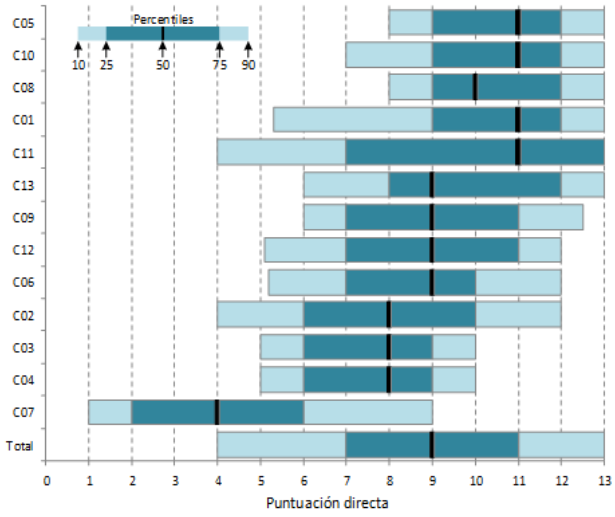

Gráfico 2. Distribución de puntuaciones por corrector

3.3.- Acuerdo entre correctores

Las tres primeras columnas de la tabla 5 muestran otros tantos índices por tribunal: el promedio del coeficiente de correlación entre las puntuaciones totales, el alfa de Cronbach y el CCI de acuerdo absoluto basado en el ANOVA de dos vías de efectos aleatorios. La media de las correlaciones para el conjunto de pares de correctores es .49, si bien existen importantes diferencias ya que el rango oscila entre .13 (TB08) y .70 (TB03). Por su parte, la mayoría de los tribunales presentan coeficientes alfa en torno a .75 o superiores. No obstante el CCI de acuerdo absoluto registra un descenso de 0.12 puntos en el conjunto de los tribunales, y hace que sólo tres presenten un valor superior a .75. Además, atendiendo a la magnitud del CCI se puede afirmar que en los tribunales TB07 y TB08 el grado de acuerdo es muy bajo, y en los tribunales TB09 y TB06, el acuerdo es entre débil y moderado en el mejor de los casos.

	Estadísticos de acuerdo básicos			ANOVA de medidas repetidas			
Tribunal (N casos)	r promedio	α	$CCI_{(2,k)}$	F	Sig.	Tamaño del efecto	Potencia
TB01 (N = 28)	.55	.83	.81	5.60	.002	.172	.934
TB02 (N = 26)	.36	.66	.64	3.73	.015	.130	.789
TB03 (N = 25)	.70	.89	.81	24.69	.000	.507	1
TB04 (N = 25)	.64	.88	.80	19.12	.000	.443	1
TB05 (N = 24)	.69	.89	.73	52.16	.000	.694	1
TB06 (N = 23)	.33	.69	.54	20.82	.000	.486	1
TB07 (N = 23)	.31	.55	.37	26.2*	.000	.544	1
TB08 (N = 22)	.13	.35	.33	2.38*	.078	.102	.476
TB09 (N = 34)	.53	.79	.51	94.28	.000	.741	1
TB10 (N = 36)	.44	.76	.69	16.08	.000	.315	1
TB11 (N = 36)	.60	.82	.60	72.53*	.000	.675	1
TB12 (N = 36)	.63	.87	.70	66.66	.000	.656	1
TB13 (N = 37)	.44	.73	.72	3.83*	.022	.096	.710
Promedio	.49	.75	.63				

r: correlaciones de las puntuaciones totales entre correctores del mismo tribunal
α: coeficiente de correlación intraclase de consistencia
$CCI_{(2.k)}$:coeficiente de correlación intraclase ANOVA dos vías de efectos aleatorios y acuerdo absoluto
*: En estos tribunales la prueba de Mauchly no confirma la esfericidad de la matriz de resultados. por lo que los datos que se presentan corresponden al contraste Greenhouse-Geisser

Tabla 5. Fiabilidad entre jueces y ANOVA de medidas repetidas por tribunal

En la segunda parte de la tabla 5 se recogen los resultados del ANOVA de medidas repetidas (un factor intrasujeto). En todos los tribunales, a excepción de TB08, existen diferencias estadísticamente significativas al nivel de confianza del 95%, mientras que en otros dos (TB02 y TB13) estas diferencias desaparecerían si se manejara el nivel de confianza del 99%. Atendiendo al tamaño del efecto (Cohen, 1988) habría cuatro tribunales donde estas diferencias podrían calificarse de relativamente pequeñas ($d > .20$), cuestión que se confirma sólo parcialmente en el TB13 donde la potencia de la prueba está en torno a .70. En el resto de los casos la potencia de la prueba señala que existe una alta probabilidad de rechazar la hipótesis nula si fuera falsa ($p > .70$).

4.- Discusión y conclusiones.

La *Evaluación Final de la Educación Primaria* está prevista en la Ley Orgánica 2/2006 de Educación, modificada por la Ley Orgánica 8/2013 para la Mejora de la Calidad Educativa, y se encuentra regulada por el Real Decreto 1058/2015 que establece sus características, estructura y criterios de evaluación para el conjunto del Estado. Es una evaluación con impacto social ya que, si bien no emite calificaciones individuales, los centros escolares reciben un informe con sus resultados que deben difundir a la comunidad educativa, e incluso algunas administraciones educativas hacen públicos los resultados de los centros. La estructura de la prueba incluye un ejercicio de expresión escrita cuyos estándares de evaluación, comunes a todo el sistema educativo, fueron resumidos en la tabla 2.

En este contexto la finalidad el estudio era evaluar las características técnicas del instrumento desarrollado por el Ministerio de Educación, Cultura y Deporte, y explorar las posibilidades de comparabilidad de los resultados extraídos a partir del uso su rúbrica de corrección. Por ello, el primer objetivo fue analizar la estructura factorial de la matriz resultante. Los resultados indican que los ocho estándares se pueden resumir en un único factor que explicaría algo menos del 40% de la varianza. Ello ofrece garantías para reducir a una única puntuación la competencia en expresión escrita definida normativamente en los documentos ministeriales.

En relación al segundo objetivo del estudio (analizar los efectos del corrector) se advirtieron diferentes incidencias. En primer lugar se observan fuertes variaciones en cuanto a la severidad de los correctores. Incluso después de eliminar las puntuaciones del corrector más extremo (C07) la diferencia entre el juez más severo y el más benévolo supera los 3.5 puntos (casi el 30% del rango total de la escala). Se trata de una distancia importante teniendo en cuenta que los rangos de puntuación de los correctores son estrechos. Este hallazgo está en la línea de otros estudios que habitualmente informan de diferencias estadísticamente significativas en las puntuaciones asignadas por los correctores (Congdon & McQueen, 2000; Engelhard, 1996; Leckie & Baird, 2011; Lunz, et al., 1990; Park, 2010; Wang, & Yao, 2013; Wolfe, 2004).

En el análisis de fiabilidad intra-juez se encontró que el promedio del alfa de Cronbach era de .77. Este valor está por encima del criterio que Jonsson y Svingby (2007) consideran como suficiente ($\alpha > .70$). No obstante debe señalarse que más de la mitad de los correctores no alcanzan dicho valor, y que cuando se atiende al CCI el nivel de consistencia interna de los jueces queda aún más rebajado. De hecho, se han identificado dos casos (C08 y C03) que hacen un uso muy poco consistente de la rúbrica ($CCI_{3,k} < .50$) y otros cuatro (C02, C05, C06 y C13) donde el CCI está por debajo de .60.

En relación al tercer objetivo se han encontrado diferencias entre los tribunales, tanto en los resultados en la prueba, como en el grado de acuerdo inter-jueces. Dado que en cada tribunal había cuatro puntuaciones totales, para el conjunto de los tribunales hay 78 pares de correlaciones posibles. Prácticamente el 60% de estos pares está por encima de .50. Sin embargo, se identificaron 11 correlaciones inferiores a .30, y en 10 de los 11 casos estaban involucrados C08 y C13, cuyas valoraciones, en general, son poco coherentes con el resto de correctores. Esto afecta a los estadísticos de consistencia de aquellos tribunales donde fueron asignados estos dos correctores, en especial al TB08 que fue en el que ambos coincidieron. Finalmente, los resultados del ANOVA de medidas repetidas señalan que en la mayoría de los tribunales hay una diferencia significativa entre las puntuaciones de los correctores. Es posible que el dato más llamativo sea que el tribunal donde las diferencias fueron más pequeñas (TB08) era

precisamente el tribunal con menor acuerdo entre correctores. Esto parece apuntar a que incluso con bajos niveles de acuerdo es posible encontrar promedios de puntuación similares.

En definitiva, se han identificado importantes diferencias en la severidad, uso consistente de la rúbrica y acuerdo entre correctores. Sin embargo, el diseño Youden de 13 bloques debiera controlar gran parte de estas diferencias (al menos en cuanto a severidad) al asignar a los correctores a tribunales mediante un procedimiento matricial y sistemático (Fernández-Alonso & Muñiz, 2011). Por tanto, es probable que las diferencias y la falta de consenso tengan su explicación en otros factores como el perfil de los correctores y su familiaridad con la rúbrica, la naturaleza de la tarea o el propio diseño de la rúbrica. Los correctores responden a un perfil similar: tenían poca experiencia docente, eran noveles en este tipo de tareas y recibieron formación y entrenamiento específico con la rúbrica a emplear. No obstante los datos disponibles señalan que si bien la experiencia previa, entrenamiento, familiaridad con las rúbricas mejoran el consenso no eliminan definitivamente las diferencias (Congdon & McQueen, 2000; Linacre, Engelhard, Tatum, & Myford, 1994; McNamara, 1996). Con respecto a la naturaleza de la tarea a evaluar hay evidencias que señalan que los ensayos escritos y los exámenes de las materias socio-lingüísticas –el tipo de rúbrica de este trabajo- producen valoraciones menos fiables que los ejercicios físico-manuales o los exámenes científico-matemáticos (Cuxart Jardí, 2000). Finalmente la escala presenta algunas características que pueden dificultar el acuerdo entre correctores. Es cierto que se trata de una rúbrica analítica que, en evaluaciones de alto impacto, es preferible a la holística (Kuo, 2007). Sin embargo, para un funcionamiento óptimo las rúbricas analíticas deben incluir descripciones detalladas, criterios de separación de las puntuaciones entre los niveles de ejecución claramente especificados y ejemplos concretos codificados (Baird, Meadows, Leckie & Caro, 2017; Kuo, 2007). Revisando la rúbrica publicada por el Ministerio de Educación, Cultura y Deporte (2016) no es fácil encontrar muchas de las características apuntadas. Una última propiedad de la rúbrica que pudo afectar a los niveles de acuerdo entre correctores encontrados es el número de niveles puntuación. La puntuación final, es decir, la suma de las puntuaciones de los ocho estándares contenía 14 niveles de puntuación (rango total de 0 a 13 puntos), cuando se sabe que a medida que aumenta el número de niveles disminuye se producen valoraciones menos fiables (Jonsson & Svingby, 2007).

Finalmente, el estudio presenta algunas limitaciones que, por otro lado marcan líneas de trabajo futuro con importantes implicaciones prácticas. Los métodos clásicos para estudiar la fiabilidad entre correctores tienen una limitación operativa clara: no están diseñados para incluir las variaciones entre los correctores dentro del modelo de estimación de la competencia del alumnado. Sin embargo, las diferencias encontradas en las calificaciones apuntan a la necesidad de incluir el efecto del corrector en el modelo de estimación, por lo que en el futuro será necesario implementar otros modelos de análisis que puedan corregir las estimaciones de la competencia de los sujetos teniendo en cuenta las diferencias entre jueces (Adams & Wu, 2010; Eckes, 2009; Lunz et al., 1990; Prieto, 2011). En general los estudios se limitan a señalar diferencias entre correctores pero son escasos los trabajos que se aventuran a revisar y comparar las calificaciones originales con las obtenidas un vez descontado el efecto del corrector (Eckes, 2005). Hasta donde alcanza nuestro conocimiento esta sería una tarea pendiente en España. Por último, se sospecha que parte de las diferencias en las puntuaciones encontradas pueden tener su origen en el diseño de la rúbrica de corrección empleada. Por ello, una segunda línea de trabajo sería la mejora de la definición de los criterios

rúbricas y una separación de los niveles de puntuación que ayuden a mejorar la consistencia y el acuerdo entre correctores.

Anualmente cientos de miles de estudiantes se inscriben en pruebas para la obtención de una titulación oficial y compiten por premios o por acceder a estudios de secundaria postobligatoria, enseñanzas de régimen especial y estudios superiores. Todos estos procesos incluyen ejercicios de ejecución que son evaluados mediante rúbricas y donde existen potenciales efectos de corrección. En comparación con el volumen de candidatos implicados, las evidencias sobre la adecuación de las rúbricas y la fiabilidad de los correctores parecen ciertamente escasas.

5.- Referencias

Abad, F. J., Olea, J., Ponsoda, V., & García, C. (2011). *Medición en ciencias sociales y de la salud*. Madrid: Síntesis.

Adams, R., & Wu, M. (2010). *The analysis of rater effects*. Recuperado en diciembre de 2017 de: https://www.acer.org/files/Conquest-Tutorial-3-RaterEffects.pdf.

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

Barrett, P. (2001). *Conventional interrater reliability: definitions, formulae, and worked examples in SPSS and STATISTICA*. Recuperado en diciembre de 2017 de: http://www.pbarrett.net/techpapers/irr_conventional.pdf.

Bravo-Arteaga, A. & Fernández del Valle, J. C. (2000). La evaluación convencional frente a los nuevos modelos de evaluación auténtica. *Psicothema*, *12*(S. 2), 95-99.

Byrne, B. M. (2001). *Structural Equation Modeling with AMOS*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hilsdale. NJ: Lawrence Earlbaum Associates.

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

Cuxart Jardí, A. (2000). Modelos estadísticos y evaluación: tres estudios en educación. *Revista de Educación*, 323, 369-394.

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

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (Section H)*. Strasbourg, France: Council of Europe/Language Policy. Recuperado en septiembre de 2017 de https://rm.coe.int/1680667a23#search=eckes.

European Commission/EACEA/Eurydice (2016). *Structural indicators on achievement in basic skills in Europe – 2016. Eurydice Report*. Luxembourg: Publications Office of the European Union. doi:10.2797/092314.

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

European Commission/EACEA/Eurydice (2014). *Modernisation of higher education in Europe: Access, retention and employability*. Luxembourg: Publications Office of the European Union. doi: 10.2797/72146.

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

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

Engelhard, G. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement, 33*(1), 56-70. doi: 10.1111/j.1745-3984.1996.tb00479.x.

Feeley, T. H. (2002). Comment on Halo Effects in Rating and Evaluation Research. *Human Communication Research*, 28: 578-586. doi:10.1111/j.1468-2958.2002.tb00825.x

Fernández-Alonso, R. & Muñiz, J. (2011). Diseño de cuadernillos para la evaluación de las competencias básicas. *Aula Abierta, 39*(2), 3-34.

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

Gwet, K. L. (2014). *Handbook of inter-rater reliability. The definitive guide to measuring the extent of agreement among raters*. Gaithersburg, MD: Advanced Analytics, LLC.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23-34.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179-185. https://doi.org/10.1007/BF02289447.

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

Koo. T. K., & Li. M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2). 155-163. http://doi.org/10.1016/j.jcm.2016.02.012.

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

LaRoche, S., Joncas, M., & Foy, P. (2016). Sample design in TIMSS 2015. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in TIMSS 2015* (pp. 3.1-3.37). Recuperado en diciembre de 2017 de: Boston College, TIMSS & PIRLS International Study Center website: http://timss.bc.edu/publications/timss/2015-methods/chapter-3.html.

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

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

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

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

McGraw. K. O., & Wong. S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1). 30-46.

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

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

OECD (2014). *PISA 2012 Technical Report*. Paris: OECD Publishing. Recuperado en Septiembre de 2017 de: https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf .

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

Pérez-Gil, J. A., Chacón Moscoso, S. y Moreno Rodríguez, R. (2000). Construct Validity: The Use of Factor Analysis. *Psicothema, 12*(2), 441-446.

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

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

Shrout . P. E.. & Fleiss. J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2). 420-428.

Stemler, S.E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4). Recuperado en Septiembre de 2007 de: http://PAREonline.net/getvn.asp?v=9&n=4.

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

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

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, *16*(2), 209-220. http://dx.doi.org/10.1037/a0023353.

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

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