El uso del Marco Común Europeo de Referencia para las Lenguas para evaluar las redacciones en el examen de inglés en las Pruebas de Acceso a la Universidad[1]

The use of the Common European Framework of Reference for Languages to evaluate the compositions in the English exam in the University Entrance Examination

María Belén Díez-Bedmar
Universidad de Jaén. Facultad de Humanidades y Ciencias de la Educación. Departamento de Filología Inglesa. Jaén, España.

Traducción de María Belén Díez-Bedmar

Resumen

El Marco Común Europeo de Referencia para las Lenguas (MCER) (Consejo de Europa, 2001) se ha convertido en el estándar para describir y evaluar el dominio que los estudiantes tienen en una segunda lengua o lengua extranjera. Sin embargo, no se ha utilizado aún para evaluar el dominio de inglés que los alumnos tienen antes de entrar en la Universidad, es decir, cuando escriben su examen de inglés en las Pruebas de Acceso a la Universidad. Este artículo tiene como objetivo cubrir la laguna existente por medio de un doble objetivo. Primero, usar el MCER para evaluar las redacciones escritas en dicho examen, y explorar los niveles en los que se encuentran las redacciones. En segundo lugar, analizar la fiabilidad interjueces cuando se utiliza el MCER para descubrir los problemas que se encuentran al utilizarlo. Para poder alcanzar estos objetivos, se seleccionó una muestra representativa de las redacciones escritas sobre un tema en el examen, y se pidió a dos evaluadoras que clasificaran las redacciones de acuerdo con el MCER. Los resultados de este artículo muestran que la fiabilidad interjueces es muy baja (k= ,245) cuando se utiliza el MCER. Los datos obtenidos también apuntan a que, teniendo en cuenta los casos en los que las dos evaluadoras tuvieron acuerdo total, la mayoría de las redacciones escritas por los alumnos (91,33%) se encuentra en el nivel B1. Estos resultados pueden informar a los evaluadores, a los diseñadores de pruebas y a las autoridades educativas de los niveles otorgados a las redacciones. Los datos aportados también señalan qué aspectos necesitarían revisión o adaptación si el Marco Común de Referencia de las Lenguas se va a utilizar en el futuro para evaluar las redacciones en el examen de inglés de la Prueba de Acceso a la Universidad.

Palabras clave: evaluación, expresión escrita, lenguas extranjeras, Marco Común Europeo de Referencia para las Lenguas, niveles, Prueba de Acceso a la Universidad, fiabilidad interjueces

Abstract

The Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) has become the standard used to describe and evaluate the students' command of a second or foreign language. However, it has not been used yet to evaluate the students' command of English as a

[1] La autora quiere agradecer la ayuda recibida de la Comisión Interuniversitaria Andaluza para las Pruebas de Acceso a la Universidad, y especialmente de su secretario, Dr. Bengoa Díaz, por permitirle el acceso a los exámenes de inglés escritos para la PAU en Junio de 2008 en Jaén. También se agradece el apoyo del Vicerrectorado de Estudiantes e Inserción Laboral en la Universidad de Jaén, y de los Drs. Bueno González y Pérez Paredes. De igual manera, la autora expresa su gratitud al proyecto "El sistema de acceso a la Universidad: propuestas en la gestión, decisión e inferencias en el área de lenguas extranjeras" (FFI2011-22442), financiado por el Ministerio de Educación, por hacer posible la publicación de este artículo.

foreign language before entering university, i.e. when writing the English exam in the University Entrance Examination. This paper aims at bridging this gap in the literature by means of a twofold objective. First, to use the CEFR to evaluate the compositions written in that exam, and explore the levels which are awarded to those texts. Second, to analyse the inter-raters' reliability when using the CEFR to unveil the problems found when raters use it. To do so, a representative sample of the compositions written on one topic in the exam was selected, and two raters were asked to evaluate them according to the CEFR. The results in this article show that the inter-raters agreement is very low (k= .245) when they use the CEFR. The findings also highlight that, when the cases in which there was total inter-rater agreement, most of the compositions (91.33%) were placed at B1 level. The results of this paper may inform raters, test designers and education authorities of the levels which are awarded to the students' compositions at this stage. Similarly, the findings highlight the rating aspects which would need revision or adaptation if the CEFR is to be used in the future to evaluate the compositions in the English exam in the University Entrance Examination.

Keywords: evaluation, written expression, foreign languages, Common European Framework of Reference for Languages, levels, University Entrance Examination, inter-raters' reliability

Introducción

El establecimiento del Espacio Europeo de Educación Superior (EEES), que asegura la transparencia y la comparación de distintos sistemas de certificados en Europa, así como el reconocimiento de los estudios cursados en cualquier universidad europea (es decir, los programas de movilidad), y facilita el acceso de los estudiantes a programas de grado o postgrado en países extranjeros, requería que se fijara un criterio para determinar el nivel de competencia en lenguas extranjeras de los estudiantes.

El Marco Común de Referencia para las Lenguas (MCRL) (Consejo de Europa, 2001), fue el elegido para describir los estándares que tienen que alcanzar los alumnos de educación secundaria o universitaria. En el contexto español los niveles del Marco Común se han utilizado para determinar los niveles requeridos en un momento muy importante de la vida académica del estudiante: la Prueba de Acceso a la Universidad (PAU). Sin embargo, hay una contradicción entre el uso de los niveles del Marco y su aplicación a la evaluación de la producción escrita en lengua extranjera en el examen de inglés en la PAU: mientras que los niveles del Marco se mencionan en la documentación legal, la evaluación del examen no se lleva a cabo siguiendo el Marco y los descriptores de lo que "puede hacer" la persona que se examina. Por lo tanto, entre los próximos cambios que se podrían implementar en la evaluación del examen de inglés de la PAU estarían el uso del Marco y de los descriptores de lo que "Puede hacer" la persona que se examina.

Aunque se han publicado estudios previos sobre el examen de inglés en la PAU (veánse revisiones en García Laborda, 2006 y Díez-Bedmar, 2011), y sobre el uso de la lengua extranjera por parte de los alumnos en el examen de inglés (Díez-Bedmar, en prensa), la evaluación de la expresión escrita por medio del Marco no se ha explorado aún. Por esta razón, este artículo pretende cubrir ese vacío en la literatura con un doble objetivo. En primer lugar, utilizar el Marco para explorar los niveles que se le otorgan a una muestra representativa de las redacciones escritas en el examen de inglés de la PAU. Así, es posible saber si los estudiantes cumplen los requisitos que aparecen en los documentos legales en relación al uso de la lengua extranjera. El segundo objetivo es analizar la fiabilidad interjueces cuando se utiliza el Marco, para poder descubrir los problemas que surgen cuando las evaluadoras que trabajan en la enseñanza secundaria y universitaria (como es el caso hoy en día) utilizan el Marco Común para evaluar las redacciones.

Los resultados de este artículo se pueden utilizar para que los expertos que diseñan el examen de inglés, los evaluadores y las autoridades pertinentes conozcan los niveles del Marco que los alumnos muestran en este examen. De igual manera, los datos obtenidos desvelan los

aspectos de evaluación que necesitan revisión o adaptación si el Marco y los descriptores de lo que "puede hacer" la persona que se examina se van a utilizar en el futuro para evaluar el examen de inglés de la PAU.

El marco conceptual

La revisión bibliográfica que se presenta a continuación se dividirá en dos grandes bloques. El primero ofrecerá una panorámica de los niveles del Marco que se han establecido como requisitos para los estudiantes en España y fuera de nuestras fronteras. El segundo bloque estará dedicado a la revisión de las causas de errores que se encuentran al evaluar pruebas, en lo que se refiere al tipo de escala de evaluación y los evaluadores.

El Marco Común de Referencia par alas Lenguas: niveles y requisitos

El Marco Común de Referencia para las Lenguas es una escala analítica redactada en términos positivos que está centrada en el examinador, el responsable de elaborar las pruebas, o el usuario (Alderson, 1991). El Marco proporciona una tabla con descriptores de actividades comunicativas, así como de los niveles de la lengua extranjera relacionados con competencias específicas. Éstos se presentan en dos dimensiones, la vertical y la horizontal. En la primera se evalúa si los estudiantes han alcanzado los descriptores funcionales de lo que "puede hacer" la persona que se examina para cada nivel, normalmente con la ayuda de un examen de referencia, lo que da lugar a la asignación de un nivel del Marco para cada prueba. Sin embargo, en la segunda dimensión, los evaluadores comprueban si el estudiante ha alcanzado los criterios específicos relacionados con la variedad, coherencia y precisión en su uso de la lengua extranjera, así como los descriptores de lo que "puede hacer" la persona que se examina según el tipo de texto. Por esta razón, se puede decir que se utilizan escalas analíticas, pero la evaluación es holística ya que el evaluador necesita juzgar el texto utilizando los criterios y relacionándolos con la producción del estudiante. A pesar de las ventajas del Marco, se han criticado muchos aspectos de su diseño. Entre otros, se ha citado en la literatura la necesidad de explicitar más los niveles (Alderson y col., 2004, citado en Weir, 2005; Huhta y col., 2002; Kaftandjieva y Takala, 2002).

En el contexto español, los niveles del Marco se mencionaron por primera vez en Andalucía en 2008 en la Orden (BOJA 169, 26-08-2008) que determina el currículo para Bachillerato, es decir, los dos últimos años voluntarios de la Educación Secundaria, antes de entrar en la Universidad. Como se dice en el documento, en aquellos casos en los que el estudiante no ha cursado ninguna segunda lengua con anterioridad, se establecen los niveles A1 y A2, mientras que si el alumno ha cursado algún tipo de segundo idioma los niveles esperados son A2 y B1. De la misma manera, las directrices que se han diseñado y publicado por el Distrito Único Andaluz para el examen de Selectividad desde el año 2009-2010 en adelante (Distrito Único Andaluz, 2010, 2011) también usan los niveles del Marco y explicitan que el Bachillerato consolida un nivel de competencia B1 en una segunda lengua. Para ayudar a los evaluadores a familiarizarse con los descriptores de lo que "puede hacer" la persona que se examina, las directrices anteriormente citadas reproducen las matrices de evaluación para los dos niveles del Marco que se esperan en la producción del alumnado en el examen de inglés de la PAU, esto es, el A2 y el B1.

Los niveles que los estudiantes tienen que alcanzar antes de terminar un grado o empezar un Máster no se han establecido unitariamente en el contexto nacional. Por ejemplo, se requiere el nivel B1 como requisito mínimo para terminar el grado de Maestro en Educación Infantil y el Grado de Maestro en Educación Primaria (véanse la Orden ECI/3854/2007 y la Orden ECI/3857/2007). Este mismo nivel es un requisito de entrada para el Máster Universitario

en Profesorado de Educación Secundaria Obligatoria y Bachillerato, Formación Profesional y Enseñanza de Idiomas. Aparte de estos dos grados y este Máster, cada universidad decide los requisitos mínimos de segundas lenguas o lenguas extranjeras que los estudiantes tienen que alcanzar para terminar un grado, como indican Halbach, Lázaro Lafuente, y Pérez Guerra (2010).

Una situación parecida se encuentra en el contexto internacional. El nivel B1 se ha impuesto como requisito para terminar la educación secundaria en algunos países como Chile (Khalifa, Robinson, y Harvey, 2010) y Colombia (Gómez Montes, Mariño, Pike, y Moss, 2010). Este mismo nivel también se considera una característica meritoria en los curricula de los estudiantes en China (Xueling, Meizi, y Bateman, 2010). El nivel B1 también se ha establecido como nivel mínimo (Randall, 2010) para terminar un grado en Francia e Italia. Por ultimo, otras instituciones consideran que se requiere un nivel superior, es decir, un nivel B2 o un nivel C, para entrar en la Universidad o para ejercer una profesión en la que se requiera el uso de inglés (Green, 2008).

Como se puede ver, todavía hay aspectos que debatir acerca de los niveles de entrada o salida en diferentes universidades. Lo que parece claro es que hay un acuerdo en el uso del Marco a la hora de promover estándares comunes para establecer requisitos de entrada y salida de grados y/o postgrados, y para diseñar programas, currícula, exámenes y materiales en la enseñanza del inglés.

Causas de errores de medida al evaluar la destreza escritora: la escala de evaluación y los evaluadores

La fiabilidad es una cualidad de las calificaciones de las pruebas (Bachman, 1990). Dicho de otra manera, la fiabilidad está relacionada con el grado en el que las puntuaciones son resultado de una prueba que no tiene ningún tipo de error de medida, independientemente del momento en el que se haga la prueba, la forma de la misma, los evaluadores que la puntúen, etc. (Bachman, 1990; Hamp-Lyons, 1991b; Weigle, 2002; etc.). Debido a la importancia que tiene la fiabilidad en una prueba tan importante como es la PAU, a continuación se describen los errores de medida que pueden estar causados por los evaluadores o por la escala utilizada.

Los factores que causan los errores de medida por parte de los jueces aparecen en un buen número de publicaciones (Bachman y Palmer, 1996; Lumley, 2005; McNamara, 1996; Shaw y Weir, 2007; Weigle, 2002; Weir, 2005; etc.), ya que la fiabilidad de la evaluación normalmente se refiere a la fiabilidad de los jueces (Hamp-Lyons, 2007). De hecho, una de las dos consideraciones centrales en el proceso de evaluación (la otra es la definición de la escala de evaluación) es asegurarse de que los jueces saben entender y aplicar la escala (Weigle, 2002).

Aunque se supone que las escalas o criterios de evaluación ayudan a los jueces a interactuar con los textos de los estudiantes y evaluarlos de manera similar, puede que la benevolencia o severidad de los jueces no cambie (Kondo-Brown, 2002; McNamara, 1996; Weigle, 1998). De hecho, los jueces están influidos por varios factores cuando evalúan. Por ejemplo, la literatura especializada menciona factores como el efecto del uso de escalas diferentes, los años de experiencia del evaluador, su bagaje académico, edad y género (Vaughan, 1991; revisiones en Weigle, 2002), con resultados contradictorios en algunas ocasiones. Así, el uso de escalas holísticas por parte de jueces experimentados parece que da lugar a puntuaciones más benevolentes, mientras que la experiencia de los jueces no tiene ningún tipo de efecto cuando se utilizan escalas analíticas (Song y Caruso, 1996). Sin embargo, cuando los jueces experimentados se comparan con los menos experimentados, los experimentados son más estrictos (Sweedler-Brown, 1985), usan estrategias más eficaces a la hora de evaluar, y muestran un conocimiento más amplio de recursos para evaluar el texto (Cumming, 1990), leen el texto de una sola vez, lo evalúan incluyendo comentarios al final y utilizan una variedad de acciones más limitada que la de los jueces menos expertos (Huot, 1993; Pula y Huot, 1993; Wolfe y Ranney, 1996). Sin embargo, otros estudios han señalado que los jueces inexpertos son más estrictos (Weigle, 1998). Otro factor importante es el bagaje académico de los jueces. Así, los jueces que están especializados en inglés como segunda lengua y otros miembros de la

comunidad universitaria que no lo están evalúan las redacciones de forma diferente, o no aplican igual los criterios de evaluación (Hamp-Lyons, 1991a; Weigle, Boldt, y Valsecchi, 2003). Sin embargo, parece que el uso de escalas analíticas disminuye las diferencias entre los miembros de la comunidad universitaria de diversas disciplinas (Song y Caruso, 1996). El hecho de que el juez sea hablante nativo o no de la lengua cuya producción se evalúa también es una variable que se debe de tener en cuenta, ya que los hablantes no nativos son más estrictos que los nativos (Bueno González, 1992; Hyland y Anan, 2006). Por último, se ha señalado que la edad de los jueces también puede influir el proceso de evaluación, ya que los profesores mayores son más benevolentes en algunas ocasiones (Santos, 1988; Vann, Meyer, y Lorenz, 1984), aunque no es siempre el caso (Roberts y Cimasko, 2008).

Además de las características propias de los jueces, el uso de las escalas de evaluación resulta difícil. Como muestra DeRemer (1998), los jueces proceden de varias formas a la hora de utilizarlas. Si la escala de evaluación no ofrece suficiente información para describir los textos de los estudiantes, o si los descriptores en la escala no están lo suficientemente especificados, los jueces pueden tener dificultades a la hora de evaluar. Por lo tanto, los evaluadores desarrollan estrategias para realizar su tarea. Así, puede ser que un entendimiento parecido de los contenidos de cada categoría de evaluación no suponga una misma aplicación de los contenidos de la escala (Lumley, 2002), y es posible que los jueces no consideren o evalúen los contenidos de esa categoría de forma similar (Turner y Upshur, 2002).

Otras variables importantes relacionadas con el uso de la escala son el número de aspectos (incluidos en la escala) a los que los jueces tienen que prestar atención, o el número de escalas que hay que considerar para evaluar una prueba. El Marco sugiere que cuatro o cinco categorías por nivel son el límite para evitar que la demanda cognitiva afecte a los evaluadores (Consejo de Europa, 2001). El número de niveles en una escala también puede afectar a la tarea de los evaluadores, ya que pueden no ser capaces de discernir entre los niveles establecidos en la escala (Bachman y Palmer, 1996; Penny, Jonson, y Gordon, 2000). Por este motivo, se necesita considerar un número fiable y práctico de niveles (Bachman y Palmer, 1996). Según Penny, Johnson, y Gordon (2000), más de ocho niveles en una escala podría causar problemas en el acuerdo entre jueces, y el Marco aboga por seis niveles, que son los que se corresponden con los niveles naturales con los que los profesores están familiarizados, es decir, principiante, elemental, intermedio bajo, intermedio, intermedio alto y avanzado (Consejo de Europa, 2001).

Además de la necesidad de entrenar a los jueces, se pueden utilizar métodos de resolución de puntuaciones para incrementar la fiabilidad de los jueces cuando los efectos de las variables mencionadas anteriormente pueden sesgar los resultados. Entre los métodos más utilizados están aquellos que contemplan o bien combinar las puntuaciones de dos jueces sustituyendo la puntuación por la que da un experto, la combinación de puntuaciones de dos jueces y la de un experto y, finalmente, la combinación de la puntuación del experto con aquella del juez cuya puntuación se acerque más a la del experto (Johnson, Penny, y Gordon, 2000; Weigle, 2002). Otra opción para mejorar la fiabilidad interjueces puede ser incrementar el número de puntuaciones entre niveles añadiendo un decimal adicional (Cronbach, Linn, Brennan, y Haertel, 1995, citado en Penny, Jonson, y Gordon, 2000). Las ventajas que se obtienen al hacerlo incluyen la reducción del desacuerdo, el incremento de la fiabilidad interjueces, así como la oportunidad de dejar a los jueces expresar el grado de ambigüedad que se puede encontrar cuando se evalúa un texto. Según Penny, Johnson, y Gordon (2000), cuando se utiliza un grado elevado de la puntuación, la media y la desviación típica no cambian significativamente, pero mejora el acuerdo entre jueces (Penny, Johnson, y Gordon, 2000). Por último, también se puede utilizar *scaling*, esto es, un análisis estadístico que detecta los jueces cuyas puntuaciones no están en la media y la desviación típica de los jueces como grupo (Shaw y Weir, 2007), así como el desarrollo y el análisis de sistemas automáticos de evaluación de textos (Burstein y Chodorow, 2002).

Los efectos que las variables de los evaluadores pueden tener en su proceso de evaluación se han explorado también en el examen de inglés de la PAU. Por ejemplo, Herrera Soler (2000-2001) consideró las variables género y lugar de trabajo de los evaluadores para concluir que los hombres son más benevolentes que las mujeres y que la precisión era un aspecto más importante para el profesorado de secundaria que para las mujeres que trabajan en

la universidad. En otro estudio, Amengual Pizarro (2005) encontró diferencias entre hombres y mujeres que trabajan en institutos, ya que las mujeres eran menos estrictas que los hombres. Sin embargo, ocurría lo contrario cuando los evaluadores trabajaban en la universidad. Este último dato coincide con los resultados de Herrera Soler (2000-2001), pero difiere de los resultados que indicaban que el grupo que otorgaba puntuaciones más estrictas era el de las mujeres que trabajaban en institutos (Amengual Pizarro, 2005). El tipo de evaluación que se hace, es decir, holística o analítica, también variaba si los exámenes eran evaluados por hombres o mujeres. Así, con evaluaciones holísticas y considerando el mismo lugar de trabajo, las mujeres otorgaban puntuaciones más altas. Sin embargo, el uso de la evaluación analítica daba lugar a que los hombres puntuaran más alto (Amengual Pizarro, 2005).

También se ha analizado el acuerdo entre jueces cuando se utilizan evaluaciones holísticas o analíticas en el examen de inglés de la PAU. Así, el acuerdo entre jueces cuando se utiliza evaluación holística es bajo (Amengual Pizarro, 2003-2004; Amengual Pizarro y Herrera Soler, 2003), como se puede ver en los resultados obtenidos (k= .6556) (Amengual Pizarro, 2003). Otros estudios sobre el grado de acuerdo de los jueces cuando se utilizan evaluaciones analíticas u holísticas también muestran que el acuerdo total no es alto cuando se utiliza la evaluación holística (k= .6390), pero es un poco más alto cuando se utilizan criterios analíticos (k= .5993) (Amengual Pizarro, 2005). Por ultimo, se ha comparado la evaluación holística y la evaluación holística focalizada que tiene en cuenta ciertos criterios para el proceso de evaluación (Watts y García Carbonell, 1998, 2005). Los resultados muestran que el uso de los criterios (que incluyen seis niveles de corrección y un descriptor por nivel) da lugar a mejores resultados.

Metodología

Esta sección está subdivide en dos. En la primera se describe el examen de inglés en la PAU en el Distrito Único Andaluz en junio de 2008, y se explica la selección del corpus de estudiantes utilizado en este estudio. En la segunda sección se ofrecen bios de las dos evaluadoras que evaluaron las redacciones de los estudiantes siguiendo el Marco, y la forma en la que dicha evaluación se llevó a cabo.

El Examen de ingles en la PAU: la selección del corpus de estudiantes

El examen de inglés en 2008 en Andalucía constaba de tres partes principales: *Comprehension*, *Use of English* y *Production*. Para el propósito de esta investigación, se tuvo en cuenta sólo la última parte del examen, es decir, la parte de *Production*, en la que se les pidió a los estudiantes que escribieran 80-100 palabras sobre uno de los dos temas propuestos.

El examen de inglés para la PAU en Jaén en Junio de 2008 fue realizado por 2.611 estudiantes. Los dos temas propuestos fueron: "*Where, outside Spain, would you like to go on a short pleasure trip?*" y "*Attracting more tourists is essential for the Spanish economy. Discuss*.". Como la evaluación se podría ver afectada por los diferentes tipos de texto que cada uno de los temas elicita, se calculó el tema que había sido elegido por el mayor número de estudiantes y se tuvieron en cuenta sólo las redacciones que se habían escrito sobre dicho tema. En este caso, la mayoría de estudiantes (1.406) eligió "*Where, outside Spain, would you like to go on a short pleasure trip?*". Para obtener una muestra representativa de las 1.406 redacciones escritas sobre ese tema, se utilizó el muestreo aleatorio simple (Cochran, 1977) (CI= 95%, p= q = .50) con el programa *Stats 1.1*. El resultado obtenido indicó que se necesitaban 302, así que se seleccionaron de forma aleatoria y con ellas se formó el corpus de estudiantes que se utilizó para esta investigación. Estas redacciones se pasaron a formato electrónico, teniendo especial

Fecha de entrada: 10-07-2011 Fecha de aceptación: 25-10-2011

cuidado en mantener una copia fiel de las redacciones de los alumnos. Cuando el corpus ya estuvo transcrito, se obtuvo la cantidad total de palabras del corpus de estudiantes, 34.403.

Las evaluadoras y el proceso de evaluación

Dos evaluadoras, hablantes nativas de inglés, fueron seleccionadas para evaluar cada una de las redacciones en el corpus de estudiantes. Ya que las características de los jueces son decisivas cuando se lleva a cabo el proceso de evaluación, se detallan a continuación dos breves bios.

La evaluadora 1 habla francés y alemán como lenguas extranjeras y es, además, hablante bilingüe de español, ya que lleva viviendo en España más de 32 años. Da clases en un instituto de secundaria, y también imparte docencia en el curso "*That's English*" en la Escuela Oficial de Idiomas de Jaén. Por último, esta evaluadora también da clases privadas de inglés en varios niveles. La evaluadora 2 habla francés como lengua extranjera y es también hablante bilingüe de español. Durante tres años académicos ha impartido clases prácticas de inglés hablado en varios departamentos de la Universidad de Murcia en niveles que van desde el A2 hasta el C1. En esa universidad también ha formado parte de varios proyectos de investigación relacionados con el uso del inglés como lengua extranjera.

Aunque las dos evaluadoras utilizan el Marco en su trabajo, siguiendo la recomendación de Salamoura (2008), para que pudieran comprobar que estaban aplicando los criterios de forma correcta, se les remitió a los materiales que están disponibles, es decir, las escalas del Marco Común de Referencia para las Lenguas para la producción escrita (Consejo de Europa, 2001), y la Tabla 5.8. titulada "*Written Assessment Criteria Grid*" en el manual titulado *Relating Language Examinations to the Common European Framework of Reference* (2003).[2]

En el proceso de evaluación, se permitió el uso del grado elevado de nivel (Cronbach, Linn, Brennan, y Haertel, 1995) en todos los niveles, aunque el Marco sólo lo permite en los niveles A2, B1 y B2. Además, también se emplearon los niveles de signo menos, ya que las evaluadoras estimaban que era necesario para poder realizar el proceso de evaluación mejor. Por lo tanto, cuando una evaluadora consideraba que una redacción estaba por debajo del estándar de un nivel (por ejemplo, el B1), pero no se podía considerar dentro del nivel por debajo con grado elevado (A2+, en este caso), se le permitía utilizar el nivel B1- (véase la Tabla I).

Como se puede apreciar en la Tabla I, la flexibilidad que permite el Marco para usar varios niveles posibilitó reducir el número de niveles de 18 (cuando se considera el uso del grado elevado de nivel y de los niveles de signo menos) a los 3 niveles básicos, según se necesitara para los análisis. Por lo tanto, si se necesitaba una reducción de 18 a 12 niveles se podía hacer añadiendo aquellas redacciones que habían recibido un nivel A1- y un A1 en un nivel A1 más amplio (y lo mismo con A2- y A2, B1- y B1, B2- y B2, C1- y C1, o C2- y C2). Igualmente, los 12 niveles se podían incluir también en 6 niveles, de forma que los niveles A1 y A1+ se pudieran considerar en un nivel A1. Finalmente, los 6 niveles podían también reducirse a tres niveles más generales, si los niveles A1 y A2 se incluían en un nivel A más amplio.

TABLA I. Flexibilidad de agrupación de los niveles del Marco Común.

Marco Común: Tres niveles	Marco Común: Seis niveles	Marco Común: Doce niveles	Marco Común: Dieciocho niveles
A Usuario Básico	A1 Acceso	A1	A1-
			A1
		A1+	A1+
	A2 Plataforma	A2	A2-
			A2
		A2+	A2+
B Usuario Independiente	B1 Umbral	B1	B1-
			B1
		B1+	B1+

[2] En la Tabla C-4 aparece una versión más reciente del *Written Assessment Criteria Grid* (Council of Europe, 2009).

	B2 Avanzado	B2	B2-
			B2
		B2+	B2+
C Usuario Competente	C1 Dominio Operativo Eficaz	C1	C1-
			C1
		C1+	C1+
	C2 Maestría	C2	C2-
			C2
		C2+	C2+

Para poder utilizar los niveles del Marco en análisis estadísticos, se le asignó un número a cada nivel (teniendo en cuenta 3, 6, 12 ó 18 niveles), como puede verse en la Tabla II. Para hacerlo, se diseñaron 4 escalas. La primera tiene valores de 1 a 3 (para los tres niveles del Marco), la segunda de 1 a 6 (para los 6 niveles del Marco), la tercera de 1 a 12 (cuando se tiene en cuenta el uso del grado elevado de nivel, es decir, 12 niveles del Marco), y la última escala de 1 a 18 (cuando se consideran los grados elevados de nivel y los niveles de signo menos, es decir, 18 niveles del Marco).

TABLA II. Conversión de los niveles del Marco en datos numéricos.

Tres niveles		**Seis niveles**		**Doce niveles**		**Dieciocho niveles**	
Nivel	Número	Nivel	Número	Nivel	Número	Nivel	Número
A	1	A1	1	A1	1	A1-	1
						A1	2
				A1+	2	A1+	3
		A2	2	A2	3	A2-	4
						A2	5
				A2+	4	A2+	6
B	2	B1	3	B1	5	B1-	7
						B1	8
				B1+	6	B1+	9
		B2	4	B2	7	B2-	10
						B2	11
				B2+	8	B2+	12
C	3	C1	5	C1	9	C1-	13
						C1	14
				C1+	10	C1+	15
		C2	6	C2	11	C2-	16
						C2	17
				C2+	12	C2+	18

Resultados y discusión

El uso del Marco: fiabilidad interjueces

Una vez que cada nivel del Marco se relacionó con un número en una escala, como se muestra en la Tabla II, se pudo proceder a calcular la distancia o diferencia entre el nivel del Marco que cada evaluadora asignó a cada redacción. Este proceso se llevó a cabo como se detalla a continuación. Si se consideraban los tres niveles generales (A, B y C), la distancia entre el nivel A1 otorgado por una evaluadora y el nivel B1 de otra evaluadora era 1. Si se tenían en cuenta 6 niveles, la distancia entre A1 y B1 era 2. Sin embargo, cuando se tenían en cuenta 12 niveles, la distancia era 4, y con el uso de 18 niveles la distancia era 6.

Una vez calculadas las distancias, el análisis de los datos reveló que, cuando se utilizaron 18 niveles, las dos evaluadoras dieron el mismo nivel a la misma redacción en 47 casos (véase la Tabla III). La moda en los casos de no acuerdo fue 105 (es decir, el 34,8% del número total de casos), lo que implica la diferencia de sólo un nivel en la opinión de las evaluadoras. Así, se podían dar los siguientes casos: una evaluadora daba un B1 a una redacción mientras que la otra evaluadora daba un B1+ o un B1-, o una evaluadora daba un nivel C1-, y la otra un B2+ o un C1. Los datos en la Tabla III muestran que la mayor diferencia entre los niveles que las evaluadoras otorgaron a la misma redacción fue 7 (lo que ocurrió en dos casos). Más específicamente, en una redacción una evaluadora dio un nivel A1- a una redacción, mientras que la otra evaluadora optó por un nivel B1. En el segundo caso, una evaluadora dio un nivel A1 y la otra otorgó un nivel B1+.

TABLA III. Distancias encontradas cuando se utilizaron 18 niveles del Marco.

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	0	47	15,6	15,6	15,6
	1	105	34,8	34,8	50,3
	2	55	18,2	18,2	68,5
	3	35	11,6	11,6	80,1
	4	39	12,9	12,9	93,0
	5	7	2,3	2,3	95,4
	6	12	4,0	4,0	99,3
	7	2	0,7	0,7	100,0
	Total	302	100,0	100,0	

Si se tienen en cuenta 12 niveles, se puede observar en la Tabla IV que hubo 100 casos en los que las evaluadoras estaban de acuerdo, y 107 casos en las que las dos evaluadoras difirieron en un nivel (esto es, el 35,4% del número total de casos). Por ejemplo, si una evaluadora otorgaba un nivel A2, la otra evaluadora daba o un nivel A2+ o un nivel A1+. Cuando se utilizaron 12 niveles, la diferencia mayor que se encontró fue en un caso en el que las dos evaluadoras difirieron en cinco niveles (una evaluadora dio un nivel A1 y la otra un nivel B1+). Además de este caso, es también interesante mencionar que en 13 redacciones (4,3% de los casos) las evaluadoras difirieron en 4 niveles.

TABLA IV. Distancias encontradas cuando se utilizaron 12 niveles del Marco.

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	0	100	33,1	33,1	33,1
	1	107	35,4	35,4	68,5
	2	59	19,5	19,5	88,1
	3	22	7,3	7,3	95,4
	4	13	4,3	4,3	99,7
	5	1	0,3	0,3	100,0
	Total	302	100,0	100,0	

El uso de 6 niveles en el Marco (Tabla V) revela que la moda es el acuerdo entre las dos evaluadoras (en 196 casos, es decir, el 64,9% de la muestra total), y la mayor diferencia encontrada entre los niveles otorgados es de 2, lo que ocurre en 15 casos. De hecho, en 14 casos la distancia fue la que se encuentra entre los niveles A1 y B1, mientras que en el otro caso la diferencia se debió a la asignación de los niveles A2 y B2 a la misma redacción.

TABLA V. Distancias encontradas cuando se utilizaron 6 niveles del Marco.

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	0	196	64,9	64,9	64,9
	1	91	30,1	30,1	95,0
	2	15	5,0	5,0	100,0
	Total	302	100,0	100,0	

Por último, los datos en la Tabla VI muestran que cuando se consideraron 3 niveles del Marco, las evaluadoras estuvieron de acuerdo en 226 casos. Como se puede apreciar en la misma tabla, incluso cuando se utilizaron los tres niveles básicos, hubo 76 ocasiones (el 25,2% de la cantidad total de redacciones) en las se difirió al otorgar un nivel A o B a la misma redacción.

TABLA VI. Distancias encontradas cuando se utilizaron 3 niveles del Marco.

		Frecuencia	Porcentaje	Porcentaje Válido	Porcentaje acumulado
Válidos	0	226	74,8	74,8	74,8
	1	76	25,2	25,2	100,0
	Total	302	100,0	100,0	

A modo de resumen, la Tabla VII muestra la información obtenida al analizar el acuerdo entre jueces cuando las evaluadoras otorgaron un nivel del Marco a las 302 redacciones del corpus de estudiantes:

TABLA VII. Acuerdo entre jueces cuando se consideraron 3, 6, 12 y 18 niveles del Marco.

	18 niveles del Marco	12 niveles del Marco	6 niveles del Marco	3 niveles del Marco
Acuerdo entre jueces	47 casos (15,56%)	100 casos (33,11%)	196 casos (64,9%)	226 casos (74,83%)

Como se puede observar, el número de casos en los que las evaluadoras no coincidieron en el nivel del Marco que debía ser otorgado a una redacción fue alto (25,17%), incluso cuando se utilizaron los 3 niveles básicos. Con 6 niveles, hubo 15 casos en los que las evaluadoras difirieron en dos niveles, y 91 casos en los que difirieron en 1 nivel (véase la Tabla V). En otras palabras, hay un 30,1% de los casos en los que hay problemas a la hora de distinguir entre los niveles A1 y A2, A2 y B1, o B1 y B2. Por lo tanto, parece que las matrices de valoración funcionales disponibles para esos niveles no están lo suficientemente claras para evaluar un 25,17% de las redacciones cuando se utilizan 3 niveles, y un 35.1% cuando se utilizan 6 niveles. Tal y como afirman Alderson y colaboradores (2004), Huhta y colaboradores (2002) o Kaftandjieva y Takala (2002), parece que las escalas del Marco necesitan una descripción más detallada para que se puedan entender de una forma más fácil y se puedan aplicar mejor.

Los resultados obtenidos cuando se tienen en cuenta 12 ó 18 niveles del Marco, es decir, cuando se utilizan los niveles de signo más (o sea, 12 niveles) y los niveles de signo más y de signo menos (es decir, 18 niveles), muestran el descenso en el acuerdo total de las evaluadoras, en contra de los resultados obtenidos por Penny, Jonson, y Gordon (2000) cuando utilizaron los grados elevados de nivel con una evaluación holística. De hecho, el uso de más niveles del Marco reduce a la mitad el porcentaje de acuerdo total, lo que respalda la idea de Penny, Jonson, y Gordon (2000) de que utilizar más de 8 niveles en una escala muestra problemas en el acuerdo entre jueces.

El uso de los niveles de signo menos, debido a la sugerencia de las evaluadoras de que su uso mejoraría sus evaluaciones, supuso un acuerdo menor entre las evaluadoras. Como se puede ver en las diferencias encontradas cuando se utilizan 12 ó 18 niveles, los porcentajes de las redacciones a las que se les otorgó el mismo nivel fueron 33,11% y 15,56%, respectivamente.

Por lo tanto, se puede ver que el uso de los niveles de signo menos no mejora el acuerdo entre jueces, sino que lo reduce, ya que menos de la mitad del porcentaje de las redacciones recibieron el mismo nivel cuando se utilizó. Aparte de este resultado, es interesante mencionar que las evaluadoras no coincidieron en ningún caso en la asignación del mismo nivel de signo menos a la misma redacción, lo que parece indicar que el uso de los niveles de signo menos puede no ser recomendable.

La fiabilidad interjueces se calculó utilizando el coeficente kappa de Cohen. El bajo nivel obtenido (k= .245),[3] incluso más bajo que los que obtuvo Amengual Pizarro (2003, 2005), indica que a las evaluadoras les resulta difícil discernir qué nivel del Marco se debe otorgar a una redacción, lo que ya habían advertido con otras escalas Bachman y Palmer (1996) y Penny, Johnson, y Gordon (2000). Esto es así especialmente cuando las evaluadoras no saben cómo utilizar la escala (Weigle, 2002), cuando no hay exámenes de referencia o cuando la información en la matriz de valoración no es clara o suficiente para que las evaluadoras puedan decidir. El uso de los 6 niveles que el Marco propone, que además se corresponde con los seis niveles naturales con los que los profesores (Consejo de Europa, 2001) y los expertos están familiarizados, parece ser la mejor opción cuando se trata de encontrar un equilibrio entre el número de niveles utilizado y el acuerdo entre evaluadores. Así, se puede obtener una evaluación más fiable en una prueba tan importante como el examen de inglés en la PAU.

Otros resultados interesantes están relacionados con la forma en la que las características personales de las evaluadoras influyeron en su labor, así como los niveles que cada evaluadora otorgó a las redacciones. Aunque la variable sexo se tuvo en cuenta cuando se eligieron las evaluadoras (ambas son mujeres), su edad, experiencia docente y evaluadora son diferentes. Mientras que la evaluadora 1 tiene una amplia experiencia en secundaria, la evaluadora 2 tiene menos experiencia docente en general y ha trabajado sobre todo en la universidad.

Como se indicó en el apartado de revisión bibliográfica, la experiencia docente así como la institución en la que las evaluadoras trabajan juegan un papel fundamental en la fiabilidad interjueces. De hecho, la evaluadora 2 otorgó niveles más altos que la evaluadora 1, por lo que se comprobó que el uso de la misma escala no homogeneizó sus evaluaciones, de acuerdo con lo ya referido en las publicaciones de McNamara (1996), Weigle (1998) o Kondo-Brown (2002). Los resultados obtenidos al analizar los niveles del Marco otorgados por las evaluadoras apoyan los datos de Sweedler-Brown (1985), que indican que los evaluadores más experimentados son más estrictos, así como las indicaciones de Roberts y Cimasko (2008), que afirman que los profesores mayores no tienen porqué ser los más benevolentes. De forma similar, el lugar de trabajo parece que juega también un papel importante, ya que la evaluadora que trabaja en educación secundaria otorgó niveles del Marco más estrictos que la evaluadora 2, un resultado que está de acuerdo con Herrera-Soler (2000-2001), pero no con Amengual Pizarro (2005).

Los niveles del Marco otorgados a las redacciones de los estudiantes

Debido a la baja fiabilidad interjueces que se obtuvo, y para utilizar los datos más fiables al analizar el número de redacciones en cada nivel del Marco, se consideraron sólo aquellas redacciones en las que hubo un acuerdo entre evaluadoras del 100%.

Como se puede ver en la Tabla VIII, cuando se tuvieron en cuenta 18 niveles (es decir, 47 redacciones), a 1 redacción se le otorgó el nivel A2, a 23 redacciones el nivel B1, a 19 redacciones B1+ y a 4 redacciones B2, mientras que el uso de 12 niveles clasificó las 100 redacciones de la siguiente manera: 1 redacción en el nivel A2, 75 en el nivel B1+, 19 en el nivel B1+, 4 en el nivel B2 y 1 en el nivel C1. Cuando se utilizaron 6 niveles (196 redacciones), 10 de ellas se clasificaron en el nivel A2, 179 en el nivel B1, 6 en el nivel B2 y, por último, una redacción en el nivel C1. Finalmente, el uso de 3 niveles dividió las 226 redacciones entre 17 en el nivel A, 208 en el nivel B y sólo 1 en el nivel C.

[3] Según Altman (1991), los valores situados por encima de *k*= .8 son indicadores de un acuerdo entre jueces muy bueno.

Por lo tanto, la producción escrita de los estudiantes en el corpus analizado se caracteriza por un número elevado de redacciones en el nivel B (208), B1 (179), B1 (75) o B1 (23), considerando el uso de 3, 6, 12 y 18 niveles, respectivamente. Dicho de otra forma, si se consideran 3 niveles, el 92,03% de las redacciones se sitúan en el nivel B; si se consideran 6 niveles, el 94,43% de las redacciones están en el nivel B (niveles B1 o B2); cuando se utilizan 12 niveles, el 94% están en el nivel B (niveles B1, B1+ o B2) y, finalmente, cuando se utilizan 18 niveles, el 97,87% están en el nivel B (niveles B1-, B1, B1+, B2-, B2 o B2+).

TABLA VIII. Acuerdo entre jueces sobre las redacciones de los estudiantes, dependiendo del número de niveles del Marco utilizados.

Marco: Tres niveles	Acuerdo entre jueces (número de casos)	Marco: Seis niveles	Acuerdo entre jueces (número de casos)	Marco: Doce niveles	Acuerdo entre jueces (número de casos)	Marco: Dieciocho niveles	Acuerdo entre jueces (número de casos)
A Usuario Básico	17	A1 Acceso		A1		A1-	
						A1	
				A1+		A1+	
		A2 Plataforma	10	A2	1	A2-	
						A2	1
				A2+		A2+	
B Usuario Independiente	208	B1 Umbral	179	B1	75	B1-	
						B1	23
				B1+	19	B1+	19
		B2 Avanzado	6	B2	4	B2-	
						B2	4
				B2+		B2+	
C Usuario Competente	1	C1 Dominio Operativo Eficaz	1	C1	1	C1-	
						C1	
				C1+		C1+	
		C2 Maestría		C2		C2-	
						C2	
				C2+		C2+	

Si se analizan las diferencias entre los niveles B, B1 y B1+, los datos obtenidos muestran que, cuando se utilizan 3 niveles, las redacciones en el nivel B son el 92,03% del número total, y que cuando se tienen en cuenta 6 niveles, el 91,33% de las redacciones recibe un nivel B1. En el caso del uso de 12 niveles, las redacciones en el nivel B1, es decir, B1 o B1+, son el 94% (las redacciones en el nivel B1 representan el 75% del número total de redacciones, y aquellas en el nivel B1+ el 19%). Por último, el uso de 18 niveles revela que el 89,36% de las redacciones está en el nivel B1 (el 48,94% en el nivel B1 y el 40,2% de las redacciones en el nivel B1+).

Cuando las redacciones no recibían un nivel B1 del Marco, se situaban en los niveles A (especialmente A2) y B2. Si se utilizaban 3 niveles, el 7,52% de las redacciones se situaba en el nivel A, mientras que el 0,44% estaba en el nivel C. Cuando se consideraron 6 niveles, el 5,10% de las redacciones estaba en el nivel A1, el 3,06% en el nivel B2 y un 0,51% en el nivel C1. Por último, el uso de 12 ó 18 niveles revela que los niveles B2 son los más utilizados después de los niveles B1, con el 4% y el 8,51% de las redacciones, respectivamente. De hecho, sólo se le otorgó el nivel A2 a una redacción. Como resultado, se pueden hacer dos observaciones: primero, que hay un porcentaje muy bajo de redacciones en el nivel C (sólo una); y segundo, que cuantos más niveles del Marco se utilizan, más coinciden las evaluadoras en otorgar niveles B a las redacciones, si se compara con los casos de acuerdo total en las redacciones en el nivel A (véase la Tabla VIII).

Conclusiones

El objetivo principal de este artículo era utilizar el Marco para conseguir un doble propósito. En primer lugar, explorar qué niveles del Marco se otorgarían a una muestra representativa de las redacciones escritas para el examen de inglés de la PAU en la Universidad de Jaén en 2008, y ver si esos niveles eran semejantes a los niveles establecidos en la legislación Andaluza (BOJA 169, 26-08-2008). Segundo, ser conscientes de los problemas más importantes que se encuentran cuando los evaluadores utilizan el Marco y los descriptores de lo que "puede hacer" la persona que se examina para evaluar las redacciones de los alumnos en dicho examen.

Como se puede ver en los resultados obtenidos, el uso del Marco y de los descriptores de lo que "puede hacer" la persona que se examina presenta algunos problemas. De hecho, el bajo acuerdo entre jueces que se obtuvo, incluso cuando se consideraron los tres niveles básicos, A, B y C, puede indicar que las matrices de valoración no son suficientes para que los evaluadores puedan discernir el nivel en el que está cada redacción. Por lo tanto, sería necesario llevar a cabo un paso previo si los niveles del Marco Común y los descriptores de lo que "puede hacer" la persona que se examina se van a aplicar para evaluar las redacciones del examen de inglés. Aunque el entrenamiento de los evaluadores puede mejorar la fiabilidad interjueces, la complementación de las matrices de valoración disponibles con más descripciones del uso del lenguaje que los alumnos reflejan en cada nivel podría ser beneficiosa para que los evaluadores pudieran entender y usar mejor el Marco.

Aparte de la necesidad de complementar las matrices de valoración, los resultados de este artículo señalan algunas recomendaciones relacionadas con el uso de los niveles de signo más y de signo menos, es decir, 12 y 18 niveles del Marco, respectivamente. Como se discutió con anterioridad, su uso no es útil, ya que el acuerdo total entre jueces desciende significativamente cuando se utilizan, y los evaluadores no coinciden en el uso de los niveles de signo menos en ningún caso. Por esta razón, se propone el uso de 6 niveles del Marco si se quiere encontrar un equilibrio entre el número de niveles utilizados y el acuerdo total entre jueces.

En esta investigación se descubrió también que la edad de los evaluadores, su bagaje profesional y su experiencia son variables importantes a la hora de evaluar las redacciones, como se indica en la literatura especializada. De hecho, la evaluadora más experimentada (y de mayor edad), que trabaja en educación secundaria, es más estricta que la evaluadora menos experimentada, cuya experiencia profesional es universitaria. Por lo tanto, estas variables tienen que considerarse si se quieren evitar sesgos en el proceso de evaluación.

Una vez que las evaluadoras otorgaron un nivel del Marco a las redacciones (teniendo en cuenta el uso de 3, 6, 12 y 18 niveles), el análisis de los resultados obtenidos mostró que la mayoría de las redacciones estaban en el nivel B1. Por ejemplo, el uso de 6 niveles indica que el 91,33% de las redacciones en las que hubo un acuerdo total entre jueces estaba en el nivel B1. Por lo tanto, los estudiantes alcanzaron el nivel de competencia exigida al final de la educación

secundaria optativa y en el examen de inglés de la PAU en Andalucía. Como se indicaba en los documentos legales, los niveles que están representados en mayor medida en el examen son el nivel A2 y el nivel B1, aunque también se encontró alguna redacción en un nivel superior.

La limitación más importante de este estudio es que los resultados sólo se pueden generalizar a los estudiantes que hicieron el examen de inglés de la PAU en Junio de 2008 en Jaén, y escribieron sobre un tema específico. Por lo tanto, sería necesario realizar más investigaciones para replicar este estudio con las redacciones escritas por otros estudiantes en Andalucía y en otras universidades en España, utilizando una variedad de temas más amplia y evaluadas por otros evaluadores.

Referencias bibliográficas

ALTMAN, D. G. (1991). *Practical statistics for medical research*. London: Chapman and Hall.

ALDERSON, J. C. (1991). Bands and scores. En J.C. ALDERSON Y B. NORTH (Eds.), *Language testing in the 1990s* (71-86). London: Macmillan.

ALDERSON, J. C., FIGUERAS, N., KUIJPER, H., NOLD, G., TAKALA, S., Y TARDIEU, C. (2004). *The development of specifications for item development and classification within the common European Framework of Reference for Languages: Learning, teaching, assessment. Reading and listening. Final report of the Dutch construct project*. Unpublished document.

AMENGUAL PIZARRO, M. (2003). A study of different composition elements that raters respond to. *Estudios Ingleses de la Universidad Complutense*, 11, 53-72.

— (2003-2004). Rater discrepancy in the Spanish University Entrance Examination. *Journal of English Studies*, 4, 23-36.

— (2005). Posibles sesgos en los resultados del examen de Selectividad. En H. HERRERA SOLER Y J. GARCÍA LABORDA (Ed.), *Estudios y criterios para una selectividad de calidad en el examen de inglés* (121-148). Valencia: Universidad Politécnica de Valencia.

AMENGUAL PIZARRO, M., Y HERRERA SOLER, H. (2003). What is that raters are judging? En G. LUQUE AGULLÓ, A. BUENO GONZÁLEZ Y G. TEJADA MOLINA (Eds.), *Las lenguas en un mundo global* (11-18). Jaén: Servicio de Publicaciones de la Universidad de Jaén.

BACHMAN, L.F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

BACHMAN, L.F., Y PALMER, A.S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

BUENO GONZÁLEZ, A. (1992). Errores en la elección de palabras en inglés por alumnos de Bachillerato y COU. En A. BUENO GONZÁLEZ, J. A. CARINI MARTÍNEZ Y A. LINDE LÓPEZ (Eds.), *Análisis de errores en inglés: Tres casos prácticos* (39-105). Granada: Servicio de Publicaciones de la Universidad de Granada.

BURSTEIN, J., Y CHODOROW, M. (2002). Directions in automated essay analysis. En R.B. KAPLAN (Ed.), *The Oxford handbook of applied linguistics* (487-497). New York: Oxford University Press.

COUNCIL OF EUROPE. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

— (2003). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment, manual: Preliminary pilot version*. Strasbourg: Council of Europe, Language Policy Division.

— (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). A manual*. Strasbourg: Council of Europe, Language Policy Division.

COCHRAN, W. G. (1977). *Técnicas de muestreo*. México. Trillas

CRONBACH, L. L., LINN, R., BRENNAN, R., Y HAERTEL, E. (1995). *Generalizability analysis for educational assessments*. Los Angeles: Center for the Study of Evaluation, Standards, and Student Testing, University of California at Los Angeles.

CUMMING, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, 7, 31-51.

DEREMER, M.L. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing*, 5, 7-29.

DÍEZ-BEDMAR, M. B. (2011). The English Exam in the University Entrance Examination: An Overview of Studies. *Revista Canaria de Estudios Ingleses*, 63, 101-112.

DÍEZ-BEDMAR, M. B. (in press). Spanish pre-university students' use of English: CEA results from the University Entrance Examination. *International Journal of English Studies*, 11.

GARCÍA LABORDA, J. (2006). Analizando críticamente la Selectividad de inglés ¿Todos los estudiantes españoles tienen las mismas posibilidades? *Tesol Spain*, 30, 9-12.

GÓMEZ MONTES, I., MARIÑO, J., PIKE, N., Y MOSS, H. (2010). Colombia national bilingual project. *Research Notes*, 40, 17-22.

GREEN, A. (2008). English Profile: Functional progression in materials for ELT. *Research Notes*, 33, 19-25.

HAMP-LYONS, L. (1991a). Scoring procedures for ESL contexts. En L. HAMP-LYONS (Ed.), *Assessing second language writing in academic contexts* (241-276). Norwood, NJ: Ablex Publishing Corporation.

— (1991b). Basic concepts. En L. HAMP-LYONS (Ed.), *Assessing second language writing in academic contexts* (5-15). Norwood NJ: Ablex.

— (2007). Editorial: Worrying about rating. *Assessing Writing*, 12, 1-9.

HERRERA SOLER, E. (2000-2001). The effect of gender and working place of raters on university entrance examination scores. *Revista Española de Lingüística Aplicada*, 14, 161-180.

HUHTA, A., LUOMA, S., OSCARSON, M., SAJAVAARA, K., TAKALA, S., Y TEASDALE, A. (2002). A diagnostic language assessment system for adult learners. En J. C. ALDERSON (Ed.), *Common European Framework of Reference for Languages: Learning, teaching, assessment: Case studies* (130-146). Strasbourg: Council of Europe.

HUOT, B.A. (1993). The influence of holistic scoring procedures on reading and rating student essays. En M. M. WILLIAMSON, Y B.A. HUOT (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (206-236). Cresskill, NJ: Hampton Press.

HYLAND, K., Y ANAN, E. (2006). Teachers' perceptions of errors: The effects of first language and experience. *System*, 34, 509-519.

JOHNSON, R.L., PENNY, J., Y GORDON, B. (2000). The relation between score resolution methods and interrater reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education*, 13, 121-138.

KAFTANDJIEVA, F., Y TAKALA, S. (2002). Council of Europe scales of language proficiency: A validation study. En J. C. ALDERSON (Ed.), *Common European Framework of Reference for Languages: Learning, teaching, assessment: Case studies* (106-129). Strasbourg: Council of Europe Publishing.

KHALIFA, H., ROBINSON, M., Y HARVEY, S. (2010). Working together: The case of the English diagnostic test and the Chilean Ministry of Education. *Research Notes*, 40, 22-26.

KONDO-BROWN, K. (2002). A FACETS analysis of rater bias in measuring Japanese L2 writing performance. *Language Testing*, 19, 3-31.

LUMLEY, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, 19, 246-276.

— (2005). *Assessing second language writing: The rater's perspective*. Frankfurt: Peter Lang.

MCNAMARA, T. (1996). *Measuring Second Language performance*. London: Longman.

PENNY, J., JOHNSON, R. L., Y GORDON, B. (2000). The effect of rating augmentation on inter-rater reliability. An empirical study of a holistic rubric. *Assessing Writing*, 7, 143-164.

PULA, J.J., Y HUOT, B.A. (1993). A model of background influences on holistic raters. En M. WILLIAMSON Y B. HUOT (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (237-265). Cresskill, NJ: Hampton Press.

RANDALL, S. (2010). Cambridge ESOL's growing impact on English language teaching and learning in national education projects. *Research Notes*, 40, 2-3.

ROBERTS, F., Y CIMASKO, T. (2008). Evaluating ESL: Making sense of university professors' responses to second language writing. *Journal of Second Language Writing*, 17, 125-143.

SALAMOURA, A. (2008). Aligning English profile research data to the CEFR. *Research Notes*, 33, 5-7.

SANTOS, T. (1988). Professors' reactions to the academic writing of nonnative speaking students. *TESOL Quarterly*, 22, 69-90.

SHAW, S.D., Y WEIR, C.J. (2007). *Examining writing. Research and practice in assessing second language writing*. Cambridge: Cambridge University Press.

SONG, B., Y CARUSO, I. (1996). Do English and ESL faculty differ in evaluating the essays of Native English-Speaking, and ESL students? *Journal of Second Language Writing*, 5, 163-182.

SWEEDLER-BROWN, C.O. (1985). The influence of training and experience on holistic essay evaluation. *English Journal*, 74, 49-55

TURNER, C.E., Y UPSHUR, J.A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly*, 36, 49-70.

VANN, R.J., MEYER, D.E., Y LORENZ, F.O. (1984). Error gravity: A study of faculty opinion of ESL errors. *TESOL Quarterly*, 18, 427-440.

VAUGHAN, C. (1991). Holistic assessment: What goes on in the rater's mind? En L. HAMP-LYONS (Ed.), *Assessing second language writing in academic contexts* (11-125). Norwood, NJ: Ablex.

WATTS, F., Y GARCÍA CARBONELL, A. (2005). Control de calidad en la calificación de la prueba de lengua inglesa de selectividad. En H. HERRERA SOLER Y J. GARCÍA LABORDA (Eds.), *Estudios y criterios para una selectividad de calidad en el examen de inglés* (99-115). Valencia: Universidad Politécnica de Valencia.

WEIGLE, S.C., BOLDT, H., Y VALSECCHI, M.I. (2003). Effects of task and rater background in the evaluation of ESL student writing: A pilot study. *TESOL Quarterly*, 37, 345-354.

WEIGLE, S.C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15, 263-287.

— (2002). *Assessing writing*. Cambridge: Cambridge University Press.

WEIR, C.J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, 22, 281-300.

WOLFE, E.W., Y RANNEY, M. (1996). Expertise in essay scoring. En D.C. ADELSON Y E.A. DOMESHEK (Eds.), *Proceedings of ICLS 96* (545-550). Charlottesville, VA: Association for the Advancement of Computing in Education.

XUELING, C., MEIZI, H., Y BATEMAN, H. (2010). The use of BEC as a measurement instrument in Higher Education in China. *Research Notes*, 40, 13-15.

Recursos electrónicos

DISTRITO ÚNICO ANDALUZ (2010). *Directrices y orientaciones generales par alas pruebas de acceso a la Universidad. Curso 2010/2011. Lengua extranjera Inglés*. Retrieved on June, 16th 2010, from: http://www.ujaen.es/serv/acceso/documentos/orient_selectiv_2009_2010/ingles.pdf.

— (2011). *Directrices y orientaciones generales par alas pruebas de acceso a la Universidad. Curso 2010/2011. Lengua extranjera Inglés*. Retrieved on June 15th, 2011, from:

http://www.ujaen.es/serv/acceso/documentos/orient_selectiv_2010_2011/ingles.pdf.

HALBACH, A., LÁZARO LAFUENTE, A., Y PÉREZ GUERRA, J. (2010). *La acreditación del nivel de lengua inglesa en las universidades españolas*. British Council. Retrieved on July, 14th 2011, from: http://www.britishcouncil.org/spain_informe_acreditacion_ingles_universidades_espanolas_.pdf.

ORDEN ECI/3854/2007, de 27 de diciembre, por la que se establecen los requisitos para la verificación de los títulos universitarios oficiales que habiliten para el ejercicio de la profesión de Maestro en Educación Infantil. *Boletín Oficial del Estado*, 312. Retrieved on May, 20th, 2008, from: http://www.boe.es/boe/dias/2007/12/29/pdfs/A53735-53738.pdf

ORDEN ECI/3857/2007, de 27 de diciembre, por la que se establecen los requisitos para la verificación de los títulos universitarios oficiales que habiliten para el ejercicio de la profesión de Maestro en Educación Primaria. Boletín Oficial del Estado, 312. Retrieved on May, 20th, 2008, from: http://www.boe.es/boe/dias/2007/12/29/pdfs/A53747-53750.pdf

WATTS, F., Y GARCÍA CARBONELL, A. (1998). Rater agreement in English language assessment in the Spanish University access examination battery. *Language Testing Update* 23, Retrieved on November, 10th 2008, from: http://www.upv.es/diaal/publicaciones/watts3.pdf.

Dirección de contacto: María Belén Díez-Bedmar. Edificio D-2. Departamento de Filología Inglesa. Facultad de Humanidades y Ciencias de la Educación. Universidad de Jaén. Paraje las Lagunillas. 23071, Jaén. E-mail: belendb@ujaen.es