Evaluación psicométrica de la expresión oral en inglés de las Pruebas de Acceso a la Universidad[1]

Psychometric assessment of oral expression in English language in the University Entrance Examination

DOI: 10.4438/1988-592X-RE-2014-364-256

Javier Suárez-Álvarez
Universidad de Oviedo, Facultad de Psicología, Departamento de Psicología. Oviedo. España.
César González-Prieto
Rubén Fernández-Alonso
Consejería de Educación, Cultura y Deporte del Gobierno del Principado de Asturias. Oviedo. España.
Guillermo Gil
Instituto Nacional de Evaluación Educativa del Ministerio de Educación, Cultura y Deporte. Madrid. España.
José Muñiz
Universidad de Oviedo, Facultad de Psicología, Departamento de Psicología. Oviedo. España.

Resumen

El Real Decreto 1892/2008 establece que es obligatorio evaluar la expresión oral en lengua extranjera en la Prueba de Acceso a la Universidad (PAU), si bien hasta la fecha no se ha puesto en marcha tal evaluación. El Instituto Nacional de Evaluación Educativa (INEE) ha llevado a cabo un estudio piloto para ir perfilando la implementación de la prueba. Dentro de ese marco evaluativo, el objetivo del presente trabajo es estudiar el comportamiento psicométrico de los evaluadores encargados de examinar oralmente a los estudiantes. En este estudio participaron

(1) Agradecimientos. Parte de este trabajo ha sido financiado por el Ministerio Español de Economía y Competitividad (PSI2011-28638). Este trabajo ha sido posible gracias al apoyo recibido por parte del Instituto Nacional de Evaluación Educativa del Ministerio de Educación, Cultura y Deporte y de la Consejería de Educación y Universidades del Gobierno del Principado de Asturias.

Fecha de entrada: 01-02-2013 Fecha de aceptación: 27-09-2013

1.194 estudiantes pertenecientes a siete comunidades autónomas, con una edad media de 18,04 años, de los cuales el 57,4% eran mujeres. Cada estudiante fue evaluado por tres profesores. Se utilizaron 30 tribunales distintos, y en la evaluación participó un total de 90 profesores. Los análisis se realizaron utilizando la matriz formada por los alumnos evaluados por el mismo tribunal evaluador. Los resultados muestran que las puntuaciones asignadas por los distintos evaluadores a los estudiantes tienen unas altas correlaciones, lo que indica una elevada fiabilidad interjueces. Esta convergencia correlacional no significa que todos los evaluadores operen en la misma escala, ya que se encuentran diferencias estadísticamente significativas entre las medias de las puntuaciones asignadas por los evaluadores. Los resultados aportan datos muy importantes sobre los problemas implicados en la evaluación de la expresión oral en inglés. El estudio muestra la viabilidad técnica de examinar la expresión oral en inglés, si bien quedan aún bastantes aspectos por precisar. Así por ejemplo, los examinadores deberían ser entrenados específicamente en el material de evaluación y en las guías de calificación para minimizar el efecto diferencial de los evaluadores. Finalmente se discuten los resultados y se proponen nuevas líneas de investigación.

Palabras clave: Prueba de Acceso a la Universidad (PAU), expresión oral, Inglés, fiabilidad interjueces, validez.

Abstract

A new law enacted in Spain in 2008 makes it mandatory to assess oral expression in a foreign language as part of the standard university admissions examination (PAU), but no such test has been instituted to date. The Spanish National Institute for Educational Evaluation (INEE) has conducted a pilot study to begin outlining the implementation of such a test. Within this framework, the main goal of this research is to study the psychometric behaviour of the testers assigned to administer the oral test. A sample of 1194 students from seven Spanish autonomous communities is used. The sample is a mean of 18.04 years old and 57.4 percent women. Each student is evaluated by three teachers. Teachers are organized into thirty committees, so a total of ninety teachers are involved in the study. Data are analyzed using a matrix composed of all the students assessed by the same committee. The results show that the scores students earn from different teachers are highly correlated, which indicates good inter-rater reliability. High correlations between teacher scores do not mean that all teachers are using the same scale, however, since statistically significant differences are found between mean scores. The results of this study provide important data on the problems involved in assessing oral expression in the English language. This research demonstrates the technical viability of assessing oral expression in English, although several issues do remain unresolved. For example, testers should be specifically trained in the assessment process and scoring criteria in order to

minimize the differential effect of assessment by multiple testers. The results are discussed and new lines of research are suggested.

Key words: university admissions examination, oral expression, English language, inter-rater reliability, validity.

Contexto normativo del estudio

La Ley Orgánica 2/2006, de 3 de mayo, de Educación (LOE) dedica su artículo 38 a la Prueba de Acceso a la Universidad (PAU). En él se señala, entre otras cuestiones, que la PAU deberá adecuarse al currículo del Bachillerato y que corresponde al Gobierno, previa consulta a las comunidades autónomas, establecer las características básicas de la prueba. Desarrollando el citado artículo de la LOE, el Real Decreto 1892/2008, de 14 de noviembre, regula las condiciones para el acceso a las enseñanzas universitarias. En el artículo 9.3 introduce, por primera vez en la ordenación educativa española, un ejercicio que contempla la evaluación de la expresión oral en Lengua Extranjera en la PAU. Ahora bien, por la propia novedad que supone la introducción de este tipo de ejercicio, la disposición final segunda del Real Decreto 1892/2008 indica que antes de la implantación de la prueba oral se realizará un estudio para determinar sus características y tomar las decisiones oportunas sobre su configuración final.

Respondiendo al mandato legal, la Conferencia Sectorial de Educación decidió, a principios de 2012, realizar una serie de estudios destinados a explorar las posibilidades de integración de la expresión oral en lengua extranjera en la PAU. A finales del curso 2011-12 se llevó a cabo el primer estudio piloto en el que participaron siete comunidades autónomas: Aragón, Asturias, Baleares, Castilla-La Mancha, Comunidad Valenciana, Madrid y La Rioja. La finalidad última de dicho estudio era valorar la metodología de evaluación y el formato y características de la prueba, dejando para ensayos posteriores el estudio de los aspectos organizativos y de su integración en la PAU.

Fecha de entrada: 01-02-2013 Fecha de aceptación: 27-09-2013

Aproximación a la evaluación de la expresión oral en inglés

Vivimos en un mundo cada vez más multicultural y plurilingüe, por lo que todo lo relativo a la evaluación del lenguaje y a la traducción de unas lenguas a otras es un ámbito de estudio cada vez más consolidado (Muñiz, Elosua y Hambleton, 2013). Existen actualmente diversos sistemas y enfoques para la evaluación de las competencias en un idioma. En el caso del inglés, entre los más extendidos por todo el mundo cabe señalar el International English Language Testing System (IELTS), las University of Cambridge ESOL Examinations (Cambridge ESOL) o el Test of English as a Foreign Language (TOEFL), por citar solo tres ejemplos bien conocidos. Todos ellos tienen en común la evaluación de al menos cuatro competencias básicas en el dominio de la lengua inglesa: comprensión lectora, comprensión oral, expresión escrita y expresión oral. La evaluación de las destrezas de expresión (oral o escrita) supone una gran complejidad debido sobre todo al alto nivel de subjetividad que conlleva su medición. En cualquier caso, la evaluación de cualquier tipo de lenguaje obliga a establecer unos criterios de corrección claros y uniformes y a entrenar a los evaluadores para el manejo de dichos criterios.

Además, la evaluación de la expresión oral conlleva una dificultad añadida motivada por el propio formato del examen (un producto hablado) y por la necesidad de grabar la producción oral para su posterior revisión, cuestión bastante plausible en una prueba con alto impacto en el futuro académico de los aspirantes. En los últimos años se han propuesto distintos sistemas de evaluación de la expresión oral en inglés, en función de su frecuencia de uso cabe destacar principalmente tres métodos: pruebas automatizadas, pruebas semidirectas y entrevistas cara a cara.

Las pruebas automatizadas consisten en sistemas de reconocimiento de voz y módulos de cálculo que puntúan tanto la fluidez verbal como otros indicadores de expresión oral, tal y como hacen el Pearson Test of English (Pearson, 2009) o el TOEFL (Zechner, Higgins, Xi y Williamson, 2009). Las pruebas semidirectas consisten en el uso de ordenadores para presentar tareas y capturar las respuestas, si bien la puntuación es asignada por un evaluador humano, tal y como hacen la Simulated Oral Proficiency Interview (Lee, 2007), o el Videoconferenced Speaking Test (Kim, Craig y Seoul, 2012). En cuanto a la entrevista cara a cara, consiste en una interacción directa entre uno o dos estudiantes y uno o varios evaluadores. Este método lo utilizan entidades de reconocido prestigio como Cambridge ESOL (Macqueen y Harding, 2009) o IELTS (Brown, 2006).

Fecha de entrada: 01-02-2013 Fecha de aceptación: 27-09-2013

La aplicación automatizada tienes sus bondades. Así, los estudios de validez concurrente muestran que las puntuaciones obtenidas mediante pruebas automatizadas están altamente correlacionadas con las obtenidas mediante entrevistas orales (Bernstein, Van Moere y Cheng, 2010). En este sentido hay propuestas recientes para la PAU que abogan por una progresiva informatización de la prueba hacia sistemas automatizados de evaluación (Magal-Royo y Giménez López, 2012; Martín-Monje, 2012). Las principales ventajas de una prueba oral automatizada son la reducción de costes, el incremento en el número de destrezas en la evaluación, una mayor precisión en la medida de ciertas habilidades como la oral o la escrita (García Laborda, 2012), mayor velocidad y objetividad, nuevas posibilidades de administración (presentación aleatoria de ítems, etc.), la posibilidad de grabar la interacción o la disminución de la ansiedad entre los alumnos (Bueno Alastuey y Luque Agulló, 2012).

No obstante lo dicho, existen dos posibles justificaciones que permiten sustentar la idea de que la entrevista oral cara a cara pueda ser el método más eficaz. La primera es que, a pesar de que las pruebas automatizadas aparentemente evalúan de forma fiable parte de la expresión oral (fluidez y pronunciación), no es posible evaluar de forma completa el constructo de la forma en que lo hacen uno o varios evaluadores humanos (Bridgeman, Powers, Stone y Mollaun, 2011). Como exponen estos autores, es razonable pensar que una respuesta correctamente articulada pero sin sentido consiga una alta puntuación en un sistema automatizado y, sin embargo, no se estaría valorando la comprensión de la respuesta. En segundo lugar, las entrevistas cara a cara tienen mayor validez ecológica, ya que al configurarse como una conversación real simulan una verdadera interacción, es decir, un encuentro comunicativo donde los participantes (en este caso evaluador y evaluado) ejercen una influencia recíproca en sus acciones (Goffman, 1970). Este tipo de organización permite al examinador conocer el límite máximo de expresión oral del evaluado (Educational Testing Service, 1982, 11).

En función de todo lo expuesto, se entiende que el método más adecuado en términos de fiabilidad y validez para la evaluación de la expresión oral en inglés es la entrevista cara a cara. Sin embargo, esto no quiere decir que el método esté exento de problemas. Se pueden incluir diseños de puntuaciones que permitan mejorar su fiabilidad (Hill, Charalambous y Kraft, 2012). Siguiendo las propuestas realizadas para la expresión escrita, parece fundamental crear protocolos de actuación y

establecer criterios de evaluación, así como escalas de puntuación que sean fáciles de interpretar por los evaluadores (Díez-Bedmar, 2012). El método de la entrevista cara a cara suele contar con más de un evaluador (corrector o experto) con la intención de reducir la subjetividad de la evaluación y aumentar la fiabilidad en la medición. Sin embargo, no contar con un diseño de puntuación adecuado supondría una importante amenaza contra la fiabilidad entre evaluadores. En ese sentido, las propuestas más novedosas plantean realizar la evaluación oral siguiendo los estándares del Marco Común Europeo de Referencia para las Lenguas (MCER) (Amegual y Méndez García, 2012). En este tipo de pruebas, la fiabilidad de la evaluación normalmente se refiere a la fiabilidad de los jueces (Hamp-Lyons, 2007). Desde un punto de vista metodológico, existen diferentes formas de abordar el estudio de la fiabilidad entre evaluadores o fiabilidad interjueces. Es decir, se han propuesto multitud de indicadores para estudiar el grado en que las puntuaciones de los evaluadores están libres de error (Shoukri, 2004). En concreto, las dos principales fuentes de error son la definición de la escala de evaluación y la seguridad de que los jueces entienden y saben aplicar la escala (Weigle, 2002).

Algunos de los indicadores de fiabilidad interjueces más utilizados son el coeficiente kappa de Cohen, la correlación bivariada, la correlación intraclase o la teoría de la generabilidad (Gwet, 2012; Xiaomin y Houcan, 2005). Uno de los aspectos que se debe considerar para la selección de un indicador de fiabilidad interjueces es el nivel métrico de las puntuaciones (Abad, Olea, Ponsoda y García, 2011). De esta forma, el coeficiente kappa de Cohen permite estudiar el nivel de concordancia de dos jueces en variables nominales u ordinales (Cohen, 1960). El coeficiente de correlación intraclase permite estimar el grado de acuerdo sobre variables cuantitativas y se basa en un modelo ANOVA de medidas repetidas donde se admiten datos en una variable independiente intrasujeto (por ejemplo, diferentes ítems o diferentes jueces) (Abad et ál., 2011). Además, tal y como describen Abad et ál. (2011), se trata de un caso particular de los coeficientes de generalizabilidad. Una ventaja de este planteamiento es que permite realizar diferentes estimaciones atendiendo al modelo lineal utilizado (a saber: fijo, aleatorio o mixto en función de la aleatoriedad de la muestra) (Shoukri, 2004). En cualquier caso, la interpretación de una cuantía concreta suele ser parecida a la del coeficiente kappa (Abad et ál., 2011). Para una revisión en profundidad sobre la fiabilidad interjueces léase Gwet (2012).

Un buen ejemplo de la actualidad de esta línea de investigación son las recientes aproximaciones que se realizan desde la metodología observacional y los protocolos *think-aloud* (Crisp, 2012), la metodología cualitativa para el estudio de la cognición del evaluador (Suto, 2012) o la aplicación de los modelos de teoría de respuesta a los ítems para comparar las puntuaciones asignadas por los evaluadores (Eckes, 2009; Prieto, 2011; Wolfe y McVay, 2012). Sin embargo, a pesar de los importantes avances en esta línea de investigación, aún quedan cuestiones por resolver, por ejemplo, no hay una opinión unánime sobre cuál es el número óptimo de evaluadores/correctores, y si estos son intercambiables en sus papeles de evaluador y corrector (Myford, 2012).

En este contexto, el estudio que ahora se presenta tiene como finalidad añadir nuevas evidencias sobre la fiabilidad y validez de las puntuaciones asignadas por los evaluadores en un ejercicio de expresión oral de inglés en la PAU. Para lograrlo se proponen cuatro objetivos fundamentales: a) estimar la asociación que existe entre las dos partes de la prueba realizada (véase el apartado Instrumentos); b) estudiar la validez de constructo de la prueba, tratando de delimitar cuántos aspectos de la expresión oral del inglés han de ser evaluados; c) analizar la fiabilidad interjueces, o grado de convergencia entre las puntuaciones asignadas por los distintos evaluadores; y d) comprobar si existen diferencias a la hora de evaluar en función del papel de los examinadores, bien sean evaluadores o calificadores.

Método

Muestra

Se utilizó un muestreo en dos etapas similar al usado en la evaluación de sistemas educativos (Joncas y Foy, 2012; OCDE, 2008, 2009, 2012; Shettle et ál., 2008). En la primera etapa se seleccionaron ocho centros en cada comunidad autónoma. En esta etapa se empleó un diseño muestral de tres estratos (tamaño de la población, titularidad del centro y si era un centro bilingüe o no) y los centros se eligieron mediante un procedimiento

aleatorio y sistemático donde la probabilidad de elección de cada centro era proporcional a su tamaño. En la segunda etapa del muestreo se eligieron aleatoriamente 25 estudiantes de cada centro seleccionado en la primera etapa. El muestreo descrito se planificó para obtener una muestra de 1.400 estudiantes (200 por cada comunidad autónoma). Finalmente se consiguieron 1.194 estudiantes, es decir, el 85,3% de la muestra planificada. La media de edad de los estudiantes era de 18,04 años y el 57,4% eran mujeres.

Instrumentos

Se estableció que el nivel promedio de la prueba de expresión oral en lengua inglesa debería aproximarse al nivel B1 definido en el Marco Común Europeo para las Lenguas. Esta decisión se tomó teniendo en cuenta que en los estudios realizados sobre expresión escrita en la PAU se había encontrado este nivel (Díez-Bedmar, 2012) y también las últimas propuestas ya mencionadas de los expertos en expresión oral (Amengual y Méndez García, 2012). Sobre este supuesto se desarrollaron las especificaciones de la prueba, el material de evaluación y los criterios de valoración.

La prueba se organizó como una conversación cara a cara entre el alumno y el examinador, con una duración de ocho a 10 minutos, dividida en dos partes; la primera (entre dos y tres minutos) consistía en una conversación inicial con el examinador sobre información personal del estudiante (nombre, familia, estudios, gustos, tiempo libre, etc.). Para estructurar la interacción el examinador disponía de una serie de preguntas organizadas en bloques temáticos. El objetivo de esta primera parte era generar confianza en el alumno y que este se adaptase al contexto de la prueba. La segunda parte (entre seis y siete minutos) pretendía profundizar en la evaluación de la competencia de expresión oral del estudiante. En este caso se les pedía a los alumnos que realizaran una descripción y comentario de unos estímulos fotográficos y que respondieran a preguntas relacionadas con ellos. Al igual que en la primera parte, el examinador disponía de un listado orientativo de preguntas que podía incluir en la evaluación.

En total se emplearon 16 estímulos fotográficos organizados en cuatro temas (deportes, música, nuevas tecnologías y naturaleza). Cada estudiante

debía responder a uno de los 16 estímulos. Dentro de cada centro los estímulos se presentaron a los alumnos siguiendo un procedimiento en espiral *(spiraling method)* para asegurar que todos los estímulos se empleaban en aproximadamente las mismas ocasiones.

Para calificar cada intervención se disponía de una ficha de calificación que distinguía cinco criterios: alcance, corrección gramatical, fluidez, interacción y coherencia, extraídos del MCER (Ministerio de Educación, Cultura y Deporte, 2002). El desempeño en los cinco criterios se valoraba en una escala de 1 a 10 puntos. Además, cada criterio fue calificado doblemente, ya que se registraba una nota para la primera parte de la prueba y otra para la segunda parte.

Procedimiento

Cada conversación individual fue calificada por un tribunal o equipo evaluador de tres personas. Los evaluadores eran profesores de Bachillerato en activo de la especialidad de Inglés seleccionados por cada comunidad autónoma participante. El número de tribunales necesarios fue determinado por cada Administración educativa. Todas las comunidades autónomas utilizaron entre tres y cuatro tribunales, salvo Baleares que, debido a la insularidad, empleó siete equipos. En total participaron en la prueba piloto 90 profesores que conformaron 30 tribunales o equipos evaluadores.

Cada Administración educativa se encargó de organizar e impartir las sesiones de formación a los evaluadores reclutados. El entrenamiento consistió en presentar las finalidades del estudio, revisar el material de aplicación y trabajar y reflexionar sobre los criterios de corrección y calificación.

En el momento de la aplicación, uno de los miembros del tribunal tomaba el papel de entrevistador (evaluador) y era el encargado de mantener la conversación con el alumno y de presentar los estímulos de la evaluación. Los otros dos miembros funcionaban como observadores no participantes (calificadores). A lo largo de la sesión, las tres personas del tribunal fueron variando su papel dentro del tribunal, de tal forma que cada miembro tomó el papel de evaluador en un tercio de las aplicaciones y de calificador en los dos tercios restantes. Los tres miembros del tribunal calificaron a cada alumno de forma independiente y evitaron compartir las valoraciones y puntuaciones otorgadas.

Fecha de entrada: 01-02-2013 Fecha de aceptación: 27-09-2013

Análisis de los datos

Los análisis se llevaron a cabo utilizando la matriz formada por los alumnos evaluados por el mismo tribunal evaluador. De esta forma, la base de datos completa se subdivide en 30 matrices correspondientes a los 30 tribunales. En primer lugar, se analizó la distribución de los datos para estudiar su ajuste a una curva normal. A continuación, se especifican los análisis realizados en relación con los objetivos del estudio.

Para estudiar el primer objetivo, es decir, conocer el grado de concordancia entre las puntuaciones de la primera y la segunda parte del examen, se utilizó la correlación de Pearson.

El segundo objetivo, estudiar la dimensionalidad de los distintos criterios evaluados, se dividió aleatoriamente la muestra total en tres submuestras ($N1 = 389$; $N2 = 417$; $N3 = 388$). Con la primera submuestra se realizó un análisis de componentes principales. Con la segunda submuestra se realizó un análisis factorial confirmatorio (AFC) correlacionando los errores de medida (Byrne, 2001; Abad et ál., 2011). A continuación, se realizó otro AFC con la tercera submuestra sin modificar el modelo propuesto en la primera con la intención de hacer validación cruzada (Pérez-Gil, Chacón Moscoso y Moreno Rodríguez, 2000). El método de estimación utilizado fue el de máxima verosimilitud robusta porque mostró un mejor ajuste a los datos. La evaluación de la bondad de ajuste de los datos al modelo fue determinada en función del índice de ajuste comparativo (CFI) y la media cuadrática estandarizada de los residuales (SRMR).

En tercer lugar, para estimar la concordancia entre correctores se calculó el coeficiente de correlación de Pearson entre las puntuaciones asignadas por los tres correctores de cada tribunal. Adicionalmente, como indicador de la fiabilidad interjueces y de la relevancia de las correlaciones anteriormente obtenidas (esto es, el tamaño del efecto) se obtuvo el coeficiente de correlación intraclase para cada tribunal. La razón de utilizar este indicador y no el coeficiente kappa es que resulta más adecuado para el nivel métrico utilizado y el hecho de que, en cualquier caso, ambos son comparables (Abad et ál., 2011). Por otro lado, para estudiar las diferencias entre las puntuaciones medias otorgadas por los evaluadores, se realizó un análisis de la varianza (ANOVA) de medidas repetidas –un factor intrasujeto– puesto que cada alumno dispone de tres puntuaciones correspondientes a los tres jueces de cada tribunal.

Fecha de entrada: 01-02-2013 Fecha de aceptación: 27-09-2013

Finalmente, para estudiar el cuarto objetivo, es decir, el efecto que tiene el papel del profesor (evaluador o calificador) se realizó un ANOVA de medidas repetidas, pero esta vez, incluyendo como medida intersujeto la variable de agrupamiento papel del profesor.

Resultados

Estudio de la normalidad de la distribución de puntuaciones

En primer lugar, tanto la prueba de Kolmogorov-Smirnov como la de Shapiro-Wilk resultaron estadísticamente significativas ($p < ,001$) para todas las variables. Por lo tanto, no se puede aceptar la hipótesis nula de que los datos se ajustan a una distribución normal. Sin embargo, los estadísticos de asimetría y curtosis mostraron que, a pesar de tener apuntamiento, todas las distribuciones son simétricas (NC = 95%). En cualquier caso, los valores de los estadísticos de asimetría y curtosis estuvieron comprendidos entre -1 y 1. Teniendo en cuenta los límites de estos estadísticos cuando se utilizan muestras reales (Blanca, Arnau, López-Montiel, Bono y Bendayan, 2013), se podría considerar que estos valores no se alejan excesivamente de lo esperable en una distribución normal. Además, parece claro que pruebas estadísticas como el ANOVA son lo suficientemente robustas a violaciones de normalidad (Glass, Peckham y Sanders, 1972; Harwell, Rubinstein, Hayes y Olds, 1992; Lix, Keselman y Keselman, 1996; Schmider, Zieglel, Danay, Beyer y Bürher, 2010). Por ello, y con la intención de facilitar la interpretación de los resultados se utilizaron pruebas paramétricas.

Convergencia entre ambas partes del examen, así como entre los criterios de calificación

En relación con el primer objetivo se ha podido comprobar que la correlación entre las dos partes de la prueba fue muy elevada ($r_{xy} = ,98$). Además, no solo se comprobó que cada criterio de calificación de una parte

Fecha de entrada: 01-02-2013 Fecha de aceptación: 27-09-2013

correlacionaba de modo casi perfecto con el mismo criterio de la otra parte, sino que también se constató que los cinco criterios de calificación utilizados mostraban una alta convergencia, ya que las correlaciones entre todos ellos se encuentran entre ,91 y ,95. Este último dato ya adelanta que la escala puede ser esencialmente unidimensional.

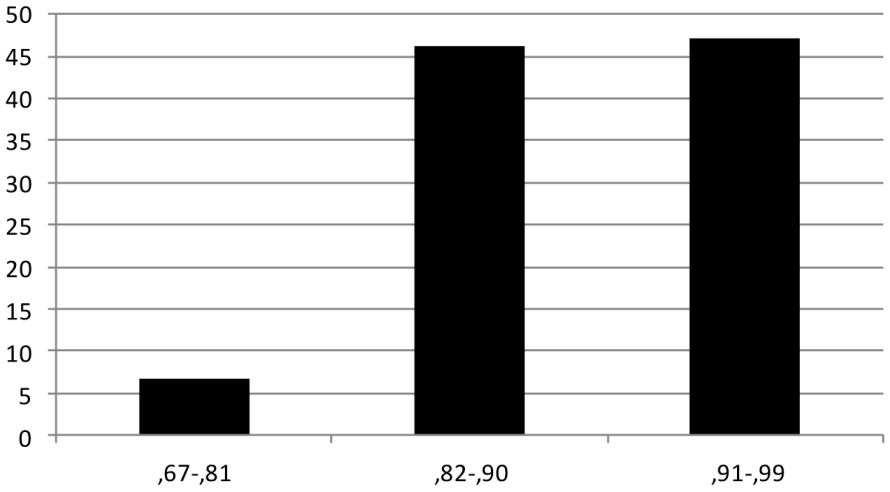
Estudio de la dimensionalidad del constructo

La medida de adecuación muestral KMO fue ,90 y la prueba de esfericidad de Bartlett fue estadísticamente significativa (p < ,001). El análisis de componentes principales (N1 = 389) indicó que las puntuaciones de todos los criterios pueden reducirse a un único valor que explicaría el 96% de la varianza total. Además, el peso de los cinco criterios en el factor es casi idéntico. El AFC realizado en la segunda submuestra (N2 = 417) presentó un buen ajuste de los datos al modelo (CFI = ,99; SRMR = ,004). Además, estos resultados se confirman en la segunda submuestra (N3 = 388) mediante validación cruzada (CFI = ,99; SRMR = ,005). Por todo ello, en función de los criterios establecidos en Hu y Bentlet (1999), los datos se ajustan adecuadamente a una estructura unidimensional.

Estudio de la fiabilidad interjueces

En relación con el tercer objetivo se han realizado dos análisis. Los resultados del primero se muestran en el Gráfico 1, donde se presenta la distribución del porcentaje de correlaciones entre los correctores organizadas en tres rangos (,67-,81; ,82-,90; ,91-,99). Todas las correlaciones son estadísticamente significativas al nivel de confianza del 99% (bilateral). La correlación mínima es de ,67 y la máxima de ,99; la correlación media es de ,89, y el 93,3% de las correlaciones están comprendidas entre ,82 y ,99. Todo ello indica que las puntuaciones asignadas por los distintos profesores mantienen una alta convergencia. Además, en función de la tabla de equivalencias entre el tamaño del efecto y las correlaciones expuesta en Becker (2000), a correlaciones superiores a ,67 les correspondería una d de Cohen superior a 1,8. Por lo tanto, las correlaciones encontradas serían relevantes (Cohen, 1988).

GRÁFICO I. Porcentaje por rango de correlaciones entre correctores



Además, en la Tabla 1 se presenta la estimación de la fiabilidad interjueces obtenida mediante el coeficiente de correlación intraclase (C_{CI}) para evaluar tanto la consistencia como el acuerdo absoluto entre los jueces. En primer lugar, se informa del grado en que los diversos jueces ordenan de la misma forma a los sujetos (el grado en que son consistentes) y resulta un valor idéntico al coeficiente α de Cronbach. En segundo lugar, el C_{CI} informa del nivel de acuerdo absoluto en las calificaciones que asignan (Abad et ál., 2011). Todos los coeficientes fueron estadísticamente significativos ($p < ,001$) Como se puede observar, todos los valores son superiores a ,90 y, por tanto, existe una alta fiabilidad interjueces.

Fecha de entrada: 01-02-2013 Fecha de aceptación: 27-09-2013

TABLA I. Fiabilidad interjueces por tribunal

Tribunal	α	C_{CI}	Tribunal	α	C_{CI}
1 (N = 24)	,987	,985	16 (N = 25)	,949	,944
2 (N = 63)	,952	,917	17 (N = 71)	,961	,961
3 (N = 35)	,990	,986	18 (N = 50)	,984	,984
4 (N = 17)	,965	,962	19 (N = 48)	,960	,958
5 (N = 25)	,983	,983	20 (N = 25)	,956	,927
6 (N = 25)	,927	,927	21 (N = 25)	,931	,931
7 (N = 25)	,961	,960	22 (N = 25)	,969	,968
8 (N = 50)	,956	,954	23 (N = 50)	,972	,970
9 (N = 48)	,912	,866	24 (N = 37)	,955	,955
10 (N = 48)	,944	,938	25 (N = 45)	,993	,993
11 (N = 31)	,941	,921	26 (N = 40)	,964	,964
12 (N = 27)	,963	,958	27 (N = 47)	,949	,946
13 (N = 34)	,899	,898	28 (N = 41)	,981	,977
14 (N = 48)	,947	,939	29 (N = 49)	,978	,977
15 (N = 66)	,977	,977	30 (N = 50)	,933	,891

(*) Nota: α = coeficiente de correlación intraclase de consistencia; C_{CI}= coeficiente de correlación intraclase de acuerdo absoluto.

El segundo análisis para estudiar el grado de convergencia entre jueces se muestra en la Tabla II. En ella se recogen los resultados del ANOVA de medidas repetidas (un factor intrasujeto). Como se puede observar, en 14

de los 30 tribunales existen diferencias estadísticamente significativas al nivel de confianza del 95%. Sin embargo, en función del tamaño del efecto y de la potencia de la prueba (Cohen, 1988) solamente nueve de esas 14 diferencias encontradas resultarían relevantes (d > ,2) y tendrían una alta probabilidad de rechazar la hipótesis nula si fuera falsa (p > ,7). En concreto, se trata de los tribunales 1, 2, 5, 6, 9, 16, 21, 25 y 30.

TABLA II. Diferencia de medias entre correctores del mismo tribunal

Tribunal	F	Sig.	Tamaño del efecto	Potencia	Tribunal	F	Sig.	Tamaño del efecto	Potencia
1 (N = 24)	9,48	,005	,292	,838	16 (N = 25)	9,933	,004	,293	,856
2 (N = 63)	27,95	,000	,311	,999	17 (N = 71)	5,935	,017	,078	,671
3 (N = 35)	1,853	,182	,052	,263	18 (N = 50)	,979	,327	,020	,163
4 (N = 17)	1,000	,332	,059	,156	19 (N = 48)	10,23	,002	,179	,88
5 (N = 25)	6,819	,015	,221	,707	20 (N = 25)	3,291	,082	,121	,414
6 (N = 25)	6,742	,002	,348	,93	21 (N = 25)	7,064	,014	,227	,723
7 (N = 25)	,001	,976	,000	,05	22 (N = 25)	5,576	,027	,189	,62
8 (N = 50)	1,3	,26	,026	,201	23 (N = 50)	4,63	,036	,086	,559
9 (N = 48)	,007	,000	,857	1	24 (N = 37)	,81	,373	,022	,142
10 (N = 48)	2,015	,162	,041	,285	25 (N = 45)	23,47	,000	,348	,997
11 (N = 31)	1,100	,303	,035	,174	26 (N = 40)	2,78	,103	,067	,37
12 (N = 27)	1,196	,284	,044	,184	27 (N = 47)	2,978	,091	,061	,394
13 (N = 34)	1,767	,193	,051	,252	28 (N = 41)	1,586	,215	,038	,233
14 (N = 48)	7,762	,008	,142	,779	29 (N = 49)	4,125	,048	,079	,512
15 (N = 66)	2,506	,118	,037	,345	30 (N = 50)	34,26	,000	,411	1

Fecha de entrada: 01-02-2013 Fecha de aceptación: 27-09-2013

Estudio del efecto que tiene el papel del profesor (corrector o evaluador)

La Tabla III muestra los resultados del ANOVA de medidas repetidas donde la medida intersujeto es la variable papel del profesor (corrector o evaluador). En este caso, hay dos tribunales de los que se desconoce el papel del profesor y, por tanto, las comparaciones de medias se realizaron con 28 tribunales. A la vista de los resultados (Tabla III), solo se encuentran diferencias estadísticamente significativas en dos casos. Además el tamaño del efecto es prácticamente nulo y la potencia de la prueba muy baja (Cohen, 1988).

TABLA III. Diferencias de medias en función del papel del profesor (examinador o corrector)

Tribunal	F	Sig.	Tamaño del efecto	Potencia	Tribunal	F	Sig.	Tamaño del efecto	Potencia
1 (N = 24)	,422	,523	,020	.095	16 (N = 25)	,004	,947	,000	,050
2 (N = 63)	,900	,347	,015	,154	17 (N = 71)	,192	,662	,003	,072
3 (N = 35)	,588	,449	,018	,115	18 (N = 50)	4,939	,031	,095	,586
4 (N = 17)	,199	,662	,014	,070	19 (N = 48)	,000	,999	,000	,050
5 (N = 25)	1,382	,252	,059	,203	20 (N = 25)	1,666	,210	,070	,235
6 (N = 25)	,621	,439	,027	,117	21 (N = 25)	1,975	,174	,082	,270
7 (N = 25)	,148	,704	,007	,066	23 (N = 50)	,238	,628	,005	,077
8 (N = 50)	,491	,487	,010	,105	24 (N = 37)	,625	,435	,018	,120
9 (N = 48)	,215	,645	,005	,074	25 (N = 45)	1,559	,219	,036	,230
10 (N = 48)	,008	,928	,000	,051	26 (N = 40)	,261	,613	,007	,079
12 (N = 27)	,942	,341	,038	,154	27 (N = 47)	,138	,712	,003	,065
13 (N = 34)	,037	,049	,001	,054	28 (N = 41)	,827	,369	,021	,144
14 (N = 48)	,001	,973	,000	,050	29 (N = 49)	,628	,432	,013	,121
15 (N = 66)	1,299	,259	,020	,202	30 (N = 50)	,000	,984	,000	,050

Fecha de entrada: 01-02-2013 Fecha de aceptación: 27-09-2013

Discusión y conclusiones

El estudio piloto ha demostrado la viabilidad técnica de examinar la expresión oral en lengua extranjera (inglés), sin embargo quedan aún bastantes aspectos por precisar. En primer lugar, es necesario replicar esto mismo para el resto de las lenguas extranjeras que serán evaluadas en la PAU. También habría que pensar en las posibles adaptaciones necesarias para los alumnos con necesidades educativas derivadas fundamentalmente de deficiencias auditivas (hipoacusias y sorderas profundas). En relación con el acuerdo interevaluadores también hay margen para seguir investigando, y sería necesario modificar el diseño para la configuración de los calificadores. Esto supondría que los tribunales no pueden ser fijos, sino que habría que emplear un diseño matricial que permita combinar los examinadores en diferentes tribunales (Fernández-Alonso y Muñiz, 2011; Fernández-Alonso, Suárez-Álvarez y Muñiz, 2012).

Si bien, desde el punto de vista técnico la evaluación es viable, sería necesario valorar si desde el punto de vista logístico también lo es. Nótese que rebajando el tiempo de examen a unos cinco o seis minutos, cada tribunal podría evaluar unos 10 estudiantes por hora. Es decir, una jornada intensiva supondría evaluar unos 80 estudiantes por tribunal. Incluso para evaluar los miles de estudiantes de una comunidad autónoma pequeña, serían necesarios decenas de tribunales que además habría que combinar adecuadamente si se pretende equilibrar los posibles efectos del tribunal. Finalmente, los examinadores deberían ser entrenados específicamente en el material de evaluación y en las guías de calificación para minimizar el efecto diferencial de los evaluadores.

Desde un punto estrictamente legal, el Real Decreto 1892/2008 establece la evaluación de la expresión oral en lengua extranjera en la PAU. Por tanto, queda fuera de estas conclusiones debatir sobre la oportunidad o conveniencia de incluir la mencionada evaluación en la PAU. Sin embargo, el formato actual del examen de lengua extranjera en el acceso a la universidad no incluye la evaluación de la expresión oral de inglés. Esto supone una grave limitación en la medida que incide negativamente en la validez consecuencial de la prueba, pues condiciona la enseñanza del idioma en el Bachillerato. Es decir, los profesores enseñan en función de cómo se va a evaluar en la PAU, y si no se incluye prueba oral no harán hincapié en este aspecto. Varios autores han cuestionado el constructo, la validez y la fiabilidad de la PAU en general (Amengual, 2006; Herrera Soler,

Fecha de entrada: 01-02-2013 Fecha de aceptación: 27-09-2013

1999; Sanz, 1999), si bien sus trabajos han tenido una repercusión limitada en las administraciones responsables de la prueba (Bueno Alastuey y Luque Agulló, 2012). Además, esta evaluación no se corresponde con los objetivos que el currículo establece para el Bachillerato (Amengual, 2006), y está sesgado hacia los aspectos gramaticales, léxicos y la compresión y expresión escrita (Herrera Soler, 1999; Bueno Alastuey y Luque Agulló, 2012). Todo ello sin olvidar, como ya se ha señalado, que dicho sesgo acarrea efectos indeseados en los contenidos y metodología del Bachillerato, mientras que los aspectos orales quedan relegados en favor de los contenidos evaluados en la PAU (Amengual, 2006, 2010; Amengual y Méndez García, 2012; Bueno Alastuey y Luque Agulló, 2012). Todo ello reclama la necesidad urgente de incluir la evaluación de los aspectos orales de la lengua extranjera en cualquier proceso de acceso a las enseñanzas universitarias.

En este contexto, el objetivo general de este trabajo fue añadir nuevas evidencias sobre la fiabilidad y la validez de las puntuaciones asignadas por los evaluadores en una prueba piloto de expresión oral de inglés de la PAU. Aceptando pues que la evaluación de la expresión oral en lengua extranjera en la PAU es irrenunciable, en lo que sigue se ofrecerán algunas orientaciones que, a la vista de los resultados, sirvan para perfilar la organización de la prueba.

Respecto al primer objetivo, los resultados mostraron que existe una alta convergencia entre ambas partes del examen, así como entre los cinco criterios de calificación utilizados en la evaluación (alcance, corrección gramatical, fluidez, interacción y coherencia). Esta alta convergencia parece indicar que puntuar por separado ambas partes es, hasta cierto punto, redundante. Esta evidencia permitiría reducir el examen a una sola parte o bien limitar el tiempo de aplicación de las dos partes sin que por ello se viera afectada la precisión de la estimación. Al fin, las calificaciones en la primera parte (que, recordemos, ocupa entre dos y tres minutos) ya parecen un predictor fiable, tanto de las calificaciones de la segunda parte como de la puntuación final.

Respecto al segundo objetivo –la validez de constructo de la prueba–, tanto el análisis de componentes principales, como el análisis factorial confirmatorio han mostrado un buen ajuste de los datos a una estructura unidimensional y han llegado a explicar el 96% de la varianza. Además, los cinco criterios contribuyen al factor único de forma similar. Este dato permitiría defender el uso de una calificación sintética y global para el

conjunto de la prueba. Sin embargo, tanto por coherencia con el MCER como por una cuestión de validez aparente, parece más adecuado seguir manteniendo un sistema de calificación analítico, en el que cada criterio se puntúe por separado y, de este modo, estimar la nota final como la media aritmética de los cinco criterios de calificación.

Como es habitual en este tipo de pruebas, la estimación de la fiabilidad (tercer objetivo) se realizó evaluando el grado de convergencia entre las puntuaciones asignadas por los distintos evaluadores (Hamp-Lyons, 2007). A la vista de los resultados, por un lado, se encuentra una alta correlación entre las puntuaciones asignadas por los evaluadores (Gráfico I), así como una alta fiabilidad interjueces para todos los tribunales (Tabla I); por otro, las medias de los evaluadores difieren en gran medida (Tabla II). Esto parece indicar que, si bien las puntuaciones asignadas por los evaluadores tienen una alta correlación, ello no necesariamente significa que todos operen en la misma escala, ya que se encuentran diferencias estadísticamente significativas entre las medias de las puntuaciones asignadas en nueve de los 30 tribunales. En función de las diferencias de medias encontradas, parece que los correctores utilizan niveles de exigencia distintos. El hecho de que los jueces puedan variar en benevolencia o severidad no es nada nuevo (Kondo-Brown, 2002; McNamara, 1996; Weigle, 1998). Sin embargo, no emplear el mismo rasero tiene implicaciones prácticas muy importantes a la hora de evaluar a los alumnos de forma equitativa. Por ello, gran parte de las orientaciones de este trabajo se centran en proponer estrategias para mitigar el efecto del corrector.

Una de las acciones para minimizar este efecto es mejorar la definición de los criterios de corrección y calificación, lo que ayudará a homogeneizar la evaluación de los jueces. En este sentido parece necesario profundizar en el desarrollo de guías y rúbricas de corrección detalladas que desmenucen los criterios de corrección del MCER. Adicionalmente, los estudios piloto deberían servir para identificar producciones que fuesen representantes típicas de ciertos niveles de competencia de expresión oral (por ejemplo: insuficiente, adecuado, bueno, excelente) y que ayuden a ubicar a los miembros del tribunal en la escala decimal propuesta. A la vista de los resultados, parece necesario que todos los evaluadores reciban un entrenamiento intensivo. Estas sesiones de formación se centrarían en la revisión de los criterios del MCER y el análisis y reflexión sobre los criterios de corrección; además, podrían incluir, a modo de taller, el visionado de

evaluaciones reales y su calificación con el objetivo de acordar niveles de exigencia comunes. En definitiva, la elaboración de un protocolo de corrección más específico y un entrenamiento previo parecen ser dos de las vías para reducir las diferencias de medias encontradas entre las puntuaciones de los correctores (Díez-Bedmar, 2012).

Por otro lado, la literatura científica ha encontrado relación entre las variaciones en las calificaciones de los correctores y factores como el efecto del uso de escalas diferentes, los años de experiencia del evaluador, su bagaje académico, la edad o el sexo (Vaughan, 1991; Weigle, 2002). Una forma de controlar el efecto de estas características de los evaluadores sería aplicar los principios del diseño experimental para distribuir los evaluadores dentro de los tribunales. Esto supondría que los tribunales no pueden ser fijos, sino que habría que emplear un diseño matricial que permita rotar y combinar los evaluadores en diferentes tribunales. En la evaluación educativa los diseños matriciales se emplean, por ejemplo, para equilibrar la dificultad de los ítems y de este modo controlar los efectos de exposición a los mismos y lograr estimaciones más fiables (Fernández-Alonso y Muñiz, 2011; Fernández-Alonso, Suárez-Álvarez y Muñiz, 2012). De igual modo, sería posible establecer un sistema rotatorio que asigne y distribuya evaluadores a tribunales de tal modo que se puedan controlar o neutralizar los sesgos debidos a las características de los primeros.

Los datos del análisis de fiabilidad entre jueces desaconsejan dejar la calificación de la prueba en manos de un solo corrector, ya que esto parece muy arriesgado, al menos desde el punto de vista psicométrico. En este sentido podría afirmarse que cuanto más numeroso fuese el tribunal, mayores serían las garantías de neutralizar el efecto de un evaluador demasiado benévolo o demasiado exigente. Sin embargo, por cuestiones logísticas tampoco esta orientación parece realista. Seguramente, tres miembros, tal y como se ha ensayado en este estudio piloto, puede ser un número adecuado. Si, por razones de viabilidad, los tribunales se conforman con dos personas siempre cabría la posibilidad de que las calificaciones se otorgaran por consenso entre los miembros con el fin de minimizar los riesgos de puntuaciones dispares.

La última recomendación para reducir el efecto del corrector sería la eliminación de puntuaciones divergentes. En una escala de 0 a 10 en la que se califica con números enteros, y con unos criterios de calificación claros y un entrenamiento adecuado, no debieran encontrarse distancias entre los miembros del tribunal de más de dos puntos. Para aquellos casos

en los que dos miembros de un mismo tribunal presentan valoraciones separadas por tres o más puntos debería preverse alguna forma de neutralizar la divergencia, como podría ser la eliminación de puntuaciones extremas, tal y como se hace en procedimientos de evaluación de similar naturaleza.

Finalmente, el cuarto objetivo fue comprobar si existen diferencias a la hora de evaluar en función del papel de los examinadores. Los resultados mostraron que no existen diferencias estadísticamente significativas (NC = 95%) entre el papel de evaluador y el de calificador (Tabla III). Ello permite aventurar que el evaluador, además de llevar el peso de la entrevista, puede también asignar puntuaciones a los aspirantes. Parece pues que la organización de los tribunales ensayada en este estudio piloto es adecuada. Por tanto, es posible mantener en futuras aplicaciones el sistema de rotación en el desempeño de los dos papeles. Lo contrario, es decir, hacer recaer el papel de evaluador siempre sobre el mismo miembro del tribunal supondrá un desgaste mayor para quien ejerza dicha función.

Referencias bibliográficas

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

Amengual, M. (2006). Análisis de la prueba de inglés de selectividad de la Universitat de les Illes Balears. *Ibérica,* 11, 29-59.

— (2010). Exploring the Washback Effects of a High-Stakes English Test. *Revista Alicantina de Estudios Ingleses,* 23, 149-170.

— y Méndez García, M. C. (2012). Implementing the Oral English Task in the Spanish University Admission Examination: An International Perspective of the Language. *Revista de Educación,* 357, 105-127.

Becker, L. A. (2000). *Basic and Applied Research Methods.* Colorado Springs (Colorado): Colorado University. Recuperado de: http://www.uccs.edu/~faculty/lbecker/default.htm (en Course Content: Part II, *Lecture Notes: Effect Size*).

Bernstein, J., Van Moere, A. y Cheng, J. (2010). Validating Automated Speaking Tests. *Language Testing,* 27 (3), 355-377. DOI: 10.1177/0265532210364404

Fecha de entrada: 01-02-2013 Fecha de aceptación: 27-09-2013

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R. y Bendayan, R. (2013). Skewness and Kurtosis in Real Data Samples. *Methodology,* 9 (2), 78-84. DOI: 10.1027/1614-2241/a000057

Bridgeman, B., Powers, D., Stone, E. y Mollaun, P. (2012). TOEFL iBT Speaking Test Scores as Indicators of Oral Communicative Language Proficiency. *Language Testing,* 29 (1), 91-108.

Brown, A. (2006). An Investigation of The Rating Process in the IELTS Oral Interview. En M. Milanovic y C. Weir (Eds. Col.) y L. Taylor y P. Falvey (Eds. Vol.), *Studies in Language Testing, Vol. 19. IELTS Collected Papers: Research in Speaking and Writing Assessments,* 316-377. Cambridge: Cambridge University Press. Citation.

Bueno Alastuey, M. C. y Luque Agulló, G. (2012). Foreign Language Competences Required in the University Admission Examination: A Proposal for the Evaluation of Oral Aspects. *Revista de Educación,* 357, 81-104.

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

Cohen, J. (1960). A Coefficient of Agreement for Nominal Tables. *Educational and Psychological Measurement,* 20, 37-46.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2.ª ed.). Hillsdale (Nueva Jersey): Lawrence Earlbaum Associates.

Crisp, V. (2012). An investigation of rater cognition in the assessment of projects. *Educational Measurement, 31*(3), 10-20.

Díez-Bedmar, M. B. (2012). The Use of the Common European Framework of Reference for Languages to Evaluate the Compositions in the English Exam in the University Entrance Examination. *Revista de Educación,* 357, 55-80.

Eckes, T. (2009). Many-Facet Rasch Measurement. En S. Takala (Ed.), *Reference Supplement to the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (sección H). Estrasburgo (Francia): Council of Europe, Language Policy Division.

Educational Testing Service (1982). *Oral Proficiency Testing Manual.* Princeton (Nueva Jersey): Educational Testing Service.

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

Fernández-Alonso, R., Suárez-Álvarez, J. y Muñiz, J. (2012). Imputación de datos perdidos en las evaluaciones diagnósticas educativas. *Psicothema,* 24 (1), 167-175.

García Laborda, J. (2012). Introduction. From Selectividad to the University Admission Examination: Past, Present and a Not-Very-Distant Future. *Revista de Educación,* 357, 17-27.

Glass, G. V., Peckham, P. D. y Sanders, J. R. (1972). Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance. *Review of Educational Research,* 42 (3), 237-288.

Goffman, E. (1970). *Ritual de la interacción. Ensayos sobre el comportamiento cara a cara.* Buenos Aires: Tiempo Contemporáneo.

Gwet, K. (2012). *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters* (3.ª ed.). USA: Advanced Analytics, LLC.

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

Harwell, M. R., Rubinstein, E. N., Hayes, W. S. y Olds, C. C. (1992). Summarizing Monte Carlo Results in Methodological Research: The One- and Two-Factor Fixed Effects ANOVA Cases. *Journal of Educational and Behaviorial Statistics,* 17 (4), 315-339.

Herrera Soler, H. (1999). Is the English Test in Spanish University Entrance Examination as Discriminating as it Should be? *Estudios Ingleses de la Universidad Complutense,* 7, 89-107.

Hill, H. C., Charalambous, C. Y. y Kraft, M. A. (2012). When Rater Reliability is no Enough: Teacher Observation Systems and a Case for the Generalizability Study. *Educational Researcher,* 41 (2), 56-64. DOI: 10.3102/0013189X12437203

Huntley, F. L., Palmer, E. J. y Wakeling, H. C. (2012). Validation of an Adaptation of Levenson's Locus of Control Scale with Adult Male Incarcerated Sexual Offenders. *Sex Abuse,* 24 (1), 46-63. DOI: 10.1177/1079063211403163

Joncas, M. y Foy, P. (2012). Sample Design in TIMSS and PIRLS. En Martin, M.O. y Mullis, I.V.S. (Eds.), *Methods and Procedures in TIMSS and PIRLS 2011*. Chestnut Hill (Massachusetts): TIMSS and PIRLS International Study Centre, Boston College. Recuperado de: http://timssandpirls.bc.edu/methods/pdf/TP_Sampling_Design.pdf

Kim, J., Craig, D. A. y Seoul, K. (2012). Validation of a Videoconferenced Speaking Test. *Computer Assisted Language Learning,* 25 (3), 257-275.

Kondo-Brown, K. (2002). A FACETS Analysis of Rater Bias in Measuring Japanese L2 Writing Performance. *Language Testing,* 19, 3-31.

Lee, Y. J. (2007). The Multimedia Assisted Test of English Speaking: The SOPI Approach. *Language Assessment Quarterly,* 4 (4), 352-366.

Lix, L. M., Keselman, J. C. y Keselman, H. J. (1996). Consequences of Assumptions Violations Revisited: A Quantitative Review of Alternatives to the One-Way Analysis of Variance F Test. *American Educational Research Association,* 66 (4), 579-619.

Macqueen, S. y Harding, L. (2009). Review of the Certificate of Proficiency in English (CPE) Speaking Test. *Language Testing,* 26 (3), 467-475.

McNamara, T. (1996). *Measuring Second Language Performance.* Londres: Longman.

Magal-Royo, T. y Giménez López, J. L. (2012). Multimodal Interactivity in the Foreign Language Section of the Spanish University Admission Examination. *Revista de Educación,* 357, 163-176.

Martín-Monje, E. (2012). The New English Oral Task in the Spanish University Admission Examination: A Methodological Proposal. *Revista de Educación,* 357, 143-161.

Ministerio de Educación, Cultura y Deporte (2002). *Marco Común Europeo de Referencia para las Lenguas: aprendizaje, enseñanza, evaluación.* Madrid: MEDC, Anaya.

Muñiz, J., Elosua, P. y Hambleton, R. K. (2013). Directrices para la traducción y adaptación de los tests: segunda edición. *Psicothema,* 25 (2), 151-157.

Myford, C. M. (2012). Rater Cognition Research: Some Possible Directions for the Future. *Educational Measurement,* 31 (3), 48-49.

OCDE (2008). *School Sampling Preparation Manual. PISA 2009 Main Study.* París: PISA, OECD Publishing. Recuperado de: http://www.oecd.org/pisa/pisaproducts/pisa2009/49023542.pdf

— (2009). *PISA Data Analysis Manual. SPSS, Second edition.* París: PISA, OECD Publishing. Recuperado de: http://browse.oecdbookshop.org/oecd/pdfs/free/9809031e.pdf

— (2012). *PISA 2009 Technical Report*. París: PISA, OECD Publishing. Recuperado de: http://www.oecd.org/pisa/pisaproducts/pisa2009/50036771.pdf

Pearson (2009). *Official guiDe to Pearson Test of English Academic.* Londres: Longman.

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 many-facet Rasch measurement. *Psicothema,* 23, 233-238.

Sanz, I. (1999). El examen de selectividad a examen. *Greta: Revista para Profesores de Inglés,* 7 (2), 16-29.

Schmider, E., Zieglel, M., Danay, E., Beyer, L. y Bürher, M. (2010). Is it Really Robust? Reinvestigating the Robustness of ANOVA against Violations of the Normal Distribution Assumption. *Methodology,* 6 (4), 147-151. DOI: 10.1027/1614-2241/a000016

Shettle, C. et ál. (2008). *The 2005 High School Transcript Study. User's Guide and Technical Report.* Washington, D. C.: National Center for Education Statistics, Institute of Education Sciences, U. S. Department of Education. Recuperado de: http://nces.ed.gov/nationsreportcard/pdf/studies/2009480rev.pdf

Shoukri, M. M. (2004). *Measures of Interobserver agReement.* Boca Ratón (Florida): Chapman & Hall.

Suto, I. (2012). A Critical Review of some Qualitative Research Methods used to Explore Rater Cognition. *Educational Measurement,* 31 (3), 21-30.

Vaughan, C. (1991). Holistic Assessment: What Goes on in the Rater's Mind? En L. Hamplyons (Ed.), *Assessing Second Language Writing in Academic Contexts,* 11-125. Norwood (Nueva Jersey): Ablex.

Weigle, S. C. (1998). Using FACETS to Model Rater Training Effects. *Language Testing,* 15, 263-287.

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

Wolfe, E. W. y McVay, A. (2012). Application Of Latent Trait Models To Identifying Substantively Interesting Raters. *Educational Measurement,* 31 (3), 31-37.

Xiaomin, S. y Houcan, Z. (2005). A Comparative Study on Methods Used in Estimating the Inter-Rater Reliability of Performance Assessment. *Psychological Science,* 28 (3), 646-649.

Zechner, K., Higgins, D., Xi, X. y Williamson, D. (2009). Automatic Scoring of Non-Native Spontaneous Speech in Tests of Spoken English. *Speech Communication,* 51 (10), 883-895.

Fecha de entrada: 01-02-2013 Fecha de aceptación: 27-09-2013

Dirección de contacto: Javier Suárez-Álvarez. Facultad de Psicología, Universidad de Oviedo. Plaza Feijoo, s/n; 33003, Oviedo, España. E-mail: suarezajavier@uniovi.es

Fecha de entrada: 01-02-2013 Fecha de aceptación: 27-09-2013