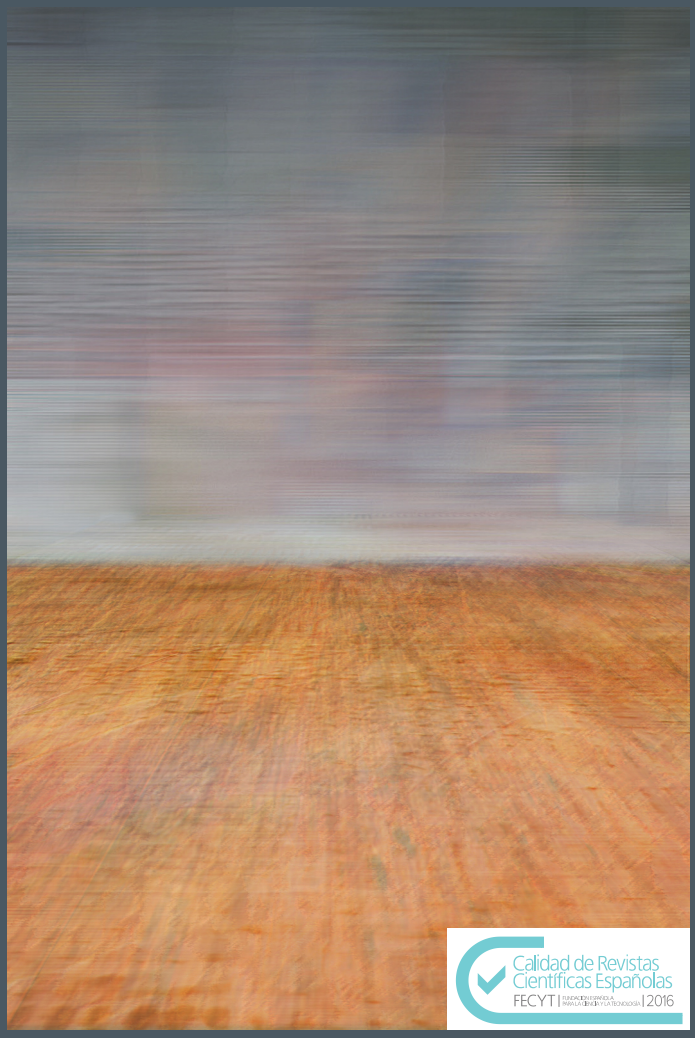
revista de EDUCACIÓN
Nº 381 JULIO-SEPTIEMBRE 2018

Vinculación de la prueba de comprensión oral del examen *CertAcles* de la Universidad de Granada con el MCER

Linking the University of Granada *CertAcles* listening test to the CEFR

Caroline Shackleton



GOBIERNO DE ESPAÑA
MINISTERIO DE EDUCACIÓN, CULTURA Y DEPORTE

Calidad de Revistas Científicas Españolas
FECYT | 2016

Vinculación de la prueba de comprensión oral del examen *CertAcles* de la Universidad de Granada con el MCER[1]

Linking the University of Granada *CertAcles* listening test to the CEFR

DOI: 10.4438/1988-592X-RE-2017-381-380

Caroline Shackleton
Universidad de Granada

Resumen

Este estudio, como parte del proceso de validación fehaciente del uso e interpretación de las puntuaciones del examen de inglés binivel de la Universidad de Granada (UGR), pretende relacionar, oficialmente, la prueba de comprensión oral de este examen con el Marco Común Europeo de Referencia (MCER). Con tal fin seguimos las recomendaciones del Consejo de Europa (2009) para determinar unos puntos de corte representativos de los niveles B1 y B2 del MCER. La prueba utilizada en el estudio fue la convocatoria de marzo del 2017 (N=464), que demostró ser un instrumento de medición fiable y apropiado. En el estudio participaron diez expertos, quienes, tras una sesión para familiarizarse con el MCER, realizaron varias sesiones de fijación de puntos de corte. Se utilizaron dos métodos: el método *Basket,* o de la cesta, para familiarizar a los participantes con los ítems de la prueba. después del cual se analizaron los resultados empleando un modelo de múltiples facetas de Rasch (MFR); por otro lado, el método *Bookmark*, o del marcador, que incorporó parámetros de dificultad del ítem, resultantes de un análisis Rasch de los puntajes de la prueba. La mediana final de los puntos de corte aportados por los expertos se calculó empleando una probabilidad de respuesta de 0.67, y se transfirió a la escala de capacidad zeta del modelo Rasch para determinar las puntuaciones brutas que representarán

(1) Agradecimientos: La autora quiere agradecer al Centro de Lenguas Modernas de la Universidad de Granada por sus aportaciones en apoyo de este estudio, así como a los dos revisores anónimos por sus valiosos comentarios y sugerencias.

los dos niveles de dominio del MCER en cuestión. También se realizaron otros controles de fiabilidad y de validez para proporcionar una documentación trasparente sobre el proceso completo, recomendado por el Consejo de Europa (2009). El método de equiparación de puntuaciones del modelo Rasch permite que los puntajes sean reproducibles en futuras versiones de la prueba y, por tanto, apoyan la *inferencia de generalización* del argumento de validez para la prueba binivel de la UGR.

Palabras Clave: pruebas de competencia lingüística, fijación de puntos de corte, validez, la teoría de respuesta al ítem

Abstract

As part of the ongoing collection of validity evidence to support the interpretation and use of the UGR bi-level English test scores this study aims to formally map the oral comprehension part of the test to the Common European Framework of Reference (CEFR). The advice provided by the Council of Europe (2009) was followed in order to determine cut scores which are representative of CEFR B1 and CEFR B2 levels. The test used in the study was from the March 2017 administration (N=464) which was shown to be a reliable and appropriate measurement instrument. The study involved ten participant judges who, after a CEFR familiarisation session, took part in standard setting sessions. Two methods of standard setting were employed. The *Basket* method was used mainly in order to familiarise the participants with the test items and was analysed using a Many-Facet Rasch measurement model. The *Bookmark* method allowed for the incorporation of item difficulty parameters produced by a Rasch analysis of test scores. The final median cut scores of the judges using a response probability of .67 were mapped back to the Rasch theta ability scale in order to determine the raw scores which represent the two CEFR ability levels. Several other validity and reliability checks were also carried out following CoE (2009) and transparent documentation on the whole process is provided. By using Rasch common item equating, the scores are reproducible on future versions of the test and so support the *generalisation inference* of the UGR test's validity argument.

Keywords: proficiency tests (language), standard setting, validity, item response theory

Introducción

En los últimos años, como parte del proceso de adaptación al Plan Bolonia y a la puesta en marcha del Espacio Europeo de Educación Superior, el Ministerio de Educación, Cultura y Deportes de España ha fomentado

Recibido: 11-05-2017 Aceptado: 02-02-2018

una política que resalta la comunicación en una segunda lengua como competencia fundamental para aumentar la movilidad universitaria y la empleabilidad. A tal fin, es de suma importancia que se proceda al diseño y planificación de cualquier examen de carácter nacional de manera que permita la vinculación de la actuación del candidato con los niveles de competencia del Marco Europeo Común de Referencia para las lenguas (MECR) si deseamos respetar los principios de trasparencia, comparabilidad y coherencia. Dichos principios juegan, sin duda, un papel de máxima repercusión para el fomento de la movilidad educacional y profesional dentro de la Unión Europea en cumplimiento de las actuales políticas de plurilingüismo europeas. En efecto, las recomendaciones del Consejo de Ministros sobre el empleo del MECR incluyen la exhortación a que los países miembros:

> … aseguren que todos aquellos exámenes, pruebas, o procedimientos de evaluación que conduzcan a la obtención de títulos de lengua de oficial reconocimiento tomen en debida consideración los aspectos relevantes del uso de la lengua, así como las competencias lingüísticas expuestas en el MECR, que se lleven a cabo en concordancia con principios de buenas prácticas y gestiones de calidad internacionalmente reconocidos, y que los procedimientos para vincular dichos exámenes y pruebas a los niveles de referencia común (A1–C2) del MCER se realicen de manera fiable y trasparente. (Consejo de Europa, 2008, p.4)

Como respuesta a este llamamiento, un gran número de los centros de lenguas universitarios en España han proporcionado exámenes de acreditación cuya elaboración sigue pautas internacionales, para así poder producir pruebas que sean válidas y de fiable vinculación al MCER. Además, en un intento de agilizar la coordinación y regulación de dichos esfuerzos, la Asociación de Centros de Lenguas en la Enseñanza Superior (ACLES) ha establecido un procedimiento de carácter nacional para certificaciones vinculadas al MCER: los exámenes *CertAcles*. Dichos exámenes han sido aprobados por la Conferencia de Rectores de las Universidades Españolas (CRUE) a nivel nacional, y además gozan de reconocimiento a nivel internacional a través de la Confederación Europea de Centros de Lenguas en la Enseñanza Superior (CercleS) desde el 2014. El modelo *CertAcles* (http://www.acles.es/es) requiere que cada destreza de capacidad lingüística sea evaluada de manera individual, que

sea desarrollada a partir de especificaciones basadas en el MCER, y que haya información proporcionada sobre los controles de calidad que siga las pautas establecidas por *los estándares* (*AERA, APA y NCME,* 2014).

Desde el 2009, el Centro de Lenguas Modernas (CLM) de la Universidad de Granada (UGR) administra tres veces por año un examen *CertAcles* binivel de B1/B2, de manera gratuita para los universitarios de la UGR. Desarrollado con la colaboración de la Dra. Rita Green, miembro experto de EALTA, el examen ya constituye un sistema de evaluación estable, y sigue todas las comprobaciones de validez recomendadas por ACLES. En el presente estudio describimos y reseñamos los procedimientos realizados para la vinculación del examen al MCER y el establecimiento de puntos de corte actualmente empleados para la prueba de comprensión oral del examen binivel B1/B2 de la UGR.

Marco teórico

El MECR, con sus seis niveles descriptivos (A1 hasta C2), se ha convertido en el sistema de referencia de los niveles de capacidad lingüística no solo en Europa, sino en otras muchas partes del mundo (Deygers, Zeidler, Vilcu, y Carlsen, 2017; Figueras, 2012). La vinculación de los resultados de una prueba con niveles del MCER amplía el nivel significativo de los resultados de los usuarios de dichas pruebas (Kane, 2012). Por ejemplo, cuando un resultado informa sobre la capacidad de comprensión oral a nivel B1, el usuario destinatario de los resultados de la prueba recibe información sobre los tipos de actividades que un usuario de la lengua con estas competencias sería capaz de realizar (Tannenbaum y Cho, 2014). En este sentido, cualquier prueba que afirme medir las competencias del MCER, tiene la obligación de fundamentar dichas alegaciones en pruebas de que esto es así. En otras palabras, una prueba debe demostrar su validez.

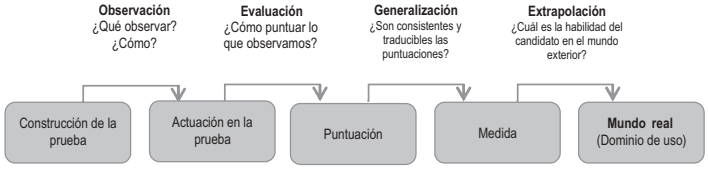
La validez no se percibe como propiedad intrínseca de la prueba —una prueba no se puede validar como tal (Chapelle, 2012; Cizek, 2016)— sino más bien como un concepto polifacético y unitario que se basa en múltiples fuentes de evidencia para comprobar las inferencias sacadas sobre la actuación de un candidato y así justificar las decisiones resultantes tomadas en base a los resultados de pruebas (Messick, 1989). Es necesaria la recopilación de evidencias que respalde la interpretación

Recibido: 11-05-2017 Aceptado: 02-02-2018

y uso de resultados de pruebas; por tanto, es recomendable el empleo de un enfoque *basado en argumentos de validez*, el cual proporciona un marco que sirve de guía para los organismos de evaluación sobre los tipos de evidencia que deben recoger (*AERA, APA and NCME,* 2014; Bachman, 2005; Bachman y Palmer, 2010; Chapelle, 2012; Kane, 2012).

Las decisiones sobre qué tipos de evidencia deben ser recopilados se deben basar necesariamente en la finalidad de la prueba (Fulcher y Owen, 2016) y, en el contexto europeo, el marco a seguir ha sido proporcionado por ALTE (2011) con el fin de guiar a los organismos de evaluación en el diseño y desarrollo de pruebas vinculadas al MCER. Dicho documento define, según líneas generales, los tipos de evidencias que deben ser recopiladas a lo largo del ciclo de desarrollo de la prueba (véase figura1), con el objetivo de incorporar la validación de pruebas a la totalidad del proceso de diseño e implantación de la prueba. A diferencia de las recomendaciones de Bachman (2005), y Bachman y Palmer (2010), la cadena de razonamiento comienza con el constructo de la prueba, no en decisiones basadas en los resultados de prueba. Como tal, vincula la actuaciones en las tareas de la prueba a inferencias sobre las capacidades lingüísticas del candidato en situaciones del mundo exterior, esto es, la inferencia de extrapolación. No obstante, el mundo más allá de la prueba permanece sin especificarse; se trata de una afirmación general con vínculos al MCER. Por tanto, se puede argumentar que el modelo ALTE dirige nuestro enfoque hacia la interpretación de los resultados, es decir, que pregunta '¿se puede interpretar un resultado de prueba como un nivel de capacidad del MCER?' Tal postura se alinea con la conceptualización de la validez propuesta por Cizek (2016), cuyo marco diferencia entre la validación de inferencias de resultados de pruebas por un lado, y las justificaciones para la implantación de la prueba por el otro.

FIGURA I. Cadena de razonamiento para un argumento de validez



Fuente: ALTE (2011, p. 15)

Para un enfoque basado en el argumento, la evidencia de la validez del establecimiento de los puntos de corte es un componente esencial (Papageorgiou y Tannenbaum, 2016, p.110). Debido a que los candidatos tienen que alcanzar el punto de corte de una prueba para demostrar que han conseguido el nivel de capacidad requerido, los resultados de esta quedarían claramente expuestos a la crítica en el caso de que dicho punto no se estableciese de manera apropiada. A tal fin, el Consejo de Europa (2009) ha proporcionado un manual que describe, en líneas generales, el proceso necesario para que los organismos de evaluación vinculen sus pruebas al MCER. Dicho proceso de vinculación consta de las siguientes cuatro actividades interrelacionadas, todas ellas necesarias para poder relacionar los resultados de prueba con el MCER.

- **Familiarización:** Los miembros de cualquier panel de examinadores deben estar familiarizados con el contenido del MCER y sus escalas.
- **Especificaciones:** Estas deben incluir una descripción detallada de la prueba y su relación con las categorías del MCER para así construir un argumento de vinculación sobre la relevancia de esta para el MCER.
- **Estandarización y Establecimiento de normas:** Toda prueba debe estar provista de un punto de corte u otra puntuación mínima como requisito necesario para aprobarla.
- **Validación empírica:** El Consejo de Europa (2009) puntualiza tres formas de evidencia de validez —la de procedimiento, la interna y la externa— las cuales tienen que proporcionarse como parte del proceso de establecimiento de normas.

Son muchos los organismos de evaluación que han seguido los consejos del manual desde su publicación: la mayoría de los proveedores de exámenes de dominio internacionales para la lengua inglesa han realizado estudios de validación (p.ej. véase Brunfaut & Harding, 2014; Kanistra & Harsch, 2017; Tannenbaum & Wylie, 2008), y proveedores menores a) nivel nacional también siguen los pasos descritos en el manual para poder reivindicar su vinculación al MCER (p.ej. véase Downey & Kolias, 2010; Kantarcioglu, Thomas, O'Dwyer, & O'Sullivan, 2010). El enfoque principal de dichos estudios es el proceso de establecimiento de normas y su validación empírica. Como viene expuesto en el Consejo de Europa (2009), el momento decisivo en el proceso de vincular una

Recibido: 11-05-2017 Aceptado: 02-02-2018

prueba a cualquier nivel del MCER es el establecimiento de una norma de decisión para poder asignar alumnos a uno de los niveles del MCER en función de su actuación en la prueba (p.11); es decir, la determinación del punto de corte.

En resumen, cualquier prueba que afirme tener vinculación con el MCER debe validarse y debe demostrarse que es fiable y representativa del nivel de competencia MCER en cuestión, ya que si no se demuestra empíricamente que una prueba es válida y fiable, su vinculación al MCER carece de sentido (Alderson, 2012). Por consiguiente, durante el ciclo de desarrollo de una prueba, el proceso de estandarización se suele realizar con posterioridad. Son precisamente estos tipos de evidencia de validez los que se han utilizado en la prueba de la UGR; el enfoque del presente estudio pretende describir el proceso oficial de vinculación expuesto en el Consejo de Europa (2009) y su empleo en la prueba de la UGR para establecer puntos de corte a niveles B1 y B2.

Metodología

Los diferentes estudios de investigación clasifican las técnicas de establecimiento de normas según dos principales tipologías: la primera se enfoca en la prueba; la segunda, en el candidato. En el presente estudio, la carencia de datos sobre la población examinada requirió que el estudio principal adoptara un enfoque en torno a la prueba. No obstante, los datos tomados de un número reducido de candidatos previamente matriculados en un curso de familiarización del examen en el CLM fueron utilizados como punto de comparación para así proporcionar cierta evidencia triangulada y por tanto, contribuir a la evidencia de la validez externa.

De los diferentes métodos que tienen un enfoque en torno a la prueba, se descartaron los de tipo de probabilidad *Angoff*: no solo llevan mucho tiempo, sino que estudios previos además han señalado problemas respecto a la incapacidad de los expertos para entender y articular correctamente las probabilidades condicionales (Ferrara y Lewis, 2012; Hambleton y Jirka, 2006; Reckase, 2010). En su lugar, se decidió utilizar una combinación de dos otros métodos que actualmente tienen una gran popularidad en diferentes investigaciones: El método *Basket* y el método *Bookmark* (que en español corresponden a 'de cesta' y 'marcador', respectivamente).

A efectos prácticos, el método *Basket* es un *método de emparejamiento de descriptores de ítems*, el cual especifica el nivel de competencia necesario para que se conteste cada ítem de la prueba. Los expertos analizan los ítems de la prueba para responder a la pregunta '¿En qué franja del MCER tendría que estar un candidato para poder contestar este ítem de forma correcta?'. Se considera el más práctico y sencillo de todos los métodos de establecimiento de puntos de corte y, además, un método que no solo refleja la importancia de los descriptores de los niveles de competencia, sino que también enfatiza el contenido de la prueba. Sin embargo, no proporciona ninguna información sobre la dificultad de los ítems y, como tal, uno de sus mayores problemas es la falta de consistencia a la hora de comparar los juicios con las medidas de dificultad empíricas (Kaftandjieva, 2010). A pesar de esta desventaja, el método ha sido empleado en el presente estudio no para llegar a establecer el punto de corte final, sino principalmente como medio de familiarización con el contenido y para fomentar una mayor discusión sobre ello entre los expertos. Los resultados fueron analizados empleando el modelo de medición de múltiples facetas de Rasch (MMFR) en programa FACETS (Linacre, 2017). Aunque se ha argumentado que el análisis mediante MMFR no es del todo apropiado para las fases posteriores del proceso de establecimiento de puntos de corte, cuando los expertos deben esforzarse para llegar a un consenso (Eckes, 2009), dicha crítica no se consideró relevante para el contexto del presente estudio, debido a que su empleo se limitaba exclusivamente a la fase de familiarización inicial.

Para que el proceso de establecer puntos de corte tenga un valor verdaderamente significativo, los expertos deben recibir información estadística sobre el funcionamiento de los ítems de la prueba. En el método *Bookmark* (Mitzel, Lewis, Patz, y Green, 2001), los ítems son mostrados en un cuadernillo donde aparecen ordenados del más fácil al más difícil distribuyéndose, por tanto, según parámetros de dificultad. Los expertos avanzan por el cuaderno de ítems ordenados (CIO) desde el ítem más fácil hasta el más difícil, colocando un 'marcador' en el punto exacto en el que consideran que un candidato que cumple los requisitos mínimos tendría menos probabilidades de dar la respuesta correcta que la mínima probabilidad especificada. Se concibe la capacidad del candidato para dominar un ítem en términos probabilísticos, o '*probabilidad de respuesta'* (PR); esta debe establecerse de antemano, aunque en la práctica se suele establecer en 0.67, una probabilidad del 67 % —es decir, 2/3— de que

Recibido: 11-05-2017 Aceptado: 02-02-2018

el candidato conteste el ítem de manera correcta. El marcador se coloca entre dos ítems en la escala latente representada por los parámetros de dificultad de *Rasch* (logits) para los ítems (β-parameters). Finalmente, se toma la mediana de los marcadores de los expertos en la variable latente para así llegar a un estándar en común. En resumen, el método *Bookmark* no solo es fácil de implantar en pruebas que se desarrollan utilizando el modelo de medición *Rasch*, y donde la capacidad del candidato y la dificultad del ítem se colocan en la misma escala, sino que refleja de forma clara la naturaleza continua de las escalas del MCER; dos ventajas que le confieren idoneidad para su uso en el presente estudio.

Instrumentos

La prueba empleada para el presente estudio proviene de la convocatoria de marzo de 2017. La prueba de comprensión (42 ítems) comprende cinco tareas: dos a nivel B1, una en el umbral entre B1 y B2, y dos a nivel B2. Todas las tareas fueron desarrolladas de acuerdo con las especificaciones de la prueba y con la colaboración de nuestra asesora; por tanto, se puede considerar que su nivel de dificultad se ha establecido de manera correcta por expertos. Además, todas habían sido pilotadas de antemano (N = entre 144 y 434), con la subsiguiente revisión o eliminación de los ítems que resultaron no funcionar de forma correcta. La prueba presenta un buen muestreo del constructo basado en los descriptores de niveles B1 y B2, e incluye una variedad de métodos de evaluación/tipos de tarea.

El análisis de la prueba con teoría clásica de los tests (N = 464) produjo un Alfa de Cronbach de 0.934 (EEM = 2.5) y evidencia la fiabilidad de la prueba. El análisis *Rasch* de los resultados de la prueba demuestra que la prueba tiene unidimensionalidad, y las estadísticas *Rasch* de *infit mean square (MNSQ)* quedan entre los valores satisfactorios de 0.74 y 1.24. La separación de ítems es de 10.2 y la fiabilidad de la separación es de 0.99. Por consiguiente, se puede afirmar que la prueba incluye un amplio abanico de diferentes dificultades de los ítems, del mismo modo que podemos confiar en que las estimaciones de dificultad sean reproducibles. El mapa de la variabilidad de los ítems reproducido en la figura II muestra los ítems ordenados según su dificultad, junto con la capacidades de los candidatos. Utilizamos esta información para crear el CIO para el estudio que usó el método marcador. Aquí, la media de los

valores *logit* de las dificultades de los ítems y la de las capacidades de los candidatos resultaron muy parecidas, lo cual demuestra que la dificultad de la prueba se ajusta bien a la población evaluada. Dado que el presente estudio no pretende investigar la prueba en sí, consideramos que la susodicha información es evidencia suficiente de que el instrumento de medición utilizado en el estudio es tanto fiable como pertinente.

FIGURA II. Mapa de variabilidad de ítems

```
     PERSON - MAP - ITEM
            <more> <rare>
 4            .#  +
            .####
                .# T   L35    L40

 3            .##  +

            .####   T
             .###     L39

 2           .###  +  L42
            ##### S
             ####     L38
    .############   S
             ####     L29D
 1          #####  +  L9A
         .########     L22    L37
             ###      L25    L26C   L33A
         ########     L16    L28A
             ### M    L12B   L18    L19    L32D
 0     .########  +M L10B   L11D   L24    L34    L41
    ############     L1C    L30B   L4A
         .####       L27B   L36    L6D
    ############     L21    L23    L7B
          .#####     L20
-1  .############  +  L31C
             ####
          .#### S S L3H    L8D
         .#####     L15B   L2F
          #####     L13C
-2          ###  +  L14C
            .#
                    L17    L5I
              #   T
             .# T
-3              +
              .

              .

-4            #  +
            <less> <frequ>
EACH "#" IS 3. EACH "." IS 1 TO 2
```

Fuente: Creada por la autora

De acuerdo con Cizek (2012), se les administró un cuestionario a los expertos tras las sesiones de establecimiento de puntos de corte para así suministrar evidencia de la validez de procedimiento. El cuestionario

Recibido: 11-05-2017 Aceptado: 02-02-2018

empleaba una escala *Likert* de cuatro puntos para eliminar la posibilidad de respuestas neutras, y permitió la recopilación de las opiniones de los expertos respecto a sus interpretaciones del proceso global, además de sus pensamientos sobre las decisiones finales tomadas con respecto a los puntos de corte.

Participantes

Resulta importante que los participantes sean expertos tanto en la aplicación del MCER, y por tanto en sus descriptores de niveles de competencia, como en la enseñanza y evaluación de la lengua evaluada (Tannenbaum y Cho, 2014). Este es, en efecto, el caso de los diez expertos que participaron en el estudio. Como la tabla I muestra, todos son profesores en el CLM con buena experiencia de la enseñanza del inglés como segunda lengua, y con buenos conocimientos del MCER. Además, conocen bien a la población de candidatos, y son conscientes de las consecuencias que los puntos de corte tienen para dicha población.

TABLA I. Información biográfica de los expertos

Experto	Sexo	Títulos	Años enseñando inglés	Número de horas de formación recibida en relación con el MCER
1	Mujer	Licenciatura, Master	35	300+
2	Hombre	Licenciatura	28	80
3	Hombre	Licenciatura, Diploma de posgrado TEFL	35	100+
4	Hombre	Licenciatura, Master, Diploma de posgrado CELTA	16	70
5	Hombre	Licenciatura	22	70
6	Mujer	Licenciatura, Master	36	400+
7	Mujer	Licenciatura, Posgrado PGCE, PhD	28	750+
8	Hombre	Licenciatura, Master, Diploma de posgrado CELTA	18	500+
9	Mujer	Licenciatura, PGCE, Master	28	750+
10	Mujer	Licenciatura	22	500+

Fuente: Creada por la autora

Resultados

A continuación, se procederá a informar sobre los resultados de cada fase de las sesiones de establecimiento de puntos de corte, incluidas las fases de familiarización y de validación externa.

Familiarización

Como preparación para el estudio, se les pidió a todos los expertos que realizaran sesiones de formación para la destreza de comprensión oral que han sido proporcionadas por el proyecto *Ceftrain* (véase http://www.helsinki.fi/project/ceftrain). Se procedió a realizar un ejercicio de ordenar descriptores del MCER para la destreza de comprensión oral, en el cual los expertos recibieron una mezcla de descriptores y tenían que asignar cada uno a un nivel del MCER. Los resultados de este ejercicio dieron un Alfa de Cronbach de 0.98. La tabla II muestra la correlación entre los descriptores del MCER y las estimaciones de los participantes obtenida mediante el empleo de un coeficiente de correlación de Spearman. Estos resultados fueron trasmitidos a los participantes, para que se pudieran debatir más a fondo con el fin de llegar a un consenso sobre las características significativas de los niveles del MCER.

TABLA II. Correlación entre descriptores del MCER y asignaciones de los expertos

Experto	1	2	3	4	5	6	7	8	9	10
	.851	.875	.835	.889	.896	.836	.957	.985	1.00	.870

(Nota: todas las correlaciones fueron significativas: $p \leq .01$)
Fuente: Creada por la autora

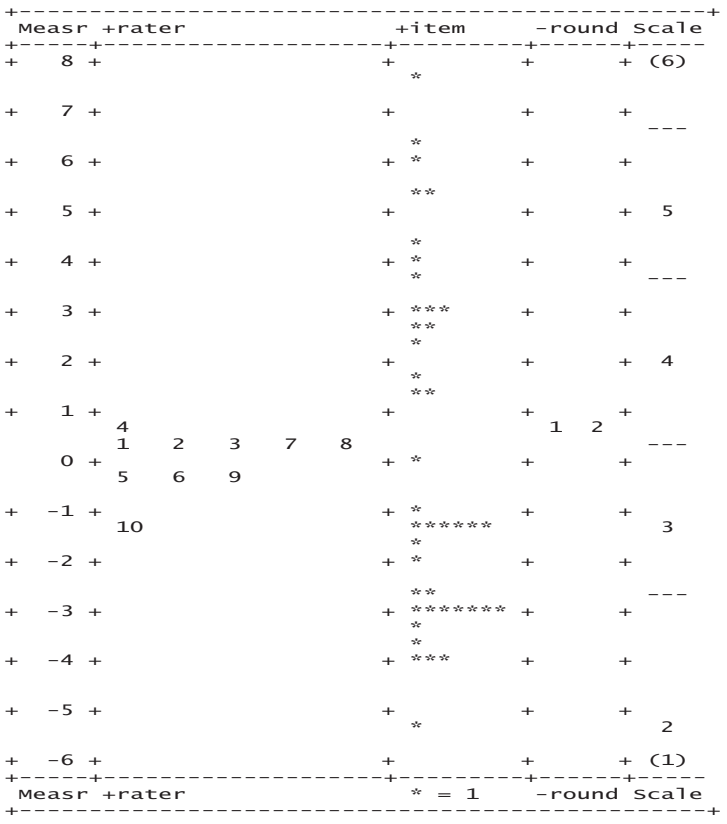
El método *Basket* de establecimiento de puntos de corte

Los expertos recibieron una sesión informativa inicial sobre el constructo del examen y las especificaciones de la prueba, seguida por una explicación específica del método *Basket*. La figura III muestra un

Recibido: 11-05-2017 Aceptado: 02-02-2018

análisis MFR de los resultados de todos los expertos para cada ítem de la prueba. El análisis permite tomar en cuenta múltiples aspectos de las estimaciones, y calibra ítems, expertos y la escala de calificación usando la misma escala de intervalos iguales. La escala empleada para representar el nivel mínimo del MCER que un candidato debería poseer para poder contestar a cada ítem es la siguiente: 1=A2, 2=B1 mínimo, 3= B1 holgado, 4=B2 mínimo, 5=B2 holgado, 6=C1. Los resultados de las dos rondas difieren mínimamente, y demuestran que los expertos confiaban en sus decisiones.

FIGURA III. Regla vertical de los resultados del método *Basket* en FACETS



Fuente: Creada por la autora

Recibido: 11-05-2017 Aceptado: 02-02-2018

Queda patente que los expertos consideraron que el contenido de los ítems tenía una distribución equilibrada a lo largo de la escala desde B1 del MCER hasta B2. Los umbrales Rasch-Andrich muestran que los expertos colocaron los ítems en distintas categorías con un aumento progresivo de dificultades, lo que concuerda con las pretensiones originales de los desarrolladores de la prueba. Los expertos discreparon ligeramente, y el experto 10 asignó los ítems a un nivel de dificultad levemente inferior al los demás expertos. No obstante, el análisis MFR toma este tipo de variaciones en cuenta y es capaz de dar una calificación a cada uno de los ítems mediante el empleo de un 'promedio justo', el cual realiza ajustes por la lenidad o severidad de los expertos. Si aplicamos los umbrales de Rasch Thurston, que miden el límite en el que un ítem tiene un 50 % de probabilidades de ser asignado a una categoría superior o inferior (i.e. el umbral de probabilidad cumulativa de 50 % (la mediana)), llegamos a unos puntos de corte iniciales de 13 para B1 y 34 para B2.

Método *Bookmark* de establecimiento de puntos de corte

Antes de pasar al estudio principal, los participantes recibieron retroalimentación, que incluía datos de impacto, sobre el método *Basket*. Posteriormente, los expertos realizaron una discusión final, llegando a la conclusión general de que el punto de corte para B2 resultaba algo alto (solo 81 candidatos, o el 17%, habría conseguido el B2). En este punto, se les explicó que estos puntos de corte no serían los finales y que el ejercicio anterior se había realizado con el fin de permitir a los participantes que se familiarizaran con la prueba antes de ver los valores de dificultad reales que se iban a emplear para el establecimiento de los puntos de corte.

A continuación, los participantes recibieron una explicación del método *Bookmark*, además de una copia del CIO. Al terminar la primera ronda, las estimaciones de los expertos fueron presentadas al grupo, junto con los datos de impacto. Los expertos procedieron a discutir los resultados más a fondo antes de pasar a la segunda ronda, en la cual se les permitió cambiar la asignación del marcador. La tabla III muestra los resultados de las dos rondas[1].

Recibido: 11-05-2017 Aceptado: 02-02-2018

TABLA III. Resultados de los puntos de corte establecidos mediante el método *Bookmark*

1ª Ronda	Número de páginas (número de expertos)	Media (DE)	Mediana	Habilidad (θ) después de ajustar a la PR de .67	Punto de corte final
B1	8/9 (3) 10/11 (2) 12/13 (2) 13/14 (1) 15/16 (1)	11 (2.87)	10/11	-.13 logits	22
B2	25/26 (1) 28/29 (2) 29/30 (5) 30/31 (1)	28.44 (1.42)	29/30	1.04 logits	29
2ª Ronda					
B1	8/9 (7) 9/10 (1) 10/11 (1)	8.33 (.71)	8/9	-.70 logits	16
B2	28/29 (6) 29/30 (3)	28.56 (.73)	28/29	.93 logits	28

Fuente: Creada por la autora

La decisión global para el conjunto de los expertos se obtiene de la mediana del segundo grupo de marcadores de todos los participantes; para establecer el punto de corte, se recomienda emplear el valor zeta inferior (Consejo de Europa, 2009). La mediana en este caso corresponde a un nivel de capacidad de θ = -.7 *logits* para B1 y θ = 0.93 *logits* para B2, después de haber realizado ajustes por la PR de 0.67. Si aplicamos este nivel de capacidad a la curva de características de la prueba, resulta equivalente a una puntuación bruta de 16 para el B1 y de 28 para el B2.

Finalmente, los expertos realizaron una última discusión en la cual consideraron tanto los datos de impacto como aquellos de validez externa. Los participantes se mostraron satisfechos con las decisiones finales y de acuerdo con que los puntos de corte establecidos deberían ser utilizados para la prueba. Los subsiguientes datos de impacto para la convocatoria de marzo 2017 dan los siguientes resultados: 135 No aptos (el 29 %), 166 Aptos de nivel B1(el 36 %) y 163 Aptos de nivel B2 (el 35 %).

Validez del procedimiento

Los resultados del cuestionario posterior al proceso de establecimiento de puntos de corte se muestran en la tabla IV. Para cada una de las preguntas, los participantes contestaron usando una escala *Likert* de cuatro puntos, donde 1 representa total acuerdo. En términos generales, se evidenció un alto nivel de acuerdo, los miembros del panel confiaban en el procedimiento y en sus decisiones finales. En particular, todos los miembros consideraban que las decisiones de establecimiento de los puntos de corte finales representaban los niveles B1 y B2 del MCER para la destreza de compresión oral de manera justa y precisa.

TABLA IV. Resultados del cuestionario

	Mean	SD
Creo que comprendo bien los niveles del MCER después de las sesiones de familiarización.	1.3	.48
Me han ayudado las explicaciones del constructo de la prueba y de sus especificaciones.	1.5	.71
He comprendido los procedimientos del método *Basket* para el establecimiento de puntos de corte.	1.2	.42
Confiaba en mis decisiones a la hora de contestar a la pregunta, '¿En qué nivel del MCER debe estar para poder contestar a este ítem?'	1.6	.52
He entendido el concepto de 'candidato de mínima capacidad'	1	.00
El uso del método *Basket* para el establecimiento de puntos de corte me ha ayudado a comprender los contenidos de la prueba.	1.2	.42
He comprendido los procedimientos del método *Bookmark* para el establecimiento de puntos de corte.	1.33	.71
He comprendido el concepto de PR y el significado de la probabilidad del 67%	1.11	.33
Confiaba en la colocación de mi marcador en la primera ronda.	1.67	.50
Confiaba en la colocación de mi marcador en la segunda ronda.	1.11	.33
Los puntos de corte finalmente recomendados por el grupo representan los niveles B1 y B2 del MCER para la destreza de compresión oral de manera justa.	1	.00

Fuente: Creada por la autora

Recibido: 11-05-2017 Aceptado: 02-02-2018

Validez interna

La validez interna de los procedimientos para el establecimiento de los puntos de corte trata de la exactitud y la consistencia de los resultados así como de la calidad de las estimaciones determinadas y publicadas. Para el método *Basket*, los resultados mostraron que los expertos tenían una consistencia interna con *Infit* MNSQ medio de 0.98 y el DE de 0.28. Los *Infit* MNSQ y *Zstd* muestran la existencia de un experto con *overfit* y otro con *underfit*. No obstante, solo uno de los expertos muestra un ligero *underfit* en términos de *outfit* y había una alta similitud entre los resultados esperados y observados, como sería de esperar de evaluadores actuando como expertos independientes (Linacre, 2017).

Con el fin de reforzar la vinculación de la prueba con el MCER, las decisiones de los expertos sobre la dificultad de los ítems tomadas con el método *Basket* fueron comparadas con las dificultades observadas mediante el análisis *Rasch*. Este análisis muestra más evidencia en la calidad de la estimaciones de establecimiento de puntos de corte. El conjunto de datos es limitado, con rangos vinculados en los ítems estimados, y una correlación Tau de Kendall de $\tau = .57, p < .001$ demuestra una relación significativa entre la estimación de la dificultad del ítem (empleando valores medianos de la segunda ronda) y los parámetros de dificultad *Rasch*. Este resultado proporciona evidencia de que los expertos reconocieron un aumento progresivo de la dificultad de los ítems.

Es necesario que proporcionemos evidencia de la exactitud de la clasificación de los puntos de corte. Los puntos de corte finalmente establecidos después de la segunda ronda del método *Bookmark* mostraron un alto grado de consenso, con DEs muy pequeños. La validez de estos puntos de corte puede ser evidenciada mediante el informe del error estándar de estimación (EE_E), una estimación de las probabilidades de replicar los puntos de corte recomendados (Tannenbaum y Cho, 2014, p.245). Cohen, Kane y Crooks (1999, p.364) sostienen que el EE_E debería ser ≤ 1⁄2 EEP. La tabla V muestra los resultados de este cálculo para las dos rondas.

Recibido: 11-05-2017 Aceptado: 02-02-2018

TABLE V. Resultados para la exactitud de la clasificación

	Ronda 1		Ronda 2	
	B1	**B2**	**B1**	**B2**
Deviación estándar para el punto de corte medio (DE)	2.87	1.42	0.71	0.73
Error estándar de la prueba (EEP)	2.5	2.5	2.5	2.5
Error estándar de estimación (EE_E)	1.01	0.50	0.25	0.26
$\frac{EE_E}{EEP}$	0.4	0.2	0.1	0.1

Fuente: Creada por la autora

Se puede ver que la EE_E era siempre inferior a la mitad del EEP, y por tanto los puntos de corte cumplen el criterio de calidad. Sin embargo, es importante destacar aquí que se ha aducido que las estimaciones de la fase final del proceso no suelen exhibir mucha variabilidad debido a que esta parte del proceso anima a los expertos a converger y llegar a una decisión consensuada (Linn, 2003).

Validez externa

La validez externa se refiere al uso y a la comparación de diferentes métodos para el establecimiento de puntos de corte, además de las comparaciones o triangulaciones con medidas obtenidas de otros estudios. Al igual que han informado previamente otros estudios, los dos procedimientos produjeron puntos de corte diferentes tanto para nivel B1 como para nivel B2. Para nivel B2, los resultados de la relación entre la puntuación bruta y los valores *zeta* muestran que existe cierto grado de discrepancia entre los dos métodos, una diferencia de poco más de *2 errores estándar*. A nivel B1, sin embargo, el punto de corte del método *Basket* está a menos de *1 error estándar* del punto de corte establecido con el método *Bookmark*, lo cual evidencia que los dos métodos produjeron resultados bastante consistentes para esta parte de la prueba, a pesar de que el método *Basket* solo se empleara para que los expertos se familiarizaran con el contenido.

Para aquellos candidatos que habían realizado cursos de familiarización del examen en el CLM antes de la convocatoria, se procedió a analizar sus

Recibido: 11-05-2017 Aceptado: 02-02-2018

resultados con más detalle, mediante la comparación de sus notas finales con los resultados pronosticados por sus profesores. A pesar del número reducido de candidatos en cuestión, la comprobación de la exactitud de clasificación de estos candidatos permitió cierto grado de triangulación y aportó otra contribución más al conjunto de evidencias de validez. La tabla VI muestra los resultados.

Este estudio, de carácter limitado, y enfocado en el criterio de estimaciones de profesores, muestra que existe una alta correlación entre los pronósticos de los profesores y la puntuación obtenida en la prueba $\tau = .85$, $p < .001$. El acuerdo de clasificación para todos los pronósticos de resultados B1 y B2 es del 100 %, lo cual nos proporciona algo más de evidencia de validez externa sobre la exactitud de los puntos de corte. Vemos que, de los candidatos a los que sus profesores pronosticaron un resultado en umbral entre los niveles de A2/B1 y B1/B2, alrededor del 50 % obtuvo el nivel superior. No obstante, cabe entender que el presente estudio es de una escala reducida, y que lo ideal sería otro estudio de mayor escala que pudiera proporcionar mejor evidencia de validez externa.

TABLA VI. Correspondencia entre los pronósticos de profesores y las decisiones finales basadas en los puntos de corte establecidos.

Nivel MCER pronosticado para la prueba de comprensión oral	Puntuación en la prueba	Nivel MCER recibido en la prueba
No Apto	1	No Apto
No Apto	11	No Apto
A2/B1	12	No Apto
A2/B1	12	No Apto
A2/B1	14	No Apto
A2/B1	17	B1
A2/B1	19	B1
B1	17	B1
B1	19	B1
B1	19	B1
B1	22	B1
B1	22	B1
B1	23	B1

B1	25	B1
B1	26	B1
B1	27	B1
B1	27	B1
B1/B2	26	B1
B1/B2	27	B1
B1/B2	33	B2
B1/B2	33	B2
B1/B2	36	B2
B2	30	B2
B2	31	B2
B2	31	B2
B2	35	B2
B2	36	B2
C1	40	B2

Fuente: Creada por la autora

Análisis de los resultados

Este estudio ha proporcionado evidencia que fortalece afirmaciones sobre el uso e interpretación de los resultados de la prueba UGR B1/B2 en cuanto a la vinculación con el MCER se refiere. El proceso para el establecimiento de puntos de corte se ha detallado como parte elemental del proceso de vinculación y, de esta manera, queda documentado el seguimiento de un proceso razonable y sistemático para llegar a establecer unos puntos de corte recomendados. Se ha dado pruebas sólidas para justificar los puntos de corte, cuyo establecimiento se puede considerar una mezcla de estimaciones, sicometría, y pragmática (Hambleton y Pitonoak, 2006, p.435), y donde la cuestión no es que el punto de corte sea correcto o no, sino que las decisiones basadas en los puntos de corte sean razonables, aceptables en términos generales, y tengan consecuencias positivas (Kane, 2017, p.11).

El punto de partida del manual de adecuación con el MCER (Consejo de Europa, 2009) es, sin duda, el uso de estimaciones de expertos, una metodología muy recomendada y aplicada en investigaciones sobre el establecimiento de puntos de corte. No obstante, algunos estudios

Recibido: 11-05-2017 Aceptado: 02-02-2018

han informado sobre las metodologías que hacen uso de estimaciones de expertos como no fiables, indudablemente como consecuencia del hecho de que el MCER no proporciona suficientes descripciones con precisión de sus niveles de capacidad (Alderson et ál., 2006; Fulcher, 2004; Weir, 2005). Existe la posibilidad de que los expertos en un estudio de adecuación interpreten los descriptores del MCER de manera diferente o que tengan interiorizada su propia idea de exactamente qué significa estar en una franja específica del MCER (Eckes, 2012; Harsch y Hartig, 205; Papageorgiou, 2010). En efecto, North y Jones (2009) sostienen que ni la familiarización con el MCER, ni la estandarización, ni las estimaciones de índices de consistencia y acuerdo, podrán demostrar que un grupo particular de expertos involucrados en la estimación de niveles para un determinado idioma no lleven consigo a la tarea sus propias interpretaciones, fruto de su particular perspectiva cultural (p.16).

No obstante, y a pesar de estas limitaciones, los expertos en el presente estudio se mostraron satisfechos con sus resultados y consideran que los puntos de corte son representativos de un candidato que pueda realizar tareas asociadas con el nivel. Por tanto, el estudio ha proporcionado más evidencia hacia la validez sustantiva y la validez de constructo de la prueba. Las dos metodologías empleadas requirieron discusiones extensas sobre la relación de la dificultad de ítems individuales con el MCER; como tal, refuerzan la afirmación de que el contenido de la prueba se adecúa con el MCER, en la línea de Kanistra y Harsch (2017) en su estudio de vinculación sobre la prueba ISE de Trinity. De hecho, el problema de cómo definir al candidato que actúe en un nivel mínimo de actuación resulta ser una tarea mucho más fácil para una prueba que ha sido desarrollada para poner en práctica el modelo del MCER; debido a que el estándar ya ha sido incorporado en la prueba, los puntos de corte adquieren un carácter más significativo (Tannenbaum y Wylie, 2008). En cambio, un estudio de vinculación retroactiva tendría que prestar particular atención a la fase de especificación de contenidos del proceso de adecuación con el MCER para que proporcione evidencia de que los contenidos de la prueba miden las capacidades lingüísticas descritas por el marco (Tannenbaum y Cho, 2014). Varios estudios de vinculación al MCER han señalado problemas a la hora de relacionar contenidos con los descriptores (p.ej. Brunfaut y Harding, 2014). Además, se ha argumentado que son muy pocos los creadores de pruebas que prestan atención a esta fase del proceso de vinculación con el MCER (Green, 2017); si los

contenidos no se alinean con el MCER, pues existe poca justificación para la realización de un estudio de establecimiento de puntos de corte. Aquí cabe destacar que la prueba de la UGR ha sido desarrollada de forma específica para que sea representativa del MCER; como tal, el estudio de vinculación no tiene carácter retroactivo, sino que forma parte del mismo proyecto de desarrollo de la prueba, y el estándar está ya incorporado a la prueba de forma implícita. Las especificaciones de la prueba incluyen descriptores del MCER y las directrices para los desarrolladores de ítems abarcan varias categorías presentadas en el Consejo de Europa (2009). Dicho proceso tiene parecido con el enfoque *a priori* de la prueba *Pearson Test of English Academic* (De Jong y Zheng, 2016).

En cuanto a la metodología empleada para el estudio, no existe el mejor método para el establecimiento de puntos de corte; los métodos deben adecuarse a la situación. Como ya se ha observado, el estudio principal, el del método *Bookmark*, se considera apropiado para usar en contextos de evaluación que se basan en el modelo *Rasch* para el desarrollo de pruebas. Por otra parte, el empleo del método *Basket* también tuvo un valor inestimable durante la primera fase del estudio, ya que permitió a los expertos que se familiarizaran con los contenidos de la prueba. En este caso, habría sido imposible comenzar con el estudio *Bookmark* de inmediato, y esto resulta de particular relevancia para una prueba de comprensión oral. El CIO no presenta los ítems en el mismo orden que la prueba original; por lo tanto, pasar por los archivos de audio de manera artificial para encontrar y tratar cada uno de los ítems individuales, sin haber tenido conocimientos previos de la prueba, habría sido de suma dificultad. El Consejo de Europa (2009) recomienda que, para una prueba de comprensión oral, los expertos deben estar provistos de un ordenador para poder hacer esta tarea. En cambio, es la firme convicción de la autora de que la metodología empleada en este estudio supone una solución mucho más pragmática al problema, y que cualquier estudio que emplee el método *Bookmark* debe realizar un estudio parecido al del método *Basket* primero. Un enfoque similar fue adoptado por Harsch y Hartig (2015), aunque su razonamiento era distinto; querían desvincular la tarea de separación de contenidos y la tarea de emparejamiento de dificultad de los ítems porque las consideraban tareas que requerían tipos de estimación muy diferentes.

De hecho, un resumen de estudios en los que se empleó el método *Bookmark* (Peterson, Schulz, y Engelhard, 2011) concluyó que los

Recibido: 11-05-2017 Aceptado: 02-02-2018

expertos confiaban en los puntos de corte resultantes. Indudablemente, los puntos de corte del presente estudio cuentan con el firme respaldo de la evidencia de validez externa que ha sido proporcionada. Dicha evidencia necesitará fortalecerse con su continuación en otros estudios en el futuro; con el tiempo, se podrían recopilar muestreos más amplios y quizás también proceder al empleo de un método del tipo *Prototype group*, similar al del estudio realizado por Eckes (2012).

Otro comentario que cabe mencionar es el hecho de que algunos ítems del CIO tenían unos valores empíricos de dificultad que no coincidían con las estimaciones de los expertos. Este fenómeno también se ha encontrado en estudios previos (p.ej. Figueras, Kaftandjieva, y Takala, 2013). Además, se ha señalado ampliamente que a menudo los expertos no son capaces de identificar la dificultad de un ítem de forma correcta (Alderson, 1993). El presente estudio no está libre de este problema, como evidencia la divergencia de los puntos de corte para el nivel B2 obtenidos por los métodos *Basket* y *Bookmark*. No obstante, es la convicción de la autora que dicha discrepancia se debe al hecho de que los participantes ya habían recibido mucha más información cuando tomaron sus decisiones finales empleando el método *Bookmark*. Como consecuencia, ellos pudieron tomar decisiones más informadas que tuvieron en cuenta las dificultades reales de los ítems, y no solo estimaciones subjetivas basadas en su opinión. Para el punto de corte a nivel B1, los expertos colocaron el marcador en un punto algo bajo del continuo de dificultad, debido a la presencia de ítems que creían ser más representativos de capacidades en umbrales entre B1 y B2. No obstante, después de aplicar el ajuste por la PR, el punto de corte para el nivel B1 obtenido por los dos métodos de establecimiento era muy parecido (menos de un error estándar), lo cual es una confirmación más de que la recopilación de múltiples fuentes de evidencia puede aumentar el nivel de confianza que tenemos en la decisiones cualitativas que tomamos.

Conclusión

Respecto al enfoque del argumento por la validez presentado por ALTE, el desarrollo de la prueba de la UGR se ha atenido a unas especificaciones detalladas y basadas en el MCER (inferencia de observación). Muestra *validez de puntuación* hasta el punto de que todas las tareas han sido

pilotadas y que los ítems muestran buenas propiedades sicométricas (inferencia de evaluación). El presente estudio ha reforzado la inferencia de extrapolación; se puede considerar que los puntos de corte son traducibles y consistentes y que los resultados se pueden emplear para establecer el nivel de capacidad en todas la futuras versiones de la prueba. Mediante el uso de un sistema de equiparación de ítems-ancla comunes (véase Kolen y Brennan, 2014; North y Jones, 2009; Wright y Stone, 1979), todas la tareas de la prueba de la UGR se pueden calibrar según la misma escala de medición *Rasch*; por consiguiente, la dificultad de la prueba será la misma en cada una de sus versiones, lo que respalda la inferencia de generalización.

Los puntos de corte no solo tienen consecuencias para los candidatos sino también para un amplio grupo de otras partes interesadas, como son los padres, educadores y legisladores. El establecimiento de puntos de corte es una parte fundamental del proceso de desarrollo de pruebas; por consiguiente, no solo debe tratarse como si fuera una actividad aislada sino también como un componente esencial e integral del continuado proceso de validación (Papageorgiou y Tannenbaum, 2016). En el contexto de la prueba de la UGR, esto incluye la implantación de futuros estudios para controlar las consecuencias de la aplicación de los puntos de corte que han sido decididos en el presente estudio, algo que forma parte de todo un abanico de controles de validez implantados por los desarrolladores de la prueba. De igual manera, se ha realizado un proceso de establecimiento de puntos de corte para la prueba de comprensión escrita de la UGR, y otros estudios parecidos se realizan con regularidad como parte de las propias sesiones internas de *benchmarking* (homogeneización de criterios) y formación en el CLM para las pruebas de producción escrita y oral. En efecto, la validación de la interpretación y uso de los resultados de la prueba binivel B1/B2 de la UGR es un proceso continuo: es precisamente esta recopilación de evidencia de validez la que proporciona el necesario respaldo para que todas las partes interesadas confíen en cualquier decisión que se fundamente en los resultados.

En España, los exámenes *CertAcles* han contribuido hasta cierto punto a responder a las recomendaciones hechas por Halbach, Lafuente y Guerra (2013) respecto a la acreditación de lenguas y la homogenización de criterios. Estas pruebas están diseñadas para que sean consistentes con el MCER, y exigen la provisión de evidencia de fiabilidad y validez.

Recibido: 11-05-2017 Aceptado: 02-02-2018

No obstante, como advierten Deygers et ál. (2017), la presión sobre los organismos de evaluación para que demuestren que sus pruebas se han adecuado con el MCER ha conducido al mal uso de pruebas en muchos contextos donde dicha afirmación no está corroborada. A tal respecto, sugeriría a todos los creadores y administradores de pruebas en el contexto actual que lleven a cabo estudios de establecimiento de puntos de corte para que los resultados oficiales de sus pruebas sean más significativos. Aunque muchas instituciones han visto el punto de corte como un estándar normativo (p.ej. el 60 %), los desarrolladores de pruebas deben cambiar esa perspectiva hacia el uso de puntos de corte que se puedan defender como representación del estándar de dominio, para que así podamos fortalecer la afirmación de que nuestras pruebas gozan de una buena vinculación con el MCER.

Referencias bibliográficas

AERA (American Educational Research Association), APA (American Psychological Association) & NCME (National Council on Measurement in Education). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Alderson, J. C. (1993). Judgements in language testing. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 46-57). Alexandria, VA: TESOL.

— (2012). Principles and Practice in Language Testing: Compliance or Conflict? Presentation at TEA SIG Conference: Innsbruck. Recuperado de http://tea.iatefl.org/inns.html

Alderson, J.C., Figueras, N, Kuijper, H., Nold, G, Takala, S & Tardieu, C. (2006). Analysing Tests of Reading and Listening in Relation to the Common European Framework of Reference: The Experience of The Dutch CEFR Construct Project, *Language Assessment Quarterly*, *3*(1), 3-30.

ALTE/Council of Europe (2011) Manual for Language Test Development and Examining. For use with the CEFR. Recuperado de http://www.coe.int/t/dg4/linguistic/ManualtLangageTest-Alte2011_EN.pdf

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly 2*(1), 1–34.

Bachman, L.F., & Palmer, A. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Brunfaut, T., & Harding, L. (2014). *Linking the GEPT listening test to the Common European Framework of Reference*. Taiwan: Language Training and Testing Centre.

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple... Language Testing, 29(1), 19–27.

Cizek, G. J. (2012). The forms and functions of evaluations in the standard setting process. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 165 - 178). New York: Routledge.

— (2016). Validating test score meaning and defending test score use: Different aims, different methods. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 212-225.

Cohen, A. S., Kane, M. T., & Crooks, T. J. (1999). A generalized examinee-centered method for setting standards on achievement tests. *Applied Measurement in Education*, *12*(4), 343-366.

Conferencia De Rectores De Las Universidades Españolas (CRUE). (2011, 8 de septiembre). Propuestas sobre la acreditación de idiomas. Recuperado de

http://www.acreditacion.crue.org/

Council of Europe. (2008). *Recommendation CM/Rec (2008)7 of the Committee of Ministers to member states on the use of the Council of Europe's Common European Framework of Reference for Languages (CEFR) and the promotion of plurilingualism*. Strasbourg, France: Council of Europe. Recuperado de http://www.coe.int/t/dg4/linguistic/Conventions_EN.asp

— (2009). Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment. Strasbourg, France: Council of Europe. Recuperado de http://www.coe.int/T/DG4/Linguistic/Manuel1_EN.asp

De Jong, J. & Zheng, Y. (2016) Linking to the CEFR: validation using a priori and a posteriori evidence. In, Banerjee, J. and Tsagari, D. (eds.) *Contemporary Second Language Assessment*. London, GB, Bloomsbury Academic pp. 83-100. (Contemporary Applied Linguistics, 4).

Recibido: 11-05-2017 Aceptado: 02-02-2018

Deygers, B., Zeidler, B., Vilcu, D., & Carlsen, C. H. (2017). One Framework to Unite Them All? Use of the CEFR in European University Entrance Policies. *Language Assessment Quarterly*, 1-13.

Downey, N. & Kollias, C. (2010). Mapping the Advanced Level Certificate in English (ALCE™) examination onto the CEFR. Aligning Tests with the CEFR, Reflections on using the Council of Europe's draft Manual, Martyniuk, W. (ed). Cambridge University Press. Cambridge. 119-129.

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Section H). Strasbourg, France: Council of Europe/Language Policy Division.

— (2012). Examinee-centered standard setting for large-scale assessments: The prototype group method. *Psychological Test and Assessment Modeling*, 54, 257–283.

Ferrara, S., & Lewis, D. (2012). The Item-Descriptor (ID) Matching method. In G. J. Cizek (Ed.), Setting performance standards: Foundations, methods, and innovations (2nd ed., pp. 255-282). New York: Routledge.

Figueras, N. (2012). The impact of the CEFR. *ELT journal*, *66*(4), 477-485.

Figueras, N., Kaftandjieva, F., & Takala, S. (2013). Relating a Reading Comprehension Test to the CEFR Levels: A Case of Standard Setting in Practice with Focus on Judges and Items. *Canadian Modern Language Review/La Revue Canadienne Des Langues Vivantes, 69*(4), 359-385.

Fulcher, G. (2004). "Deluded by artifices? The Common European Framework and harmonization." *Language Assessment Quarterly*, *1*(4), 253 - 266.

Fulcher, G., & Owen, N. (2016). Dealing with the demands of language testing and assessment. *The Routledge Handbook of English Language Teaching*, 109-120.

Green, A. (2017). Linking Tests of English for Academic Purposes to the CEFR: The Score User's Perspective, *Language Assessment Quarterly*, 1-16.

Halbach, A., Lazaro Lafuente, A., & Perez Guerra, J. (2013). La lengua inglesa en la nueva universidad española del EEES: The role of the English language in post-Bologna Spanish universities. *Revista de Educación*, 362, 105-132.

Hambleton, R., & Jirka, S. (2006) Anchor-based methods for judgmentally estimating item statistics. In S. Downing and T. Haladyna (Eds.), *Handbook of test development* (pp. 399-420). Mahwah, Nj: Erlbaum.

Hambleton, R., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Westport, CT: American Council on Education/ Praeger Publishers.

Harsch, C. & Hartig, J. (2015). What Are We Aligning Tests to When We Report Test Alignment to the CEFR?, *Language Assessment Quarterly*, 12:4, 333-362.

Kaftandijeva, F. (2010). *Methods for setting cut scores in criterion-referenced achievement tests. A comparative analysis of six recent methods with an application to tests of reading in EFL.* Cito, Arnhem: EALTA. Recuperado de www.ealta.eu.org/documentsresources/FK_second_doctorate.pdf

Kane, M. (2012). Validating score interpretations and uses. *Language Testing*, *29* (1), 3 –17.

— (2017). Using Empirical Results to Validate Performance Standards BT - Standard Setting in Education: The Nordic Countries in an International Perspective. In S. Blömeke & J.-E. Gustafsson (Eds.), (pp. 11–29). Cham: Springer International Publishing.

Kanistra, V. & Harsch, C. (2017). *Using the Item Descriptor Matching method to enhance validity when aligning test to the CEFR.* 4th Meeting of the EALTA CEFR Special Interest Group. Recuperado de http://www.ealta.eu.org/events/SIG_CEFR_london2017/presentations/Presentations/1400-1530/CEFR%20SIG%20Voula%20and%20Claudia.pdf

Kantarciouglu, E, Thomos, C, O'Dwyer, J & O'Sullivan, B. (2010) 'Benchmarking a high-stakes

proficiency exam: the COPE linking project', in W. Martyniuk (ed) *Aligning Tests with the CEFR: reflections on using the Council of Europe's draft Manual.* Cambridge University Press. Cambridge.

Kolen, M. J., & Brennan, R. L. (2014). Test equating, scaling, and linking: methods and practices (3rd ed.). New York: Springer-Verlag.

Linacre, J. M. (2017). Facets computer program for many-facet Rasch measurement, version 3.80.0. Beaverton, Oregon: Winsteps.com

Linn, R. L. (2003). Performance Standards: Utility for Different Uses of Assessments. *Education policy analysis archives*, *11*, 31.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Mitzel, H.C., Lewis, D.M., Patz, R.J., & Green, D.R. (2001). The bookmark procedure: Psychological perspectives. In G.J. Cizek (Ed), Setting

performance standards: Concepts, methods and perspectives (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum Assoc.

North, B., & Jones, N. (2009). Relating language examinations to the Common European Framework of Reference for Languages. Further material on maintaining standards across languages, contexts and administrations by exploiting teacher judgment and IRT scaling. *Strasbourg: Language Policy Division.*

Papageorgiou, S. (2010). Investigating the decision-making process of standard setting participants. *Language Testing*, 27(2), 261–282.

Papageorgiou, S., & Tannenbaum, R. J. (2016). Situating Standard Setting within Argument-Based Validity. *Language Assessment Quarterly*, *13*(2), 109–123.

Peterson, C. H., Schulz, E. M. and Engelhard Jr., G. (2011), Reliability and Validity of Bookmark-Based Methods for Standard Setting: Comparisons to Angoff-Based Methods in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice*, 30: 3–14.

Reckase, M. D. (2010). NCME 2009 presidential address: What I think I know. *Educational Measurement: Issues and Practice, 29*(3), 3–7.

Tannenbaum, R, J. & Cho, Y. (2014). Critical Factors to Consider in Evaluating Standard-Setting Studies to Map Language Test Scores to Frameworks of Language Proficiency, *Language Assessment Quarterly, 11*(3), 233-249.

Tannenbaum, R, J. & Wiley E. C. (2008). Linking English-language test scores onto the Common European Framework of Reference: An application of standard-setting methodology. *ETS Research Report Series*, *2008*(1).

Weir, C. J. (2005). Limitations of the council of Europe's framework of reference (CEFR) in developing comparable examinations and tests. *Language Testing*, *22*(3), 281–300.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA.

Información de contacto: Caroline Shackleton. Centro de Lenguas Modernas de la Universidad de Granada. Placeta del Hospicio Viejo s/n 18009, Granada, Spain. E-mail:csarah@ugr.es

Linking the University of Granada *CertAcles* listening test to the CEFR[1]

Vinculación de la prueba de comprensión oral del examen *CertAcles* de la Universidad de Granada con el MCER

DOI: 10.4438/1988-592X-RE-2017-381-380

Caroline Shackleton
Universidad de Granada

Abstract

As part of the ongoing collection of validity evidence to support the interpretation and use of the UGR bi-level English test scores this study aims to formally map the oral comprehension part of the test to the Common European Framework of Reference (CEFR). The advice provided by the Council of Europe (2009) was followed in order to determine cut scores which are representative of CEFR B1 and CEFR B2 levels. The test used in the study was from the March 2017 administration (N=464) which was shown to be a reliable and appropriate measurement instrument. The study involved ten participant judges who, after a CEFR familiarisation session, took part in standard setting sessions. Two methods of standard setting were employed. The *Basket* method was used mainly in order to familiarise the participants with the test items and was analysed using a Many-Facet Rasch measurement model. The *Bookmark* method allowed for the incorporation of item difficulty parameters produced by a Rasch analysis of test scores. The final median cut scores of the judges using a response probability of .67 were mapped back to the Rasch theta ability scale in order to determine the raw scores which represent the two CEFR ability levels. Several other validity and reliability checks were also carried out following CoE (2009) and transparent documentation on the whole process is provided. By using Rasch common item equating, the scores are reproducible on future versions of the test and so support the *generalisation inference* of the UGR test's validity argument.

(1) Acknowledgements: The author would like to thank the Centro de Lenguas Modernas of the University of Granada for supporting this study, and the two anonymous reviewers for their helpful comments and suggestions.

Received: 11-05-2017 Accepted: 02-02-2018

Keywords: proficiency tests (language), standard setting, validity, item response theory

Resumen

Este estudio, como parte del proceso de validación fehaciente del uso e interpretación de las puntuaciones del examen de inglés binivel de la Universidad de Granada (UGR), pretende relacionar, oficialmente, la prueba de comprensión oral de este examen con el Marco Común Europeo de Referencia (MCER). Con tal fin seguimos las recomendaciones del Consejo de Europa (2009) para determinar unos puntos de corte representativos de los niveles B1 y B2 del MCER. La prueba utilizada en el estudio fue la convocatoria de marzo del 2017 (N=464), que demostró ser un instrumento de medición fiable y apropiado. En el estudio participaron diez expertos, quienes, tras una sesión para familiarizarse con el MCER, realizaron varias sesiones de fijación de puntos de corte. Se utilizaron dos métodos: el método *Basket,* o de la cesta, para familiarizar a los participantes con los ítems de la prueba. después del cual se analizaron los resultados empleando un modelo de múltiples facetas de Rasch (MFR); por otro lado, el método *Bookmark*, o del marcador, que incorporó parámetros de dificultad del ítem, resultantes de un análisis Rasch de los puntajes de la prueba. La mediana final de los puntos de corte aportados por los expertos se calculó empleando una probabilidad de respuesta de 0.67, y se transfirió a la escala de capacidad zeta del modelo Rasch para determinar la puntuación bruta que representarán los dos niveles de dominio del MCER en cuestión. También se realizaron otros controles de fiabilidad y de validez para proporcionar una documentación trasparente sobre el proceso completo, recomendado por el Consejo de Europa (2009). El método de equiparación de puntuaciones del modelo Rasch permite que los puntajes sean reproducibles en futuras versiones de la prueba y, por tanto, apoyan la *inferencia de generalización* del argumento de validez para la prueba binivel de la UGR.

Palabras Clave: pruebas de competencia lingüística, fijación de puntos de corte, validez, la teoría de respuesta al ítem

Introduction

As part of continuing policy adaptation to the Bologna process and the introduction of the European Higher Education Area, recent Spanish Ministry of Education policy has placed emphasis on communication in a foreign language as a fundamental competence for enabling

Received: 11-05-2017 Accepted: 02-02-2018

greater student mobility and employability. To this end, any national test must therefore be designed in such a way as to allow the linking of candidate performance to CEFR competence levels if it is to respect principles of transparency, comparability and coherence. Such principles are unquestionably of the utmost importance in the promotion of educational and professional mobility within the EU in accordance with current European plurilingualism policies. Indeed, the recommendation on the use of the CEFR by the Council of Ministers includes the call for countries to:

> '…ensure that all tests, examinations and assessment procedures leading to officially recognised language qualifications take full account of the relevant aspects of language use and language competences as set out in the CEFR, that they are conducted in accordance with internationally recognised principles of good practice and quality management, and that the procedures to relate these tests and examinations to the common reference levels (A1-C2) of the CEFR are carried out in a reliable and transparent manner.' (Council of Europe, 2008, p.4)

In Spain, many university language centres have responded to this call by providing accreditation exams following international guidelines in order to produce valid and reliable CEFR-related tests. Furthermore, in an attempt to better coordinate and regulate such efforts, the Association of Higher Education Language Centres (ACLES) has provided a procedure for nationally recognised, CEFR-related certification: the *CertAcles* exams. These exams have not only been approved by the Committee of University Rectors (CRUE 08/09/2011) but have also been recognised at an international level by the *CercleS* organisation since 2014. The *CertAcles* model (http://www.acles.es/es) requires that each language use ability is tested separately, should be developed from CEFR informed specifications, and that quality control information following the recommendations of *the standards* (*AERA, APA and NCME,* 2014) be provided.

The University of Granada's *Centro de Lenguas Modernas* (CLM) administers a bi-level B1/B2 *CertAcles* exam three times a year, provided free of charge for University of Granada (UGR) students. Developed in conjunction with EALTA expert member Dr. Rita Green, and initiated in 2009, it is now a stable exam system which follows all the validity checks

recommended by ACLES. The present study will outline and report on the procedures for CEFR linking and cut score determination currently in use on the listening section of the UGR bi-level test at B1 and B2 levels.

Literature review

The CEFR and its six descriptor scales (A1 through to C2) has become the standard currency in language proficiency levels both in Europe and beyond (Deygers, Zeidler, Vilcu, & Carlsen, 2017; Figueras, 2012). Through the association of a test score with a CEFR level, meaning may be added for test users (Kane, 2012). For example, if a score reports B1 proficiency in listening, the score user is informed about the kind of activities a person receiving such a result should be able to perform (Tannenbaum & Cho, 2014). As such, tests which claim to measure CEFR competencies must provide evidence to support these claims: the test must be valid.

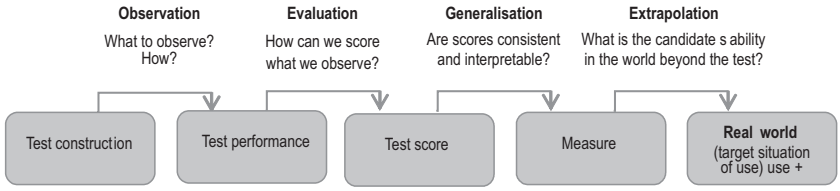
Validity is seen not as an innate property of the test — a test cannot be validated (Chapelle, 2012; Cizek, 2016) — but rather as a unitary, multi-faceted concept which draws upon multiple sources of evidence in order to substantiate the inferences made about a candidate's performance and so justify the resulting decisions taken on the basis of test scores (Messick, 1989). Evidence to support the interpretation and use of test scores is necessary, and a *validity argument-based* approach, which provides a framework to guide testing bodies on the types of evidence to be collected, is therefore recommended (*AERA, APA and NCME,* 2014; Bachman, 2005; Bachman & Palmer, 2010; Chapelle, 2012; Kane, 2012).

The types of evidence to be accumulated should be driven by test purpose (Fulcher & Owen, 2016), and in the European context such a framework has been provided by ALTE (2011) as an aid to testing bodies in the design and development of CEFR-related tests. This document outlines evidence which should be collected throughout the test development cycle (see figure 1) with the goal of building test validation into the whole process of test design and implementation. In contrast to recommendations by Bachman (2005) and Bachman and Palmer (2010), the chain of reasoning begins with the test construct rather than the decisions based on test scores. As such, it links performance on test tasks to an inference about candidate's language ability beyond the test

Received: 11-05-2017 Accepted: 02-02-2018

and the argument ends with the 'extrapolation inference'. However, the world beyond the test remains unspecified: it is a general claim linking to the CEFR. We can therefore argue that the ALTE model directs our interest to the interpretation of the score, which is to ask, 'can the score be interpreted as a CEFR proficiency level?'. This would follow the conceptualisation of validity proposed by Cizek (2016), who provides a framework that differentiates the validation of test score inferences and justification for test use.

FIGURE I. Chain of reasoning in a validity argument



Source: ALTE (2011, p. 15)

Within such an argument approach, 'validity evidence of the cut score is an essential component' (Papageorgiou & Tannenbaum, 2016, p.110). Candidates must meet the cut score to demonstrate that they have reached the required proficiency level, and the results of a test would clearly be open to question unless such a cut score was appropriately set. Here, the Council of Europe (2009) provides a manual which outlines the process that testing bodies should follow. The linking process consists of the following four interrelated activities necessary for relating test scores to the CEFR:

- **Familiarisation.** Members of any linking panel must be familiar with the content of the CEFR and its scales.
- **Specification.** This should include a detailed description of the test and its relationship to CEFR categories in order to build a linking claim about the content relevance of the exam to the CEFR.
- **Standardisation and Standard setting.** For any test, a cut score or score which is needed for passing the test needs to be decided.

Received: 11-05-2017 Accepted: 02-02-2018

- **Empirical validation.** CoE (2009) outlines three types of validity evidence – procedural, internal, and external – which must be provided as part of the standard setting process.

Since its publication numerous testing bodies have followed the advice of the manual in order to claim CEFR-linkage. Most international English language proficiency exam providers have carried out such studies (see for example Brunfaut & Harding, 2014; Kanistra & Harsch, 2017; Tannenbaum & Wylie, 2008) and smaller national exam providers also claim CEFR calibration by following the steps in the manual (see for example Downey & Kolias, 2010; Kantarcioglu, Thomas, O'Dwyer, & O'Sullivan, 2010). The main focus of these studies is on the standard setting procedure and its empirical validation. As CoE (2009, p.11) put it 'the crucial point in the process of linking an examination to the CEFR is the establishment of a decision rule to allocate students to one of the CEFR levels on the basis of their performance in the examination', that is to say the determination of the cut score.

In sum, a valid claim for CEFR linkage must show that the test is reliable and representative of proficiency at a CEFR performance level as 'if an exam is not valid or reliable, it is meaningless to link it to the CEFR' (Alderson, 2012). Consequently, the process is normally carried out late in the test development cycle. The UGR test provides just such evidence, and the present study will focus on the official linking process as laid out by CoE (2009) in order to create valid cut score decisions for B1 and B2 mastery for its bi-level listening paper.

Methodology

The literature classifies standard setting techniques into two main types: test- and examinee-centred. For the present study, no information about candidates taking the test was available, and it was therefore decided that the main study would use a test-centred method. Nevertheless, information about a small number of test takers previously enrolled on an exam preparation course at the CLM was also used for comparison in order to give some supporting triangulation evidence (and therefore evidence towards external validity).

Of the various test-centred methods, it was decided not to use *Angoff-type* probability methods. Not only are these methods time-consuming,

Received: 11-05-2017 Accepted: 02-02-2018

but previous studies report problems concerning judges' inability to understand and correctly articulate conditional probability (Ferrara & Lewis, 2012; Hambleton & Jirka, 2006; Reckase, 2010). It was decided instead to use a combination of two other methods currently popular in the literature: the *Basket method* and the *Bookmark method*.

The *basket method* is effectively an *item descriptor matching method* specifying the level of competence necessary to answer each item on the test. Judges analyse test items and answer the question, 'what CEFR level must a test taker be to answer this item?' It is considered to be the most simple and practical of all standard setting methods and is one which both reflects the importance of the performance level descriptors and places emphasis on test content. However, no information is provided about the difficulty of the items and as such, one of the main problems is a lack of consistency when comparing judgements to empirical difficulty measures (Kaftandjieva, 2010). However, the method was used in the present study as a means of familiarizing judges with the test content and for promoting discussion, rather than to set the final standard. The results of this part of the study were analysed using a Many-Facet Rasch measurement model (MFRM) in the programme FACETS (Linacre, 2017). While it has been argued that MFRM is not appropriate at the later stages of standard setting, where efforts are being made to reach consensus (Eckes, 2009), this was deemed irrelevant to the present study as it would only be used for analysis at the initial familiarisation stage.

If the process of setting cut scores is to have real meaning, judges must be given statistical information about test items. In the *Bookmark method* (Mitzel, Lewis, Patz, & Green, 2001), items are placed in an ordered item booklet (OIB), ordered by their difficulty parameters. Judges move through the OIB booklet from easiest item to most difficult, placing a bookmark at the point where they believe a minimally competent candidate will have less than a specified probability of giving a correct response. The candidate is considered to have the ability to master an item in probabilistic terms, known as the response probability (RP), which should be decided in advance, but it is often set at 0.67, a 67%, or 2/3, chance of the candidate answering the item correctly. The bookmark is placed between two items on the latent scale represented by the Rasch logit difficulty parameters of the items (β-parameters). Finally, the median of judges' scores is taken to give the group standard on the latent variable. The Bookmark method has the advantage of both being easily

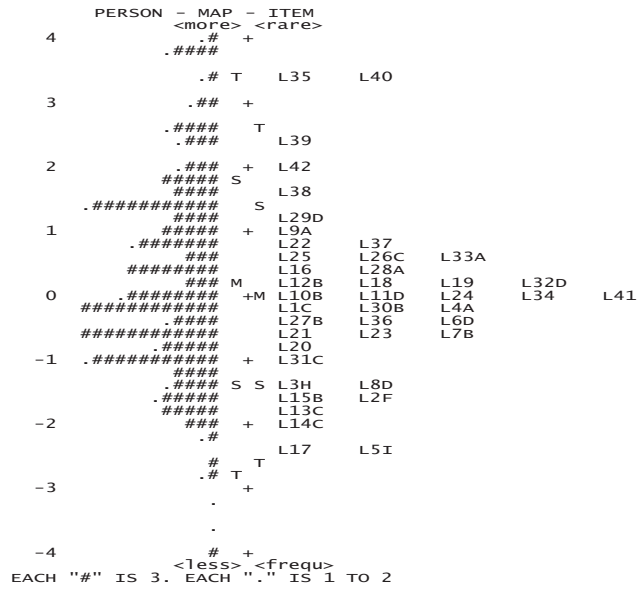
Received: 11-05-2017 Accepted: 02-02-2018

used with tests developed using Rasch measurement, where test taker ability and item difficulty are placed on the same scale, and of clearly reflecting the continuous nature of the CEFR scales, making it ideal for use in the present study.

Instruments

The test used in the present study was taken from the March 2017 administration. The listening paper (42 items) consists of five tasks - two at B1, one at B1/B2 and two at B2 level. All tasks were developed following the test specifications and in collaboration with our expert consultant and may therefore be considered as having been expertly judged to have the correct level of difficulty. Furthermore, they have all been previously piloted (N = 144 to 434) and items found not to function correctly have been removed or revised. The construct is well sampled based on B1 and B2 CEFR descriptors and a mix of test methods are used.

FIGURE II. Item variable map



Source: Created by the author

Received: 11-05-2017 Accepted: 02-02-2018

A classical test analysis of the results (where $N = 464$) gave a Cronbach's Alpha of 0.934 (SEM = 2.5), evidencing test reliability. The Rasch analysis of test scores showed the test to be unidimensional, with Rasch *Infit Mean Square (MNSQ)* statistics for all the items falling between the acceptable values of 0.74 and 1.24. The item separation is 10.2 and item separation reliability is 0.99. Therefore, the test can be said to include a good spread of item difficulties and we can be confident that these difficulty estimates are reproducible. The item variable map reproduced in figure II shows ordered item difficulty along with candidate ability. This information was used to create the OIB for the Bookmark method study. Here, the mean logit values for item difficulty and candidate ability are very similar, demonstrating that the test is appropriate in terms of difficulty for the population tested. As the present study does not intend to investigate the test itself, the above information is considered sufficient as evidence that the measurement instrument used in the study is both reliable and appropriate.

In order to provide evidence for procedural validity a post standard setting questionnaire was administered following Cizek (2012). The questionnaire contained a four-point likert scale to eliminate the possibility of neutral responses. This allowed for the collection of judge's opinions concerning their understanding of the whole process and their overall beliefs about the final cut score judgements.

Subjects

It is important that the participants have the relevant expertise in both the CEFR framework and hence its performance level descriptors, as well as in the instruction and assessment of the language being tested (Tannenbaum & Cho, 2014). This is indeed the case of the ten judges who took part in the study. As table I shows, all are experienced TEFL teachers at the CLM and are well-acquainted with the CEFR. Furthermore, they are familiar with the examinee population, and appreciate the consequences of the recommended cut scores.

Received: 11-05-2017 Accepted: 02-02-2018

TABLE I. Background information about judges

Judge	Sex	Qualifications	Number of years teaching TEFL	Number of hours specific training CEFR/Assessment
1	Female	Degree, MA	35	300+
2	Male	Degree	28	80
3	Male	Degree, Postgraduate TEFL diploma	35	100+
4	Male	Degree, MA, CELTA	16	70
5	Male	Degree	22	70
6	Female	Degree, MA	36	400+
7	Female	Degree, PGCE, PhD	28	750+
8	Male	Degree, MA, CELTA	18	500+
9	Female	Degree, PGCE, MA	28	750+
10	Female	Degree	22	500+

Source: Created by the author

Results

The results of each phase of the standard setting sessions, including those results pertinent to the familiarisation stage and external validation, will now be reported.

Familiarisation

In preparation for the study, all the judges were asked to carry out the training sessions for listening provided by the *Ceftrain* project (see http://www.helsinki.fi/project/ceftrain/index.html). A CEFR familiarisation descriptor sorting exercise was carried out in which the judges were given a mix of listening descriptors from the CEFR scales and asked to allocate a CEFR level for each one. The results gave a Cronbach's Alpha of 0.98. Table II shows the correlation between CEFR descriptors and participants' judgment using a non-parametric Spearman's rank correlation coefficient. These results were presented to the participants

Received: 11-05-2017 Accepted: 02-02-2018

and further discussion took place in order to reach consensus on salient features of CEFR levels.

TABLE II. Correlation between CEFR descriptors and judges allocations

Judge	1	2	3	4	5	6	7	8	9	10
	.851	.875	.835	.889	.896	.836	.957	.985	1.00	.870

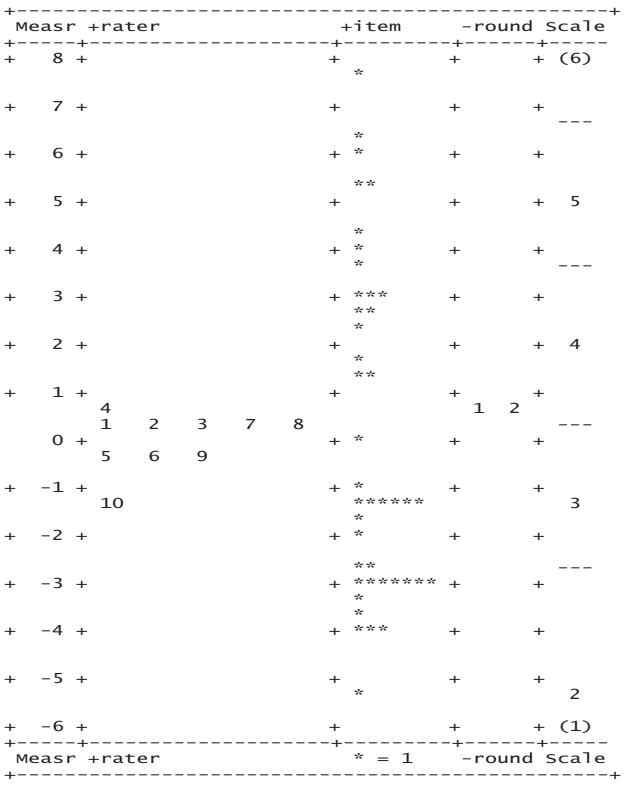
(Note: all correlations were significant at $p \leq .01$)
Source: Created by the author

Basket method of standard setting.

Judges were first given an input session on the exam construct and test specifications, which was followed by an explanation of the Basket method itself. Figure III shows an MFRM analysis of the results from all judges for each item on the test. The analysis allows for multiple aspects of the judgements to be taken into account, and calibrates items, judges and the rating scale onto the same equal-interval scale. The scale used to represent the minimum CEFR level that a candidate should have in order to answer each item is as follows: 1=A2, 2=minimum B1, 3=strong B1, 4= minimum B2, 5=strong B2, 6=C1. The results from the two rounds differed only minimally, showing that the judges were confident in their decisions.

It can be clearly seen that the judges believed the content of the items to be well spread along the ability scale from CEFR B1 to CEFR B2. The Rasch Andrich thresholds showed that judges placed items into distinct categories, which increase in difficulty level. These categories are very much in agreement with the test developers' intentions. The judges differed slightly, with judge 10 placing the items at a marginally lower difficulty level. However, the MFRM analysis takes such variations into account and, by making adjustments for the leniency or severity of the judges, is able to give a 'fair average' score for each item. Using the Rasch Thurston thresholds, which measure the boundary at which an item has a 50% chance of being rated in a higher or lower category (i.e. the 50% (median) cumulative probability threshold), we reach preliminary cut scores of 13 for B1 and 34 for B2.

FIGURE III. Vertical ruler from FACETS analysis for Basket method results



Source: Created by the author

Bookmark method of standard setting.

Before moving on to the main study, participants were given feedback, including impact data, from the Basket method. Discussion was finalised with the general conclusion that the judges felt the B2 cut score to be rather high (only 81 students, or 17%, would receive a B2). At this point, it was explained that these would not be the final cut scores and that the Basket exercise was simply intended to allow the participants to become familiar with the test before seeing the real item difficulty values that would be used to allocate cut scores.

Received: 11-05-2017 Accepted: 02-02-2018

The bookmark method was then explained to the participants and they were given the OIB. After the first round, the judge's decisions were presented to the group along with impact data. This was followed by further discussion before they moved on to the second round, where they were permitted to change the position of the bookmark. Results of both rounds are shown in table III [2].

TABLE III. Cut score results for bookmark method

Round 1	Page number (number of judges)	Mean (SD)	Median	Ability (θ) after adjusting for RP of .67	Final cut score
B1	8/9 (3) 10/11 (2) 12/13 (2) 13/14 (1) 15/16 (1)	11 (2.87)	10/11	-.13 logits	22
B2	25/26 (1) 28/29 (2) 29/30 (5) 30/31 (1)	28.44 (1.42)	29/30	1.04 logits	29
Round 2					
B1	8/9 (7) 9/10 (1) 10/11 (1)	8.33 (.71)	8/9	-.70 logits	16
B2	28/29 (6) 29/30 (3)	28.56 (.73)	28/29	.93 logits	28

Source: Created by the author

The median of the second set of bookmarks for all panellists is used as the group standard, and the lower value of theta should be taken in order to set the cut score (CoE, 2009). Here, the median corresponds to an ability level of $\theta = -.7$ logits for B1 and $\theta = 0.93$ logits for B2 once adjustments have been made for the RP of 0.67. This ability level is equivalent to raw scores of 16 for B1 and 28 for B2 when mapped onto the test characteristic curve.

(2) One panellist could not be present for the bookmark method and so this is the result for nine judges.

Received: 11-05-2017 Accepted: 02-02-2018

A final discussion then took place which considered both impact and external validity data. Participant judges were seen to be happy with the final decisions and in agreement that these should be used as the final cut scores for the test. The resulting impact data for the March 2017 administration give the final classifications of 135 (29%) fail, 166 (36%) CEFR B1 and 163 (35%) CEFR B2.

Procedural validity

The results from the post standard setting questionnaire are shown in table IV. For each question, participants answered a 4-point likert scale in which 1 represents complete agreement. Overall a very high level of agreement can be seen. The panellists felt confident about the procedure and their final decisions. Most notably, all panellists felt that the final cut score decisions were a fair and accurate representation of CEFR B1 and B2 for listening.

TABLE IV. Questionnaire results

	Mean	SD
I felt I had a sound understanding of CEFR levels for listening after the familiarisation sessions.	1.3	.48
The explanation of the test construct and specifications helped me.	1.5	.71
I understood the Basket method standard setting procedure.	1.2	.42
I felt confident about my decisions answering the question 'At what CEFR level must test taker be in order to answer this item.	1.6	.52
I understood the concept of 'minimally competent candidate'.	1	.00
The Basket method standard setting exercise helped me to understand test content.	1.2	.42
I understood the Bookmark method standard setting procedure.	1.33	.71
I understood the concept of RP and what the 67% probability means.	1.11	.33
I felt confident about the placing of my bookmark on round 1.	1.67	.50
I felt confident about the placing of my bookmark on round 2.	1.11	.33
The final recommended cut scores of the group are a fair representation of CEFR B1 and B2 levels for listening.	1	.00

Source: Created by the author

Received: 11-05-2017 Accepted: 02-02-2018

Internal validity

The internal validity of the standard setting procedure concerns the accuracy and consistency of results; the quality of judgments needs to be determined and reported. For the Basket method the results showed judges to be internally consistent with a mean *Infit MNSQ* of 0.98 and SD of 0.28. There is one *overfitting* judge and one *underfitting* judge shown by *Infit MNSQ* and *Zstd*. However, only one judge shows a slight *underfit* in terms of *outfit* and exact and expected observed agreements were very similar, as would be expected by raters acting as independent experts (Linacre, 2017).

In an attempt to provide stronger CEFR linkage, the decisions made by the judges about item difficulty using the Basket method were also compared with the actual item difficulties provided by the Rasch analysis. This analysis shows further evidence of the quality of the standard setting judgements. The data set is small, with tied ranks for the judged items and a Kendall's Tau correlation $\tau = .57$, $p < .001$ shows a significant relationship between judged item difficulty (using median values from Round 2) and Rasch difficulty parameters. This result gives evidence that the judges recognized the increasing difficulty of the items.

The classification accuracy of the cut scores needs to be evidenced. The final judged cut scores after Round 2 using the Bookmark method showed a high level of consensus with small SDs. The validity of these cut scores can be evidenced by reporting the standard error of judgement (SE_J), which is 'an estimate of the likelihood of replicating the recommended cut scores' (Tannenbaum & Cho, 2014, p.245). Cohen, Kane and Crooks (1999, p.364) argue that the SE_J should be ≤ 1⁄2 SEM. The results for this calculation for both rounds can be seen in table V.

TABLE V. Results for classification accuracy

	Round 1		Round 2	
	B1	**B2**	**B1**	**B2**
Standard deviation of mean cut score (SD)	2.87	1.42	0.71	0.73
Standard error of test (SEM)	2.5	2.5	2.5	2.5
Standard error of judgement (SE_J)	1.01	0.50	0.25	0.26
	0.4	0.2	0.1	0.1

Source: Created by the author

Received: 11-05-2017 Accepted: 02-02-2018

Thus, the SE_J was always smaller than one-half of SEM, and the cut scores fulfil the quality criterion. It should, however, be noted here that it has been argued that final round judgements do not usually have much variability because judges are encouraged to converge and come to a consensus decision (Linn, 2003).

External validity

External validity refers to the use and comparison of different standard setting methods, as well as comparisons or triangulation with measures from other studies. The present study includes two standard setting methods. As reported in previous studies, the two procedures produced different results for both the B1 and B2 cut scores. At B2, the Rasch score to measure table shows that there was a certain degree of discrepancy, with the difference at slightly more than *two standard errors*. At B1, however, the Basket cut score result falls within one *standard error* of the Bookmark result, giving evidence that the two standard setting methods produced quite consistent results on this part of the test, despite the Basket method being only used for content familiarisation.

The results were further analysed by comparing the test scores of those candidates who had undertaken exam preparation courses at CLM with a predicted ability level provided by their teachers. Despite the small group of candidates in question, by checking the classification accuracy of these students, this comparison allowed for a certain amount of triangulation and further contributed to validity evidence. The results can be seen in table VI.

This small-scale study based on a teacher judgement criterion shows that there is an extremely high correlation between teacher prediction and score on the test $\tau = .85$, $p < .001$. There is a classification agreement of 100% for all actual predictions of B1 and B2 and this gives us some external validity evidence about the accuracy of the cut scores. It can be seen that of the candidates predicted by their teachers to be borderline at both A2/B1 and B1/B2 level, about 50% of candidates were granted the higher level. It should, however, be understood that the present study is small in scale; ideally, better external validity evidence would be gained from a much larger-scale study.

Received: 11-05-2017 Accepted: 02-02-2018

TABLE VI. Comparison of teacher predictions and cut score decisions

Predicted CEFR Level Listening	Score on test	CEFR level received on test
Fail	1	Fail
Fail	11	Fail
A2/B1	12	Fail
A2/B1	12	Fail
A2/B1	14	Fail
A2/B1	17	B1
A2/B1	19	B1
B1	17	B1
B1	19	B1
B1	19	B1
B1	22	B1
B1	22	B1
B1	23	B1
B1	25	B1
B1	26	B1
B1	27	B1
B1	27	B1
B1/B2	26	B1
B1/B2	27	B1
B1/B2	33	B2
B1/B2	33	B2
B1/B2	36	B2
B2	30	B2
B2	31	B2
B2	31	B2
B2	35	B2
B2	36	B2
C1	40	B2

Source: Created by the author

Received: 11-05-2017 Accepted: 02-02-2018

Discussion

This study has given evidence for a stronger claim on the interpretation and use of test scores in terms of CEFR alignment for the UGR B1/B2 test. The standard setting process has been described in detail as the core part of the linking process and, in doing so, it has been documented that a reasonable and systematic process was followed in order to reach the final standard. Good validity evidence has been provided to support the setting of recommended cut scores, in which the standard setting process may be considered a 'blend of judgment, psychometrics, and practicality' (Hambleton & Pitonoak, 2006, p.435), and where 'the question is not whether the cut score is correct but whether decisions based on the cut scores are reasonable, broadly acceptable and have mostly positive consequences' (Kane, 2017, p.11).

The basis of the CEFR alignment manual (CoE, 2009) is without doubt the use of expert judgements, a methodology which is widely used and recommended in the standard setting literature. However, methodologies using expert judgement have been reported as unreliable, arguably due to the fact that the CEFR does not provide sufficient and precise descriptions of proficiency levels (Alderson et al., 2006; Fulcher, 2004; Weir, 2005). Judges in an alignment study may interpret CEFR descriptors differently or have their own internalized idea of just what it means to be at a CEFR level (Eckes, 2012; Harsch & Hartig, 2015; Papageorgiou, 2010). Indeed, North and Jones (2009, p.16) state, '..no amount of CEFR familiarisation and standardization, or estimation of indices of consistency and agreement, will prove that a given group of experts judging the level for a given language are not bringing their own culturally determined interpretation to the task'.

Despite these limitations, however, the judges in the present study expressed satisfaction with their results and believe that the cut scores are representative of a candidate who can perform tasks associated with the level. The study has therefore provided further evidence towards the test's substantive and construct validity. Both methodologies used required extensive discussion about matching individual item difficulty to the CEFR levels, and as such strengthens the claim that test content is CEFR-aligned, something which was also reported by Kanistra and Harsch (2017) in the Trinity ISE linking study. Indeed, the problem of defining a just-qualified candidate is a much easier task for a test

which has been developed to operationalize the CEFR model because the standard has already been built in, giving more meaning to the cut scores (Tannenbaum & Wylie, 2008). Conversely, a retro-fit type study would need to pay particular attention to the content specification stage of CEFR alignment in order to provide evidence that the content of the test measures the language skills described by the framework (Tannenbaum & Cho, 2014); various CEFR-linking studies have reported problems encountered when matching content to descriptors (e.g. Brunfaut & Harding, 2014). Furthermore, it has been argued that very few test providers pay attention to this stage of CEFR linkage (Green, 2017); if content is not aligned to the CEFR there is little justification for conducting a standard setting study. I would highlight here that the UGR test has been specifically developed to be representative of the CEFR; as such, the study is not a post-hoc linking study but forms part of the test development project itself, with the standard already implicitly built into the test. CEFR descriptors have been included in the test specifications and guidelines for item writers include a number of the categories presented in the CoE (2009) specification forms. This is similar to the *a priori* approach reported for the Pearson Test of English Academic (De Jong & Zheng, 2016).

In terms of the methodology used in the study, there is no best method of standard setting; the methods should be matched to suit the situation. As has been noted, the main Bookmark study is considered to be appropriate for use in test situations based on the Rasch measurement model of test development. Furthermore, the Basket method study was found to be invaluable as a primer study in that it allowed judges to become familiar with the test content. Here it would have been almost impossible to begin directly with a Bookmark study, and this is especially true for a listening test. The OIB does not present the items in the same order as the original test; as such it would have been extremely difficult to deal with separate items without prior knowledge of test tasks, moving artificially through audio files in an individual and non-linear fashion to target the item in question. CoE (2009) recommends that for a listening test, judges should be provided with a computer in order to do this. However, it is the author's firm belief that the methodology used in this study provides a far more pragmatic solution to this problem, and that any study using the Bookmark method for a listening test should first carry out a Basket method-type study. A similar approach was used by Harsch

and Hartig (2015), though their reasoning was different; they wanted to separate the content matching task from the item difficulty matching task because they were considered to be very separate judgment tasks.

Furthermore, a review of studies using the bookmark method (Peterson, Schulz, & Engelhard, 2011) reported that judges feel confident about the resulting cut scores. Certainly, in the present study the cut scores are well supported by the external validity evidence. This evidence will need strengthening with future studies; larger data samples could be collected over time and perhaps a prototype group method, such as the one reported by Eckes (2012), could be employed.

One other noteworthy comment was the fact that some items had an empirical difficulty value in the OIB which did not coincide with the judges' opinions. This phenomenon has been found in previous studies (e.g. Figueras, Kaftandjieva, & Takala, 2013). Also, it has been widely reported that judges are often unable to correctly identify item difficulty (Alderson, 1993). The present study is not immune to this, as is evidenced by the divergence in the B2 cut score using the Basket versus the Bookmark method. However, it is the author's belief that this difference may be explained by the fact that participants had subsequently received a lot more information when making their final decisions using the Bookmark method. Consequently, their decision was very much more informed and took into account actual item difficulty estimations as opposed to subjective, opinion-based estimations of difficulty only. For the B1 cut score, the judges placed the bookmark quite low on the continuum of difficulty due to the existence of items which they believed were more representative of high B1/B2 mastery. Nevertheless, after applying the RP adjustment, the final B1 cut score was very similar (within one standard error) using both methods of standard setting, further confirmation that the collection of multiple sources of evidence can increase confidence in qualitative decisions.

Conclusion

As regards the validity argument approach presented by ALTE, the UGR test has been developed following detailed test specifications based on the CEFR (observation inference). It shows *scoring validity* to the extent that all test tasks have been piloted and items show good

Received: 11-05-2017 Accepted: 02-02-2018

psychometric properties (evaluation inference). The present study has further strengthened the test's extrapolation inference; cut scores can be considered to be both interpretable and consistent and the results can be used to set the ability level on all future versions of the test. Through the use of a system of common item equating (see Kolen & Brennan, 2014; North & Jones, 2009; Wright & Stone, 1979), all tasks on the UGR test can be placed on the same Rasch measurement scale; test difficulty will subsequently be the same on every version of the test, which adds support for the generalisation inference.

Cut-scores have consequences not only for candidates but also for a whole range of other stakeholders, such as parents, educators, and educational policy makers. Standard setting is a fundamental part of the test development process; it should not therefore be treated simply as an isolated event but must instead be considered an essential and integral component of the continuing validation process (Papageorgiou & Tannenbaum, 2016). For the UGR test, this includes the implementation of future studies to monitor the consequences of applying the cut scores determined from the present study, which forms part of a whole range of validity checks implemented by the test developers. Similarly, a standard setting process has also been carried out for the UGR reading test, and similar studies are regularly carried out as part of the internal benchmarking and training sessions at the CLM for the speaking and writing parts of the test. Indeed, the validation of the interpretation and use of test scores for the UGR bi-level test is an ongoing process: it is just such an accumulation of validity evidence which provides the necessary backing for stakeholders to be confident of any decisions made based on test scores.

In Spain, the *CertAcles* exams have gone some way to responding to the recommendations made by Halbach, Lafuente and Guerra (2013) concerning language accreditation and the unification of criteria. These tests are designed to be consistent with the CEFR and require that reliability and validity evidence be provided. However, as Deygers et al. (2017) have warned, the pressure for testing bodies to report CEFR alignment has lead to test misuse in many contexts where this claim is not substantiated. Here, I would urge all test providers in the current context to carry out standard setting studies so as to give greater meaning to their reported scores. While traditionally many institutions have viewed the cut score as a normative standard (e.g. 60%), test developers need to shift

that perspective towards cut scores that are defensible as a representation of the performance standard if we are to allow a strong claim for CEFR linkage of our tests.

References

AERA (American Educational Research Association), APA (American Psychological Association) & NCME (National Council on Measurement in Education). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Alderson, J. C. (1993). Judgements in language testing. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 46-57). Alexandria, VA: TESOL.

— (2012). Principles and Practice in Language Testing: Compliance or Conflict? Presentation at TEA SIG Conference: Innsbruck. Retrieved from http://tea.iatefl.org/inns.html

Alderson, J.C., Figueras, N, Kuijper, H., Nold, G, Takala, S & Tardieu, C. (2006). Analysing Tests of Reading and Listening in Relation to the Common European Framework of Reference: The Experience of The Dutch CEFR Construct Project, *Language Assessment Quarterly*, *3*(1), 3-30.

ALTE/Council of Europe (2011) Manual for Language Test Development and Examining. For use with the CEFR. Retrieved from: http://www.coe.int/t/dg4/linguistic/ManualtLangageTest-Alte2011_EN.pdf

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly 2*(1), 1–34.

Bachman, L.F., & Palmer, A. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Brunfaut, T., & Harding, L. (2014). *Linking the GEPT listening test to the Common European Framework of Reference*. Taiwan: Language Training and Testing Centre.

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple... Language Testing, 29(1), 19–27.

Cizek, G. J. (2012). The forms and functions of evaluations in the standard setting process. In G. J. Cizek (Ed.), *Setting performance standards:*

Foundations, methods, and innovations (2nd ed., pp. 165 - 178). New York: Routledge.

— (2016). Validating test score meaning and defending test score use: Different aims, different methods. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 212-225.

Cohen, A. S., Kane, M. T., & Crooks, T. J. (1999). A generalized examinee-centered method for setting standards on achievement tests. *Applied Measurement in Education*, *12*(4), 343-366.

Conferencia De Rectores De Las Universidades Españolas (CRUE). (2011, 8 de septiembre). Propuestas sobre la acreditación de idiomas. Retrieved from:

http://www.acreditacion.crue.org/

Council of Europe. (2008). *Recommendation CM/Rec (2008)7 of the Committee of Ministers to member states on the use of the Council of Europe's Common European Framework of Reference for Languages (CEFR) and the promotion of plurilingualism*. Strasbourg, France: Council of Europe. Retrieved from http://www.coe.int/t/dg4/linguistic/Conventions_EN.asp

— (2009). Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment. Strasbourg, France: Council of Europe. Retrieved from: http://www.coe.int/T/DG4/Linguistic/Manuel1_EN.asp

De Jong, J. & Zheng, Y. (2016) Linking to the CEFR: validation using a priori and a posteriori evidence. In, Banerjee, J. and Tsagari, D. (eds.) *Contemporary Second Language Assessment.* London, GB, Bloomsbury Academic pp. 83-100. (Contemporary Applied Linguistics, 4).

Deygers, B., Zeidler, B., Vilcu, D., & Carlsen, C. H. (2017). One Framework to Unite Them All? Use of the CEFR in European University Entrance Policies. *Language Assessment Quarterly*, 1-13.

Downey, N. & Kollias, C. (2010). Mapping the Advanced Level Certificate in English (ALCE™) examination onto the CEFR. Aligning Tests with the CEFR, Reflections on using the Council of Europe's draft Manual, Martyniuk, W. (ed). Cambridge University Press. Cambridge. 119-129.

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Section H). Strasbourg, France: Council of Europe/Language Policy Division.

— (2012). Examinee-centered standard setting for large-scale assessments: The prototype group method. *Psychological Test and Assessment Modeling*, 54, 257–283.

Ferrara, S., & Lewis, D. (2012). The Item-Descriptor (ID) Matching method. In G. J. Cizek (Ed.), Setting performance standards: Foundations, methods, and innovations (2nd ed., pp. 255-282). New York: Routledge.

Figueras, N. (2012). The impact of the CEFR. *ELT journal*, *66*(4), 477-485.

Figueras, N., Kaftandjieva, F., & Takala, S. (2013). Relating a Reading Comprehension Test to the CEFR Levels: A Case of Standard Setting in Practice with Focus on Judges and Items. *Canadian Modern Language Review/La Revue Canadienne Des Langues Vivantes, 69*(4), 359-385.

Fulcher, G. (2004). "Deluded by artifices? The Common European Framework and harmonization." *Language Assessment Quarterly*, *1*(4), 253 - 266.

Fulcher, G., & Owen, N. (2016). Dealing with the demands of language testing and assessment. *The Routledge Handbook of English Language Teaching*, 109-120.

Green, A. (2017). Linking Tests of English for Academic Purposes to the CEFR: The Score User's Perspective, *Language Assessment Quarterly*, 1-16.

Halbach, A., Lazaro Lafuente, A., & Perez Guerra, J. (2013). La lengua inglesa en la nueva universidad española del EEES: The role of the English language in post-Bologna Spanish universities. *Revista de Educación*, 362, 105-132.

Hambleton, R., & Jirka, S. (2006) Anchor-based methods for judgmentally estimating item statistics. In S. Downing and T. Haladyna (Eds.), *Handbook of test development* (pp. 399-420). Mahwah, Nj: Erlbaum.

Hambleton, R., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Westport, CT: American Council on Education/ Praeger Publishers.

Harsch, C. & Hartig, J. (2015). What Are We Aligning Tests to When We Report Test Alignment to the CEFR?, *Language Assessment Quarterly*, 12:4, 333-362.

Kaftandijeva, F. (2010). *Methods for setting cut scores in criterion-referenced achievement tests. A comparative analysis of six recent methods with an application to tests of reading in EFL.* Cito, Arnhem: EALTA. Retrieved from: www.ealta.eu.org/documentsresources/FK_second_doctorate.pdf

Kane, M. (2012). Validating score interpretations and uses. *Language Testing*, *29* (1), 3 –17.

— (2017). Using Empirical Results to Validate Performance Standards BT - Standard Setting in Education: The Nordic Countries in an International Perspective. In S. Blömeke & J.-E. Gustafsson (Eds.), (pp. 11–29). Cham: Springer International Publishing.

Kanistra, V. & Harsch, C. (2017). *Using the Item Descriptor Matching method to enhance validity when aligning test to the CEFR.* 4th Meeting of the EALTA CEFR Special Interest Group. Retrieved from: http://www.ealta.eu.org/events/SIG_CEFR_london2017/presentations/Presentations/1400-1530/CEFR%20SIG%20Voula%20and%20Claudia.pdf

Kantarciouglu, E, Thomos, C, O'Dwyer, J & O'Sullivan, B. (2010) 'Benchmarking a high-stakes

proficiency exam: the COPE linking project', in W. Martyniuk (ed) *Aligning Tests with the CEFR: reflections on using the Council of Europe's draft Manual.* Cambridge University Press. Cambridge.

Kolen, M. J., & Brennan, R. L. (2014). Test equating, scaling, and linking: methods and practices (3rd ed.). New York: Springer-Verlag.

Linacre, J. M. (2017). Facets computer program for many-facet Rasch measurement, version 3.80.0. Beaverton, Oregon: Winsteps.com

Linn, R. L. (2003). Performance Standards: Utility for Different Uses of Assessments. *Education policy analysis archives*, *11*, 31.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Mitzel, H.C., Lewis, D.M., Patz, R.J., & Green, D.R. (2001). The bookmark procedure: Psychological perspectives. In G.J. Cizek (Ed), Setting performance standards: Concepts, methods and perspectives (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum Assoc.

North, B., & Jones, N. (2009). Relating language examinations to the Common European Framework of Reference for Languages. Further material on maintaining standards across languages, contexts and administrations by exploiting teacher judgment and IRT scaling. *Strasbourg: Language Policy Division.*

Papageorgiou, S. (2010). Investigating the decision-making process of standard setting participants. *Language Testing*, 27(2), 261–282.

Papageorgiou, S., & Tannenbaum, R. J. (2016). Situating Standard Setting within Argument-Based Validity. *Language Assessment Quarterly*, *13*(2), 109–123.

Peterson, C. H., Schulz, E. M. and Engelhard Jr., G. (2011), Reliability and Validity of Bookmark-Based Methods for Standard Setting: Comparisons to Angoff-Based Methods in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice*, 30: 3–14.

Reckase, M. D. (2010). NCME 2009 presidential address: What I think I know. *Educational Measurement: Issues and Practice, 29*(3), 3–7.

Tannenbaum, R, J. & Cho, Y. (2014). Critical Factors to Consider in Evaluating Standard-Setting Studies to Map Language Test Scores to Frameworks of Language Proficiency, *Language Assessment Quarterly, 11*(3), 233-249.

Tannenbaum, R, J. & Wiley E. C. (2008). Linking English-language test scores onto the Common European Framework of Reference: An application of standard-setting methodology. *ETS Research Report Series*, *2008*(1).

Weir, C. J. (2005). Limitations of the council of Europe's framework of reference (CEFR) in developing comparable examinations and tests. *Language Testing*, *22*(3), 281–300.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA.

Contact address: Caroline Shackleton. Centro de Lenguas Modernas de la Universidad de Granada. Placeta del Hospicio Viejo s/n 18009, Granada, Spain. E-mail:csarah@ugr.es

Received: 11-05-2017 Accepted: 02-02-2018