Revista Nebrija
de Lingüística Aplicada a la Enseñanza de las Lenguas

Validación del contenido de una prueba oral de clasificación de E/LE

Content validity of a Spanish oral placement Test

Begoña Martín Alonso
Universidad Nebrija.
bmartinalonso36@gmail.com

Martín Alonso, B. (2017). Validación del contenido de una prueba oral de clasificación de E/LE. *Revista Nebrija de Lingüística Aplicada* (2017) 22/.

RESUMEN

El presente estudio profundiza en la investigación sobre la evaluación de la lengua oral, en lo que respecta a la manera de incluir la validez de contenido como una evidencia más de la validez de un examen. La complejidad de medirla viene determinada, en primer lugar, por la dificultad para definir el dominio de la prueba y seleccionar tareas y muestras lingüísticas que representen adecuadamente ese dominio. En segundo lugar, por la escasez de información sobre cómo desarrollar procedimientos para estimar este tipo de evidencia de validez (Newman, Lim & Pineda 2013). Este trabajo revisa el concepto de validez de contenido y detalla el procedimiento de validación del contenido de una prueba de clasificación de español oral mediante juicio de expertos aplicando la técnica del grupo focal. Los resultados revelan la pertinencia de la metodología propuesta para obtener valoraciones de expertos unánimes y validar la representatividad y relevancia del contenido de una prueba.

Palabras clave: evaluación de lengua oral, validez de contenido, grupo focal, juicio de expertos

ABSTRACT

The present study explores on speaking language testing. Precisely it focuses on the inclusion of content validity as further evidence of the validity of an exam. The complexity of measuring content validity remains, in the first place, in defining the test domain and in selecting linguistic tasks and samples that adequately represent this domain. Secondly, this difficulty is due to the lack of information on how to develop procedures for estimating this type of validity evidence (Newman, Lim & Pineda 2013). This work reviews the concept of content validity and details the procedure for validating the content of a speaking placement test in Spanish as a Foreing Language through expert judgment applying the focus group technique. Results confirm the relevance of the proposed methodology to achieve unanimous expert judgments and to validate the representativeness and relevance of test content.

Keywords: testing speaking, content validity, placement test, focus group, expert judgements

?
Fecha de recepción: 10/11/2016
Fecha de aceptación: 15/01/2017

1. INTRODUCCIÓN

La principal preocupación cuando se desarrollan y administran exámenes de lengua reside en demostrar no solamente que los resultados de los mismos son fiables, sino también que las interpretaciones y usos que realizamos de estos son válidos (Bachman 1990).

De este modo, la validez se manifiesta como una de las cualidades necesarias para garantizar la utilidad de una prueba (Bachman & Palmer 1996).

Tradicionalmente concebida como una cualidad del test –es decir, si este mide lo que realmente pretende medir–, en esta investigación se define como una propiedad de las interpretaciones y usos que realizamos con base a los resultados del mismo (Bachman & Palmer 1996; Fulcher 2003; Messick 1986; Sireci 1998).

Por consiguiente, para validar una determinada prueba se han de recoger diferentes tipos de evidencias que respalden las conclusiones que extraemos[1] y las decisiones que tomamos[2] a partir del desempeño lingüístico de los estudiantes.

Entre los diferentes tipos de información que se pueden recolectar, se encuentran las evidencias relacionadas con el contenido (validez de contenido), con el criterio (validez de criterio) y con el constructo (validez de constructo).

En particular, la presente investigación se centra en recabar evidencias para corroborar la relevancia y representatividad del contenido de una prueba (Bachman 2002; Sireci 1998; Sireci & Faulkner-Bond 2014). En otras palabras, se valida el contenido de un examen oral de clasificación y, para ello, mediante el juicio de profesores expertos se determina el grado en que las tareas que lo conforman representan y son relevantes para el dominio o constructo de lengua oral definido en las especificaciones.

Por lo tanto, se coincide con Bachman (1990) y Bachman, Kunnan, Vanniarajan & Lynch (1988) en la necesidad de demostrar que un instrumento de evaluación es relevante y cubre adecuadamente los contenidos o la habilidad lingüística que se pretende medir para poder validar las interpretaciones que realizamos después de la actuación lingüística de los estudiantes.

Se advierte que, pese a la reconocida importancia que presenta hoy en día el concepto de validez de contenido en el ámbito de evaluación de segundas lenguas, este no fue ampliamente aceptado sino hasta finales de los sesenta, debido a que a principios del siglo S.XX los esfuerzos investigativos estaban concentrados en medir la validez de manera estadística y en la validez de constructo (Sireci 1998).

De la misma manera, se trae a colación algunos problemas que envuelven al proceso de validación entre los que se encuentran la complejidad de definir el dominio y seleccionar tareas que representen adecuadamente este constructo (Bachman 2002), la dificultad de obtener juicio de expertos unánimes (Alderson 1990; Alderson, Clapham & Wall 1998) y la escasez de información sobre los procedimientos que se pueden desarrollar para estimar este tipo de evidencia de validez (Newman, Lim & Pineda 2013).

Probablemente estas preocupaciones expliquen la escasez de investigaciones recientes que han abordado las contribuciones de profesores expertos en el diseño y validación de pruebas de lenguas (Cumming, Grant, Muchany-Ernt & Powers 2004). Sin embargo, se encontraron algunos estudios empíricos en el campo de lenguas extranjeras que miden la validez de contenido de las tareas y de la escala de evaluación mediante juicio de expertos empleando diferentes técnicas y métodos como entrevistas, talleres, grupos focales o cuestionarios (Bachman *et al.* 1988; Bachman, Davidson & Milanovic 1991; Cumming *et al.* 2004; Deygers, Van Gorp, Luyten & Joos 2013; Deygers & Van Gorp 2015; North & Schneider 1998; Robles & Rojas 2015; Wall, Clapham & Alderson 1994).

La mayoría se concentran en I/LE y en las habilidades de expresión escrita y comprensión lectora, solo cinco de ellas consideran también la lengua oral (Cumming *et al.* 2004; Deygers *et al.* 2013; Deygers *et al.* 2015; North & Schneider 1998; Robles & Rojas 2015). Por el contrario, se hallaron varios estudios empíricos sobre la destreza oral centrados en diseñar escalas de evaluación mediante análisis de discurso (Brown 2002; Frost, Elder & Wigglesworth 2011; Jin & Mak 2012; O' Sullivan, Weir & Saville 2002) o en los factores que afectan a la variabilidad de los resultados (Bachman, Lynch & Mason 1995; Del Moral 2014; Elder, Barkhuizen, Knoch, & von Randow 2007; Sawaki 2007; Wigglesworth 1993).

En lo que respecta al campo de E/LE, se encontraron algunas investigaciones que examinan la fiabilidad y la validez de las escalas de evaluación de la lengua oral (Del Moral 2014; Doquin & Martin Leralta 2014; Robles & Rojas 2015).

Fruto de lo anterior, esta investigación surge con el fin de cubrir el vacío de en el ámbito de evaluación oral en E/LE y profundizar en los procedimientos de validación al proponer el grupo focal como una técnica apropiada para validar el contenido de las tareas de una prueba de clasificación de español oral.

A continuación se define el concepto de validez de contenido y se precisa el procedimiento de validación del contenido de una prueba de clasificación de español oral en seis niveles (A1.1, A1.2, A2.1, A2.2, B1.1 y B1.2) a través de juicio de expertos utilizando la técnica del grupo focal. Por último, se discuten los alcances y limitaciones de la metodología aplicada.

2. MARCO CONCEPTUAL: LA VALIDEZ DE CONTENIDO

La validez de contenido –junto con la validez de constructo y la validez concurrente– se trata de uno de los tipos de evidencias que se puede recolectar para analizar la validez de una prueba y reside en examinar si los ítems que la conforman son una muestra representativa de todos los ítems incluidos en el dominio de interés[3] (Kerlinger 1986, cit. Wilson, Pan & Schumsky 2012: 19). Es decir, consiste en demostrar que el contenido de un examen (ítems/tareas, preguntas, materiales, etc.) es representativo y relevante para el constructo o habilidad lingüística que se quiere medir.

De esta manera, los tres aspectos que se han de tener en cuenta para valorar la validez de contenido son la definición, la representatividad y la relevancia del dominio (Sirecci 1998; Sireci & Faukner-Bond 2014).

La primera reside en definir operativamente el constructo; en otras palabras, en describir las habilidades lingüísticas que evalúa el test. Para ello, se puede tomar de referencia un modelo teórico (como el de habilidad lingüística comunicativa de Bachman 1990; véase apartado 3.3.1) o los contenidos de un programa, en el caso de los exámenes de aprovechamiento (ALTE 2005; Bachman *et al.* 1988). Esta definición se puede recoger en las especificaciones de la prueba (Sireci & Faulkner-Bond 2014).

La representatividad del dominio atañe al grado en que el instrumento de evaluación representa y mide adecuadamente el constructo definido. Para tal fin, normalmente se solicita a profesores expertos que analicen el contenido de las tareas con base a modelos teóricos de habilidad lingüística (Alderson 1990; Bachman *et al.* 1988 y Bachman *et al.* 1991) o contenidos de un determinado programa (Cumming *et al.* 2004; Wall *et al.* 1994).

La relevancia del dominio hace referencia a la medida en que cada ítem de una prueba es pertinente para el constructo detallado en las especificaciones. Esta se suele explorar solicitando a un grupo de especialistas que califique la medida en que cada ítem es relevante para determinados aspectos de las especificaciones de la prueba (Sireci & Faulkner-Bond 2014) o para los contenidos de un programa (Cumming *et al.* 2014).

No obstante, cabe destacar la complejidad que subyace al proceso de definir de manera clara y concisa el constructo que se quiere medir y de seleccionar tareas y muestras de lengua que representen adecuadamente este dominio (Bachman 2002). Del mismo modo, Alderson (1990) y Alderson *et al.* (1998) aluden a la dificultad de obtener valoraciones de expertos unánimes.

Asimismo, se coincide con Newman *et al.* (2013) en que no existe suficiente información acerca de los procedimientos para examinar la validez de contenido. Los estudios empíricos consultados suelen emplear tanto métodos cuantitativos (Li & Sireci 2013; Wilson, Pan & Schumsky 2012) como cualitativos o mixtos (Alderson 1990; Bachman *et al.* 1988 y Bachman *et al.* 1991; Cumming *et al.* 2004; Deygers *et al.* 2013; Newman *et al.* 2013; North & Schneider 1998; Robles & Rojas 2015). Entre los métodos cualitativos más utilizados se encuentra el juicio de expertos y, entre las técnicas para recoger la opinión de expertos, se destacan las entrevistas, los talleres, los grupos focales y los cuestionarios.

Finalmente, se aclara que la validez es un concepto unificado que consiste en la recolección de suficiente información que respalde las interpretaciones y decisiones que tomamos con base a los resultados de un prueba (Bachman 1990; Bachman & Palmer 1996). Por consiguiente, para determinar la validez de una prueba se han de recoger diferentes tipos de evidencias. La validez de contenido se presenta como uno de los más importantes en términos de que brinda la información necesaria para relacionar más tarde el desempeño lingüístico de los estudiantes con el constructo de la prueba (ALTE 2005; Bachman *et al.* 1988). De la misma forma, se destaca su relación con la fiabilidad en el sentido en que ambas aseguran que la prueba refleja adecuadamente los objetivos y contenidos lingüísticos definidos en el dominio lingüístico de la prueba (ALTE 2005).

Por todo lo anterior, este estudio representa un punto de partida para validar una prueba de clasificación y describe el procedimiento seguido para examinar la representatividad y relevancia de una prueba oral que servirá de base para después comparar, en un futuro estudio empírico, si la habilidad lingüística que queremos medir en la prueba se corresponde con las actuaciones lingüísticas de los estudiantes.

3. METODOLOGÍA

3.1. OBJETIVOS

El objetivo del presente estudio empírico consiste en validar la representatividad y relevancia del contenido de una prueba de clasificación de español oral mediante el juicio de expertos (Bachman *et al.* 1991; Cumming *et al.* 2004) aplicando la técnica del grupo focal.

3.2. CONTEXTO DE LA INVESTIGACIÓN

El estudio se lleva a cabo en el Centro Latinoamericano (CLAM) de la Pontificia Universidad Javeriana de Bogotá (PUJ) y se centra en la prueba oral para clasificar a los candidatos a los cursos de E/LE en los niveles A1.1., A1.2, A2.1, A2.2, B1.1 y B1.2 según el *Marco Común Europeo de Referencia para las Lenguas* (MCER) del Consejo de Europa (2002). La prueba consta de cuatro tareas de interacción oral con diferentes tipologías discursivas (narración, descripción, comparación y argumentación).

3.3. PROCEDIMIENTO DE VALIDACIÓN DEL CONTENIDO DE LAS TAREAS ORALES

Se llevaron a cabo dos procedimientos; uno teórico que consistió en definir el dominio de la prueba oral a través de la redacción de las especificaciones y otro empírico en el que se llevó a cabo un grupo focal con expertos de la PUJ para determinar el grado en el que las tareas se relacionan y son relevantes para el dominio de la prueba oral de clasificación detallado en las especificaciones.

3.3.1 DEFINICIÓN DEL DOMINIO

Para poder estudiar la representatividad y relevancia del contenido es imprescindible definir previamente el dominio de la prueba (véase apartado 2).

Para tal fin, como recomiendan Sireci & Faulkner-Bond (2014), se redactaron las especificaciones de la prueba, que se presentan como una guía indispensable que deben seguir tanto los redactores (Alderson *et al.* 1998; Bachman & Palmer 1996; McNamara 2000) cuando quieran diseñar pruebas similares, como los evaluadores para otorgar calificaciones justas. Del mismo modo, se destaca la estrecha relación entre estas y la validación de los exámenes (Alderson *et al.* 1998; Fulcher 2003; Luoma 2004).

En concreto, se siguió la propuesta de Luoma (2004) para dividir las especificaciones en tres partes: las especificaciones del constructo, las especificaciones de las tareas y las especificaciones de la evaluación.

Las primeras consisten en definir el constructo y establecer el propósito de la misma. Adicionalmente, se deben describir el contexto en que se administrará y los candidatos que se presentarán a este examen. Para definir el constructo –entendido como las habilidades que mide una determinada prueba–, Luoma (2004) aconseja que se revisen modelos teóricos como el modelo de habilidad lingüística comunicativa (HLC) de Bachman (1990) o, específicamente para la lengua oral, el de Bygate (1987).

Las especificaciones de las tareas reúnen los tipos de tareas que conforman la prueba, así como las destrezas que miden y los materiales e instrucciones que se requieren. Para definir las propiedades de las tareas de evaluación los estudios consultados (Bachman *et al.* 1988; Bachman *et al.* 1991) consideran los parámetros de Bachman (1990) y Bachman & Palmer (1996): características del contexto (participantes, tiempo), características de la rúbrica (instrucciones, duración, método), características del input (formato, lengua) y características de la respuesta (formato, lengua, tema). Del mismo modo, se realza la estrecha relación entre la definición del constructo y las tareas, en el sentido de que estas deben medir el constructo de la prueba definido en las especificaciones (Bachman & Palmer 2013).

Por último, las especificaciones de la evaluación contienen información sobre los criterios de calificación y su modo de aplicación. De nuevo, se subraya la estrecha relación entre los criterios y el constructo, en el sentido de que estos forman parte de la definición del mismo (Brown 2000; Fulcher 2003; Hughes 2003; Luoma 2004). Como resultado, las especificaciones de las tareas, las escalas y el constructo se diseñan de manera paralela (Luoma 2004).

En el presente trabajo las especificaciones de la prueba oral de clasificación siguen estas pautas y detallan, en primer lugar, la definición del constructo de lengua oral entendido como el conjunto de habilidades lingüísticas (dar información personal, describir y comparar lugares y servicios, narrar y argumentar) que el estudiante lleva a cabo para desempeñar con éxito las tareas de la prueba. Para ello, se tomaron de referencia las rutinas de información[4] de Bygate (1987), las funciones lingüísticas contempladas en el *Plan Curricular del Instituto Cervantes* (Instituto Cervantes 2006) y los programas del CLAM. Cabe recalcar la importancia de estos últimos debido a que una prueba de clasificación debe estar alineada con los contenidos de los programas de un determinado centro (Alderson *et al.* 1998; McNamara 2000; Nakamura 2007). De igual manera, se describen el propósito de la prueba y los candidatos de la prueba.

En segundo lugar, las especificaciones de las tareas albergan las propiedades de las cuatro tareas diseñadas que denominaremos Tarea 1, Tarea 2A, Tarea 2B y Tarea 3. Respecto a las características propuestas por Bachman & Palmer (1996), estamos de acuerdo con Fulcher (2003) en que no está claro qué elementos de la lista son los más importantes para describir las tareas de la prueba oral. Por lo tanto, proponemos los siguientes aspectos el propósito y el tipo de tarea, la autenticidad, la duración, las instrucciones, el material y el input lingüístico y cultural, y se definen a partir del análisis de los programas del CLAM, de las tareas de los exámenes DELE del Instituto Cervantes y LETRA de la Universidad Antonio de Nebrija (Baralo 2009; Martin Leralta 2011) y de las recomendaciones de Fulcher (2003) y Luoma (2004).

Por último, las especificaciones de la evaluación incluyen la escala analítica (criterios y descriptores) diseñada para calificar a los alumnos de E/LE, así como los procedimientos de calificación. Se tomaron como referencia las escalas del MCER (Consejo de Europa 2002), los exámenes DELE y LETRA (Estaire & Baralo 2011) y la definición del constructo de lengua oral.

Las dos primeras partes de las especificaciones resultaron las más relevantes para el estudio empírico realizado debido a que las habilidades lingüísticas (constructo de lengua oral) y algunos aspectos de las tareas descritos en las especificaciones se sometieron a juicio de expertos a través del grupo focal (véase Anexo 1).

3.3.2. RELEVANCIA Y REPRESENTATIVIDAD

Para examinar la relevancia y representatividad del contenido de las tareas de la prueba se empleó el juicio de expertos por tratarse de uno de los métodos cualitativos más utilizados para medir la validez de contenido de una prueba (Bachman *et al.* 1988; Bachman *et al.* 1991; Cumming *et al.* 2004; Deygers *et al.* 2013; Deygers & Van Gorp 2015; North & Schneider 1998; Robles & Rojas 2015; Wall *et al.* 1994).

En concreto, se aplicó la técnica del grupo focal y se solicitó a los expertos tanto el establecimiento de las habilidades lingüísticas que identificaran en el ejercicio de las tareas, como la descripción de algunas propiedades de las mismas: nivel de dominio lingüístico, uso real de la lengua y contenidos lingüísticos y culturales (véase Anexo 1). Una vez finalizado el análisis, se cotejaron las valoraciones de los expertos sobre las tareas con la descripción de las mismas recogida en las especificaciones (véase Anexo 1), para valorar si su contenido se correspondía y era relevante para cada nivel de dominio de la prueba oral de clasificación.

Se aclara que, para seleccionar el contenido de la prueba que se va a validar (habilidades y características) se tuvieron en cuenta los estudios de Bachman *et al.* (1988) y Bachman *et al.* (1991) y las propiedades que resultaban más relevantes en términos del propósito de la prueba.

3.3.2.1 TÉCNICA DEL GRUPO FOCAL

Para recolectar las valoraciones de los expertos respecto a los contenidos de la prueba oral de clasificación, al igual que Deygers *et al.* (2013) y Deygers & Van Gorp (2015), se utilizó la técnica del grupo focal, tal y como utilizaron para evaluar la validez de contenido de su estudio.

El grupo focal se llevó a cabo el 14 de julio de 2015 con seis profesores expertos del CLAM y se estructuró en cuatro fases: introducción, trabajo individual, discusión y conclusión.

En la fase introductoria, y de acuerdo con Dörnyei (2007), se recordaron las normas básicas de un grupo focal (respetar turnos de habla, no interrumpir, etc.) y se explicaron el propósito y la estructura del mismo.

Durante la fase de trabajo individual, siguiendo las recomendaciones de Dörnyei (2007), Elliot (2005) y Hernández Sampieri *et al.* (2010), se proporcionó una guía de preguntas que los docentes debían responder de manera individual. En total se formularon cuatro preguntas abiertas que pedían al especialista que identificara las habilidades lingüísticas y las características de las tareas de la prueba oral: uso real de la lengua, nivel de dominio lingüístico, contenidos lingüísticos y culturales. Como material de apoyo se les proporcionaron los programas del CLAM y los índices de *Aula Internacional*[5] (Corpas, García & Garmendia 2013).

En la fase de discusión, los expertos pusieron en común las respuestas que habían dado a la guía y discutieron sobre las habilidades y las características de las tareas de la prueba de clasificación.

Además, tal y como sugieren Dörnyei (2007), Elliot (2005) y Hernández Sampieri *et al.* (2010), se eligió un moderador (el investigador) con el fin de propiciar un ambiente distendido, orientar la discusión cuando no se hubieran tratado los objetivos y prevenir que ningún participante en particular liderase la conversación.

Por último, en la fase de conclusión, teniendo en cuenta las palabras de Dörnyei (2007), se hizo una recapitulación de los puntos principales, se resolvieron dudas y se identificaron acciones futuras.

3.3.2.2 PARTICIPANTES DEL GRUPO FOCAL

Para la validación del contenido de las tareas, siguiendo las pautas de Dörnyei (2007) y Elliot (2005) –quienes plantean que los grupos focales deben estar conformados por entre seis y doce participantes–, se eligieron seis expertos del Departamento de Lenguas de la Facultad de Comunicación y Lenguaje de la PUJ.

Al respecto de los especialistas, se coincide con Brindley (1994) y Martínez Batzán (2011) en que los profesores deberían formar parte de dicho equipo porque son los que van a realizar las calificaciones y puesto que cuentan con experiencia en la elaboración de tareas de aula y evaluación. En concreto, se está de acuerdo con Newman *et al.* (2013) en que estos suelen ser futuros usuarios de la prueba. Por consiguiente, se seleccionaron docentes con experiencia en la enseñanza y evaluación de E /LE, concretamente en el CLAM, contexto donde se administrará la prueba (véase apartado 3.2).

Del mismo modo, debido a la importancia de contar con un grupo homogéneo de participantes para contribuir a un ambiente de discusión relajado en el que pudieran intercambiar opiniones sin cohibiciones (Dörnyei 2007; Elliot 2005; Hernández Sampieri *et al.* 2005), se tuvieron en cuenta los siguientes factores: años de experiencia en la enseñanza y evaluación de ELE, concretamente en el CLAM; formación en evaluación, por ejemplo si eran evaluadores acreditados del Instituto Cervantes, y nivel de jerarquía que ocupan en el centro. Esta información se recolectó mediante un cuestionario que permitió una adecuada selección de los participantes.

4. RESULTADOS DE INVESTIGACIÓN

Para determinar la relación y relevancia de la prueba oral respecto al dominio lingüístico definido, se cotejaron las valoraciones de los expertos sobre las habilidades y las propiedades de las tareas (uso de la lengua oral, nivel de dominio lingüístico, habilidades lingüísticas, y contenidos lingüísticos y culturales) con la información recogida en las especificaciones de la prueba.

Las categorías aplicadas para el análisis del juicio de expertos son las siguientes:

- Se corresponde (C): la respuesta del expertos es igual que la información albergada en las especificaciones.
- Se corresponde parcialmente (CP): la respuesta del experto coincide en algunos aspectos pero en otros discrepa con la información incluida en las especificaciones.

A continuación, se presentan los resultados teniendo en cuenta las propiedades de las tareas analizadas y las tareas que componen la prueba oral de clasificación: Tarea 1, Tarea 2A, Tarea 2B y Tarea 3 (véase Tabla 1).

Propiedades de las tareas	Tarea 1	Tarea 2 A	Tarea 2 B	Tarea 3

Uso real de la lengua	C	C	C	C
Nivel de dominio lingüístico	C	CP	CP	C
Habilidades lingüísticas	C	CP	CP	C
Contenidos lingüísticos y culturales	CP	CP	CP	CP

Tabla 1: Resultados de la discusión del grupo focal

4.1 USO REAL DE LA LENGUA

En todas las tareas, los expertos identificaron los usos reales pero no aportaron ejemplos de contextos reales donde los estudiantes los pudiesen llevar a cabo. No obstante, el moderador del grupo focal orientó la conversación y preguntó por dichos contextos teniendo en cuenta el perfil de los candidatos de la prueba y los profesores aportaron varios ejemplos.

Producto de lo anterior se concluye que las tareas diseñadas para la prueba oral de clasificación obedecen a situaciones comunicativas reales en las que los aprendientes del CLAM tendrán que desenvolverse. Es decir, las respuestas del grupo focal se corresponden con lo expuesto en las especificaciones (véase Tabla 1 y Anexo 1).

4.2 NIVEL DE DOMINIO LINGÜÍSTICO

Los docentes identificaron el nivel de las Tarea 1, A1.1 , y la Tarea 3, B1.1 y B1.2 expuesto en las especificaciones (véase Anexo 1). Por el contrario, en las Tareas 2A y 2B surgieron discrepancias entre ellos y, por ende, las respuestas se correspondieron parcialmente con lo contenido en las especificaciones (véase Tabla 1).

De esta manera, en la Tarea 2A, los profesores identificaron solamente un nivel, el A1.1 o el A2.1. Por otro lado, el experto E #2 argumentó, con base a los programas del CLAM y a las escalas del MCER (Consejo de Europa 2002), que esta tarea medía también el nivel A2.2, nivel que no estaba incluido en las especificaciones.

Del mismo modo, en la Tarea 2B, todos los expertos coincidieron en que la tarea medía el nivel A2.2. De la misma forma, añadieron que también medía el nivel A2.1 y B1.1 justificando sus aportaciones con los contenidos de los programas y las escalas del MCER(Consejo de Europa 2002).

Por lo anterior, se revisan los niveles de las especificaciones de todas las tareas, en particular las Tareas 2A y 2B para incluir los niveles más relevantes teniendo en cuenta los resultados obtenidos en el grupo focal y las habilidades y los contenidos expuestos en las especificaciones.

4.3 HABILIDADES LINGÜÍSTICAS

Los docentes lograron identificar todas las habilidades de las Tareas 1 y Tarea 3 (véase Anexo 1) y el 75% de las habilidades de las Tareas 2A y 2B (véase Tabla 2). Como resultado, las respuestas del grupo focal se corresponden con lo establecido en las especificaciones de las Tarea 1 y Tarea 3 y se corresponden parcialmente con las Tareas 2A y 2B (véase Tabla 1).

Habilidades lingüísticas (Especificaciones)	Habilidades lingüísticas (Expertos)
TAREA 2A	
Expresar gustos	Sí
Describir barrios, pueblos y lugares	Sí
Comparar barrios, pueblos y lugares	Sí
Expresar intención	No
TAREA 2B	
Describir físicamente personas	Sí

Expresar hábitos	Sí
Describir aspectos de su pasado	Sí
Valorar aspectos de su pasado	No

Tabla 2: Resultados de las habilidades lingüísticas

De igual modo, se subraya que estos añadieron otras habilidades que no estaban presentes en las especificaciones pero sí en los programas del CLAM y los índices del manual empleado en el Centro, *Aula Internacional* (Corpas, García & Garmendia 2013).

En definitiva, se cree conveniente modificar las especificaciones de las tareas para incluir las habilidades lingüísticas más relevantes teniendo en cuenta las intervenciones de los profesores. Del mismo modo, se considera necesario revisar la pertinencia de las funciones lingüísticas incluidas en las especificaciones y que no se mencionaron en la discusión del grupo focal.

4.4 CONTENIDOS LINGÜÍSTICOS Y CULTURALES

Los expertos en ninguna tarea identificaron todos los contenidos lingüísticos y culturales: Tarea 1 (72%), Tarea 2A (71,4%), Tarea 2B (57%) y Tarea 3 (50%)(véase Anexo 2). Por tanto, las respuestas se correspondieron parcialmente con lo establecido en las especificaciones (véase Tabla 1).

Como en el caso de las habilidades, los docentes trajeron a colación otros contenidos lingüísticos que no se habían abarcado en las especificaciones pero sí en los programas del CLAM.

Por consiguiente, se considera necesario incluir en las especificaciones de las tareas los contenidos más relevantes teniendo en cuenta las valoraciones de los profesores. De igual forma, se revisa la pertinencia de las competencias que no se mencionaron en el grupo focal.

Por último, se subraya que todos los participantes criticaron la carga cultural de las imágenes incluidas en las Tareas 2A y 2B y un experto (E#2) sugirió que se ampliara el tema de la Tarea 3.

A causa de lo anterior, se modifican las imágenes de las situaciones y acciones que el aprendiente tiene que describir y comparar en las Tareas 2A y 2B y se amplía el contexto de las Tareas 2 y Tarea 3 para no limitarlo a la ciudad de Bogotá.

5. CONCLUSIONES

En líneas generales las valoraciones que realizaron los expertos en el grupo focal sobre el contenido de la prueba oral constatan que las tareas reflejan el constructo definido en las especificaciones y, por tanto, son representativas y relevantes para clasificar a los alumnos en los cursos de E/LE.

Sin embargo, atendiendo a las sugerencias de los participantes en el grupo focal se revisa la información detallada en las especificaciones en lo que respecta a los niveles de dominio lingüístico, habilidades y contenidos lingüísticos que miden las tareas, principalmente las tareas 2A y 2B. Del mismo modo, se modifican las imágenes contenidas en los materiales de las tareas para reflejar situaciones y hechos más generales que puedan extrapolarse a diferentes contextos culturales.

A pesar de que se dieron discrepancias entre los docentes al respecto de lo que medían cada tarea, en concreto a la hora de establecer el nivel de dominio al igual que acontece en el estudio de Alderson (1990), se logró alcanzar un cierto grado de acuerdo gracias a la técnica del grupo focal.

A raíz de lo anterior se desprende, primero, la importancia de contar con las opiniones de profesores expertos para mejorar las especificaciones y las tareas que conforman una prueba. Segundo, se manifiesta la pertinencia de la técnica del grupo focal para obtener valoraciones unánimes y, por ende, se coincide con Bachman *et al.* (1991) en la posibilidad de alcanzar niveles razonables de acuerdo entre los expertos.

En particular, un aspecto que ha brindado aportes a la aplicación de la técnica del grupo focal es la guía individual previa a la discusión de los participantes, en tanto que ha permitido que los participantes se forjaran una opinión en función de sus conocimientos y experiencias docentes y evaluadoras sin influencia del grupo, lo que ha favorecido una discusión posterior más rica en argumentos. No obstante, una de las limitaciones residió en las discrepancias que surgieron entre los mismos en cuanto a la concepción de uno de los parámetros que debían evaluar, el uso real de la lengua. Por consiguiente, los resultados ponen de manifiesto la necesidad de validar previamente la guía para asegurar una interpretación unívoca de los parámetros que la componen y, por ende, garantizar valoraciones válidas y fiables.

Por otro lado, se ha comprobado la adecuación de las especificaciones no solamente para definir el dominio que se quiere medir, sino también para los procesos de validación de exámenes (Alderson *et al.* 1998; Fulcher 2003; Luoma 2004); en particular, para comprobar la representatividad y relevancia de las tareas.

En este punto se resalta como aspecto positivo el hecho de que los expertos no hayan tenido acceso a las especificaciones de la prueba y debieran identificar ellos mismos los contenidos que se evalúan en las tareas. Es decir, a diferencia de lo que realizaron otros estudios como Alderson (1990) Bachman *et al.* (1988), Bachman *et al.* (1991), quienes solicitaron a los expertos que compararan las habilidades lingüísticas y características de las tareas con modelos teóricos de habilidad lingüística y características del método, en esta investigación se cotejan las valoraciones de los docentes sobre el constructo y algunos parámetros de las tareas con el dominio definido en las especificaciones para, de esta manera, no influir en sus respuestas.

En relación con el contenido de la prueba, las características de habilidades lingüísticas, uso real de la lengua, nivel de dominio lingüístico, y contenidos lingüísticos y culturales se consideran adecuadas para aportar evidencia sobre la representatividad y relevancia del contenido de la prueba oral. A partir de las limitaciones detectadas en este estudio y los resultados arrojados en Bachman *et al.* (1991), de cara a futuras aplicaciones de la metodología empleada se recomienda incluir otros componentes de la prueba oral también recogidos en las especificaciones, como las instrucciones, el material y la tipología de las tareas.

Finalmente, se percibe la pertinencia de indagar en el proceso de validación y estudiar otros tipos de evidencias debido a que la validez de una prueba solo se puede demostrar a través de la recolección e interpretación de todos los tipos relevantes de información (Bachman 1990). De esta manera, se concluye al igual que Bachman *et al.* (1988) que la información recogida en este estudio sirve de base para la formulación de hipótesis sobre las futuras actuaciones lingüísticas del candidato y para analizar la correspondencia entre estas hipótesis y el desempeño lingüístico del estudiante en la prueba.

En definitiva, el procedimiento que se ha presentado en el presente artículo pretende constituir una referencia para evaluadores, profesores, administradores o estudiosos interesados en valorar la relevancia y representatividad del contenido de una prueba de lengua oral. Asimismo, puede servir como base para analizar otros tipos de evidencias de validez o para seguir profundizando en la propia validez de contenido, ya que el procedimiento aquí expuesto constituye tan solo la primera parte de la validación del contenido de la prueba evaluadora. El paso siguiente para garantizar la validez lo constituirá la validación de la escala de calificación y de las tareas mediante las observaciones de un repertorio de actuaciones lingüísticas de candidatos a la prueba, lo que permitirá, a su vez, el análisis de la fiabilidad interevaluadora e intraevaluadora de los examinadores, cerrando así el estudio de la validez y fiabilidad de la prueba de lengua.

1. Un ejemplo de una conclusión o interpretación sería que el alumno posee un amplio abanico léxico o que no presenta suficiente fluidez.
2. Un ejemplo de una decisión que se toma con base a los resultado de una prueba, sería otorgar el certificado de un programa o clasificar al estudiante en un determinado curso.
3. “rests on demonstration that the test´s items are a representative sample of all ítems within the containt domain of interest” (Kerlinger 1986 en Wilson *et al.* 2012: 19).
4. Bygate (1987) señala que las rutinas de información consisten en estructuras de información recurrentes (contar historias, descripciones de personas o lugares, comparaciones, instrucciones, etc.) y las rutinas de interacción no dependen tanto del contenido de la información sino en las estructuras para organizar los turnos de habla en situaciones de interacción de la vida real (comprar un billete de tren, hablar por teléfono).
5. Los libros de *Aula Internacional* son los que se utilizan en el Centro Latinoamericano de la PUJ.

REFERENCIAS BIBLIOGRÁFICAS

Association of Language Testers in Europe (2005). Materials for the guidance of test item writers, 1-199. http://www.alte.org/attachments/files/item_writer_guidelines.pdf

Alderson, J. (1990). Testing Reading Comprehension Skills (Part One). Reading in a foreign language, 6 (2), 425-438. http://nflrc.hawaii.edu/rfl/PastIssues/rfl62anderson.pdf

Alderson, J., Clapham, C., Wall, D. (1998). Exámenes de idiomas: Elaboración y evaluación. England: Cambridge University Press.

Bachman, L. (1990). Fundamental considerations in language testing. England: Oxford University Press.

Bachman, L. (2002). Some reflections on task-based language performance assessment. Language Testing, 19: 453-479. http://journals.sagepub.com/doi/abs/10.1191/0265532202lt240oa

Bachman, L., Davidson, F., Milanovic, M. (1991). The use of test method characteristics in the content analysis and design of EFL proficiency tests. 13th Annual Language Testing Research Colloquium, 21-23 de marzo, 125-150.

Bachman, L., Kunnan, A., Vanniarajan, S., Lynch, B. (1988). Task and ability analysis as a basis for examining content and construct comparibility in two EFL proficiency and test batteries. Language Testing, 5: 128-159. Doi 10.1177/026553228800500203.

Bachman, L., Lynch, B., Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. Language Testing, 12: 238–57. http://files.eric.ed.gov/fulltext/ED368154.pdf

Bachman, L., Palmer, A. (1996). Language Testing in Practice. England: Oxford University Press.

Bachman, L., Palmer, A. (2013). Language Assessment in Practice. England: Oxford University Press.

Baralo, M. (2009). Implantacio?n de una certificacio?n de nivel A1 + 1 y disen?o curricular especi?fico de espan?ol para trabajadores inmigrantes. Actas del XX Congreso Internacional de ASELE, Comillas, 23-26 de septiembre de 2009.

Brindley, G. (1994). Task-Centred Assessment in Language Learning: The Promise and the Challenge. Language and Learning, papers presented at the Annual International Language in Education Conference (Hong Kong, 1993), 73-94. https://eric.ed.gov/?id=ED386045

Brown, A. (2000). An investigation of the Rating Process in the IELTS Oral Interview. IELTS Research Reports, 3: 49-84. https://www.educationaustralia.com/PDF/Vol3_Report3.pdf

Brown, A. (2002). Candidate discourse in the revised IELTS Speaking Test. IELTS Research Reports, 6: 1-18. https://www.yumpu.com/en/document/view/28597215/report-3-candidate- discourse-in-the-revised-ielts-speaking-test

Bygate, M. (1987). Speaking. England: Oxford University Press

Consejo de Europa (2002). Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación. Madrid: MEC y Anaya. Versión española en Centro Virtual Cervantes. http://cvc.cervantes.es/obref/marco/default.htm

Corpas, J., García, E., Garmendia, A. (2013). Aula Internacional 1, 2, 3, 4, Barcelona: Difusión.

Cumming, A., Grant, L., Mulcahy-Ernt, P., Powers, D. (2004). A teacher-verification study of speaking and writing prototype tasks for a new TOEFL. Language Testing, 21: 107-145. Doi 10.1191/0265532204lt278oa.

D´Este, C. (2012). New views of validity in language testing. EL.LE 1 (1), 61-76. http://virgo.unive.it/ecf-workflow/upload_pdf/1_1_12_DEste.pdf

Del Moral, F. (2014). Dificultades en el uso de los descriptores en el proceso de evaluación de la expresión oral de ELE.

Revista Nebrija de Lingüística Aplicada la Enseñanza de las Lenguas, 16. http://www.nebrija.com/revista-linguistica/dificultades-en-el-uso-de-los-descriptores-en-el-proceso-de-evaluacion-de-la-expresion-oral-de-ele

Deygers, B., Van Gorp K. (2015). Determining the scoring validity of a co-constructed CEFR-based rating scale. Language Testing, 32 (4), 521-541. http://ltj.sagepub.com/content/32/4/521.abstract?rss=1

Deygers, B., Van Gorp, K., Luyten, L., Joos, S. (2013). Rating scale design: A comparative study of two analytic rating scales in a task-based test. Exploring language frameworks, Proceedings from the ALTE, Kraków conference, julio 2011, 273–289. Cambridge: Cambridge University Press.

Doquin, A., Martin Leralta, S. (2014). Procedimientos para asegurar la validez y la fiabilidad de la evaluación de la interacción oral para la certificación lingüística de ELE. La enseñanza del Español como LE/L2 en el siglo XXI, 263-274, Asociación para la Enseñanza del Español como Lengua Extranjera. http://cvc.cervantes.es/ensenanza/biblioteca_ele/asele/pdf/24/24_261.pdf

Dörnyei, Z. (2007). Research methods in applied linguistics quantitative, qualitative, and mixed methodologies. England: Oxford University Press.

Elder, C., Barkhuizen, G., Knoch, U., von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. Language Testing 24 (1), 37–64. Doi 10.1177/0265532207071511.

Elliot, H. (2005). Guidelines for conducting a Focus group. American Journal For Reserchers, 1-10. https://assessment.trinity.duke.edu/documents/How_to_Conduct_a_Focus_Group.pdf

Estaire, S., Baralo, M. (2011). Variables socioculturales y comunicativas en el diseño curricular de una certificación de español para trabajadores inmigrantes. Lengua y migración, 3 (2), 5-41. http://dspace.uah.es/dspace/bitstream/handle/10017/19081/Variables_Baralo_LM_2011_3_2.pdf?sequence=1&isAllowed=y

Frost, K., Elder, C., Wigglesworth, G. (2011). Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test taker´s oral performances. Language Testing, 29 (3), 345-369. Doi 10.1177/0265532211424479.

Fulcher, G. (2003). Testing second language speaking. England: Pearson/Longman.

Hernández Sampieri, R., Fernández Collado, C., Baptista, M. (2010). Metodología de Investigación. México: Mc Graw-Hill.

Hughes, G. (2003). Testing for language teachers. England: University Press.

Instituto Cervantes (2006). Plan Curricular del instituto Cervantes. Niveles de referencia para el español. Madrid: Instituto Cervantes.http://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular/

Jin, T., Mak, B. (2012). Distinguishing features in scoring L2 Chinese speaking performance: How do they work? Language Testing, 30: 23-47. Doi 10.1177/0265532212442637.

Kane, M. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50: 74–83. http://onlinelibrary.wiley.com/doi/10.1111/jedm.12000/abstract

Li, X., Sireci, S. (2013). A New Method fo Analyzing Content Validity Data Using Multidimensional Scaling. Educational and Pscychological Measurement, 73 (3), 365-385. http://journals.sagepub.com/doi/pdf/10.1177/0013164412473825

Luoma, S. (2004). Assessing Speaking. England: Cambridge University Press.

Martin Leralta, S. (2011). Certificación lingüística de nivel inicial para inmigrantes en contexto laboral: ejemplo de una prueba del examen DILE. Lengua y Migración, 3 (1), 89-104. http://lym.linguas.net/Download.axd?type=ArticleItem&id=90

Martínez Batzan, A. (2011). La evaluación de lenguas: garantías y limitaciones, Barcelona: Octaedro.

Messick, S. (1986). The once and future issues of validity: assessing the meaning and consequences of measurement. New Jersey: Educational Testing Service Princenton. http://onlinelibrary.wiley.com/doi/10.1002/j.2330-8516.1986.tb00185.x/full

Nakamura, Y. (2007). A Rasch- based analysis of an in-house placement test. 6th Annual JALT Pan-SIG, 97-109.https://jalt.org/pansig/2007/HTML/Nakamura.htm

Newman, I., Lim, J., Pineda, F. (2013). Content validity using a mixed methods approach: Its application and development through the use of a table of specifications methodology. Journal of Mixed Methods Research, 7: 243-260. Doi 10.1177/1558689813476922.

North, B., Schneider, G. (1998). Scaling descriptors for language proficiency scales. Language Testing, 15 (2), 217–263. Doi 0265-5322(98)LT149OA.

O´Sullivan, B., Weir, C., Saville, N. (2002). Using observation checklists to validate speaking-test tasks. Language Testing, 19 (1), 33-56. http://lrc.cornell.edu/events/past/2008-2009/papers08/osull2.pdf

Robles, P., Rojas, M. (2015). La validación por juicio de expertos: dos investigaciones cualitativas en Lingüística aplicada. Revista Nebrija de Lingüística Aplicada, 18. https://www.nebrija.com/revista-linguistica/files/articulosPDF/articulo_55002aca89c37.pdf

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: reporting a score profile and a composite. Language Testing, 24: 355-390. http://ltj.sagepub.com/content/24/3/355.abstract

Sireci, S. (1998). The construct of contend validity. Social Indicators Research, 45: 83-117. https://www.researchgate.net/publication/227088853_The_Construct_of_Content_Validity

Sireci, S., Faulkner-Bond, M. (2014). Validity evidence based on test content. Psicothema, 26: 100-107. http://www.psicothema.com/pdf/4167.pdf

Wall D., Clapham C., Alderson J. (1994). Evaluating a placemente test. Language Testing, 11: 321-344. http://ltj.sagepub.com/content/11/3/321.full.pdf

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. Language Testing, 10 (3), 305-336. http://journals.sagepub.com/doi/abs/10.1177/026553229301000306

Wilson, F., Pan, W., Schumsky D. (2012). Recalculation of the critical values for Lawshe´s content validity ratio. Measurement and Evaluation in Counselling and Development, 45 (3), 197-210. Doi: 10.1177/0748175612440286.

ANEXO

ANEXO I: ESPECIFICACIONES DE LAS TAREAS

TAREA 1: CALENTAMIENTO	
Uso real	Presentarse/dar información personal el primer día de clase a sus compañeros/profesor/. Entrevistas de trabajo, primer día de trabajo. Situación real tanto en el ámbito académico como público/personal.
Nivel de dominio	A1.1
Habilidades	Dar información básica sobre sí mismo con respecto a la profesión/edad/nacionalidad/ domicilio. Expresar intenciones.
Contenidos	Gramática: Presente indicativo de verbos de presentación como: ser/estar/tener/querer/llamarse. Por/parar. Usos de por/para. Vocabulario: nacionalidades, profesiones, números. Cultura: saludos y expresiones de despedida.

TAREA 2: BOGOTÁ Y YO

Uso real	Tareas 2A y 2B: Conversación habitual cuando se conoce a alguien por primera vez en el ámbito laboral, académico o social. En contexto de aprendizaje de L2 en inmersión.
Nivel de dominio	Tarea 2A: A1.2 y A2.1
	Tarea 2B: A2.2
Habilidades	Tarea 2A: Expresa gustos, describe de manera sencilla y limitada pueblos, barrios y ciudades. Habla sobre intenciones y proyectos en relación a su estancia en Bogotá. Compara barrios y ciudades.
	Tarea 2B: Describe físicamente a las personas y expresa sus hábitos. Describe y valora experiencias y aspectos de su pasado.
Contenidos	Tarea 2A: Gramática: verbos gustar/hay vs estar. Artículos determinados e indeterminados. Comparativos. Ir +a + infinitivo. Vocabulario: lugares y servicios de una ciudad. Cultura: lugares emblemáticos de Bogotá y sus países de origen.
	Tarea 2B: Gramática: indefinido vs imperfecto, estar+ gerundio. Vocabulario: expresiones de temporalidad y valoración, adjetivos para la descripción de personas, tiempo libre. Cultura: situaciones cotidianas en Colombia y sus países de origen.

TAREA 3: CONTAMINACIÓN EN BOGOTÁ	
Uso real	Conversación habitual entre amigos o compañeros de trabajo. ¿Qué te parece la vida aquí? ¿Qué te parece lo del terremoto que pasó en Italia?
Nivel de dominio	B1.1 y B1.2
Habilidades	Valora situaciones y hechos. Expresa hipótesis y hace conjeturas.
Competencias	Gramática: futuro simple, conectores de probabilidad y valoración con indicativo vs subjuntivo. Condicional simple. Vocabulario: léxico para valorar situaciones: es injusto+ es normal. Léxico relacionado con el medio ambiente. Cultura: problemas sociales relacionados con la contaminación en Bogotá y en sus países de origen.

ANEXO II- RESULTADOS DE LOS CONTENIDOS LINGÜÍSTICOS Y CULTURALES

Contenidos Especificaciones	Contenidos Expertos
Tarea 1	
Presente	Sí
Ser	Sí
Estar	Sí
Tener	No
Llamarse	Sí
Por/para	Sí
Querer	No
Nacionalidades	Sí
Profesiones	Sí
Números	Sí
Saludos y despedidas	Sí

Contenidos Especificaciones	Contenidos Expertos
Tarea 2A	
Gustar	Sí
Hay y estar	Sí
Artículos	No
Comparativos	Sí
ir+a+ infinitivo	No
Lugares y servicios de una ciudad	Sí
Lugares emblemáticos de Bogotá y susc culturas de origen	Sí

Contenidos Especificaciones	Contenidos Expertos
Tarea 2B	
Indefinido	Sí
Imperfecto	No
Estar +gerundio	Sí
Vocabulario de descripción (físico, carácter)	Sí
Expresiones de temporalidad	No
Expresiones de valoración	No
Situaciones cotidianas /vida cotidiana	Sí

Contenidos Especificaciones	Contenidos Expertos
Tarea 3	
Futuro	No
Conectores de probabilidad /hipótesis	Sí
Conectores de valoración	Sí
Subjuntivo	No
Condicional	No
Medio ambiente	Sí
Léxico para valorar	No
Problemáticas sociales relacionadas con la contaminación en Colombia y sus países de origen	Sí

?