VALIDEZ Y FIABILIDAD DE UNA ESCALA DE EVALUACIÓN ORAL EN ELE[1]

VALIDITY AND RELIABILITY OF A SPEAKING SCALE FOR ASSESSING SPANISH AS A FOREIGN LANGUAGE

Begoña Martín Alonso
Pontificia Universidad Javeriana
bmartinalonso@javeriana.edu.co

RESUMEN

La presente investigación se enmarca en el campo de la Lingüística Aplicada a la Enseñanza de Español Lengua Extranjera, en lo que respecta a la validez y fiabilidad de pruebas orales con fines de clasificación. La complejidad viene determinada, en primer lugar, por los numerosos desafíos a los que todavía se enfrenta la evaluación de la lengua oral, principalmente en el campo de los exámenes de clasificación en ELE. En segundo lugar, por la escasez de investigaciones que han analizado la validez de contenido de escalas orales y a los desafíos a los que se enfrentan. A la luz de estas problemáticas, este artículo examina la validez de contenido de la escala de evaluación de una prueba oral de clasificación en ELE combinando la técnica del juicio de expertos con las observaciones realizadas a las actuaciones lingüísticas de alumnos durante el pilotaje del examen. Además, se calcula la consistencia interna y la fiabilidad interevaluador mediante los coeficientes *Alfa de Cronbach* y *Kappa de Cohen*. Los análisis cualitativos y cuantitativos revelaron la pertinencia de la metodología propuesta y permitieron concluir que la escala de evaluación diseñada es válida y fiable.

Palabras clave: validez de contenido, fiabilidad, escala de evaluación oral, juicio de expertos, actuaciones lingüísticas.

ABSTRACT

This research is part of the field of Applied Linguistics in Teaching Spanish as a Foreign Language, as regards the validity and reliability of speaking tests for classification purposes in SFL. The complexity is determined, first of all, by the fact that oral assessment faces numerous challenges, mainly in placement tests. Secondly, because the paucity of research that has examined content validity of assessing scales and the challenges it faces. In light of

---

[1] Este estudio forma parte de la Tesis Doctoral en Lingüística Aplicada a la Enseñanza de Español Lengua Extranjera de la Universidad Antonio de Nebrija "Diseño y validación de una prueba oral de clasificación en ELE". Tesis dirigida por Susana Martín Leralta e Irini Mavrou. Disponible en: https://biblioteca.nebrija.es/cgi-bin/opac/O8242/ID3e7f0cc5?ACC=161

these problems, this article aims to examine the content validity of the assessment scale for a SPF placement test by combining expert judgement methodology with the observations made to the linguistic performances of students during the piloting of the test. In addition, the internal consistency and the inter-rater reliability were calculated using the coefficients of Cronbach´s Alpha and Cohen´s Kappa. The qualitative and quantitative results verified the suitability of the proposed methodology and revealed that the assessment scale is valid and reliable.

Key words: content validity, reliability, speaking scale, expert judgement, linguistic performances.

Recibido: 01.03.2018. Aceptado: 11.10.2018

1. INTRODUCCIÓN

A pesar de la indiscutible importancia de la expresión e interacción oral en los procesos de enseñanza-aprendizaje, su trascendencia en el campo de la evaluación académica se limita, en muchos casos, a una calificación que otorgan los docentes en función del desempeño lingüístico del alumno a lo largo del curso. De hecho, se trata de la habilidad lingüística que menor atención ha recibido en el campo de la evaluación de lenguas extranjeras (LE) o segundas lenguas (L2) (Fulcher, 2003). Como resultado, se presenta como uno de los grandes desafíos tanto para los evaluadores como para los diseñadores de exámenes.

Estos problemas se agravan, concretamente, en la administración de pruebas orales de español lengua extranjera (ELE) con fines de clasificación. La frecuente incorporación de alumnos fuera de los plazos establecidos por los centros impide que se lleven a cabo los procesos de evaluación con la rigurosidad necesaria y supone, en muchos casos, que la evaluación de la lengua oral se limite a una entrevista informal con el alumno. Del mismo modo, se percibe un desequilibrio entre los contenidos de estas pruebas con los programas que se imparten en las instituciones. Tal es el caso de la prueba oral de clasificación de ELE que se emplea en la Pontificia Universidad Javeriana de Bogotá (PUJ), donde el enfoque por tareas que se sigue en los cursos de enseñanza y los contenidos que se imparten no se corresponden con los de esta prueba.

Paralelamente, la principal preocupación en el diseño de exámenes orales es comprobar que estos midan de manera consistente (sean fiables) y que las interpretaciones y usos que se realicen a partir de los resultados sean adecuados (sean válidos) (Bachman, 1990). En concreto, la validez se presenta como uno de los aspectos más destacados a la hora de desarrollar y evaluar pruebas lingüísticas (Bordón, 2015) y, sin embargo, los esfuerzos en el campo de la evaluación de la lengua oral han estado destinados a estimar la fiabilidad (Van der Walt y Steyn, 2008). Adicional a la escasez de investigaciones que se han centrado en la validez, principalmente en el ámbito de ELE (Martín Alonso, 2017b), se suman la noción disgregada y desarticulada de esta cualidad y la práctica habitual de analizar los diferentes tipos de evidencias como si se tratara de una *hoja de verificación* (Bachman, 2005).

La presente investigación sostiene, junto con Bachman y Palmer (2013) y Mendoza Ramos (2015), que en el proceso de validación se deben recoger diferentes tipos de evidencias (validez de constructo, de contenido o de criterio) que sustenten la adecuación

de las decisiones que se toman y las conclusiones que se extraen a partir de los resultados de la prueba (por ejemplo, clasificar al alumno en un determinado nivel de lengua). No obstante, se aclara que este proceso no consiste en examinar los diferentes tipos de evidencias como si fuera una hoja de verificación, sino que se necesitan considerar primero las interpretaciones y decisiones que son importantes para el propósito de la prueba y después seleccionar los tipos de evidencias que se recolectan (Kane, 1992; Bachman, 2005; Bachman y Palmer, 2013).

De igual forma, este artículo se centra en la validez de contenido y parte de la consideración de que el juicio de expertos es la metodología más adecuada para examinarla pero que también es necesario comprobar que el dominio lingüístico que queremos medir es el que realmente producen los estudiantes durante la realización de las tareas (O´Sullivan, Weir y Saville, 2002). Además, se asienta sobre la constatación de que tanto la validez como la fiabilidad son interdependientes en términos de que esta última se manifiesta como requisito para la validez de una prueba.

A la luz de las reflexiones anteriores, en el marco de la prueba oral de clasificación de la PUJ, esta investigación tiene como propósito recoger evidencias que sustenten la validez de contenido y la fiabilidad de la escala de evaluación oral. Para ello, se plantean los siguientes objetivos específicos: determinar la adecuación de los descriptores de la escala de evaluación oral en lo concerniente a su formulación; medir el grado en que los descriptores de la escala de evaluación reflejan el uso de la lengua que producen los candidatos en la realización de las tareas, y calcular la consistencia interna y la fiabilidad interevaluador de la escala.

A continuación, se definen los conceptos clave y se describe el procedimiento metodológico llevado a cabo para dar respuesta a los objetivos establecidos. Por último, se exponen los resultados de los análisis cualitativos y cuantitativos realizados y se discuten los alcances y limitaciones de la metodología propuesta.

2. LA VALIDEZ

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 junto a la fiabilidad se manifiestan como dos cualidades necesarias para garantizar la utilidad de una prueba. En concreto, en esta investigación la validez se considera uno de los primeros aspectos que se debe tener en cuenta en el diseño de exámenes puesto que, si "una prueba no es válida para el objetivo para el que se ha preparado, los resultados no significan lo que se cree que significan" (Alderson, Clapham y Wall, 1999, p.165).

El concepto de validez emergió en el campo de la evaluación de lenguas a mitad del s. XX –hasta entonces exclusivo del ámbito de la psicometría–, gracias a las aportaciones de Lado (1964) quien la definió como una propiedad de la prueba relacionada con el grado en que esta mide lo que pretende medir. Paralelamente, la validez se empezó a percibir como una cualidad relativa y específica a cada contexto de evaluación en el sentido de que no se podía hablar de validez alta, media o baja de una prueba a no ser que se tuviera en cuenta el propósito para el que fue creada.

A finales del siglo XX, la concepción de validez experimentó un cambio significativo y se empezó a concebir como un juicio evaluativo sobre la representatividad y adecuación de las conclusiones que extraemos (por ejemplo, buen dominio gramatical) y las decisiones que tomamos a partir de las actuaciones lingüísticas de los alumnos (por ejemplo, clasificar en un determinado curso de lengua). De esta manera, desde finales de los ochenta, la validez ya no se percibe como una cualidad de la prueba de si esta mide lo que pretende medir (Lado, 1964), sino como una propiedad de la interpretación y usos de los resultados de la misma (Bachman, 1990; D´Este, 2012; Kane, 1992; Van der Walt y Steyn, 2008). Por consiguiente, para validar un examen se deben presentar diversas evidencias que sustenten que estas interpretaciones y acciones, que realizamos en función del desempeño lingüístico del estudiante, son apropiadas y significativas.

De lo que se ha expuesto hasta ahora se infiere que el proceso de validación no es un evento único ni estático, sino un proceso continuo que reside en la recogida de una cantidad suficiente de evidencias que nos permita demostrar que una determinada prueba es válida (D´Este, 2012). A pesar de que la validación se puede llevar a cabo antes, durante o después del diseño de la prueba, se considera que esta debe realizarse antes de que los resultados se puedan utilizar para cualquier propósito en particular. Lo anterior debido a que cuanto más capaces seamos de definir el constructo que queremos medir en la fase a priori, más significativos serán los procesos estadísticos posteriores (Van der Walt y Steyn, 2008).

En lo que respecta a qué evidencias son más relevantes y cuántas se deben recoger en este proceso, no existe unanimidad entre los autores. Esta investigación se asienta sobre el principio de que no consiste en examinar los diferentes tipos de evidencias como si fuera una *hoja de verificación* (Bachman, 2005). Por el contrario, se fundamenta en que primero se han de considerar las interpretaciones y decisiones que son importantes para el propósito de la prueba y después seleccionar los tipos de evidencias que se recolectan (Bachman, 2005; Kane, 1992).

El propósito de la prueba objeto de este estudio es clasificar a los estudiantes en los niveles iniciales e intermedios de ELE de la PUJ. Este tipo de pruebas -al igual que las finales o las de progresos-, son de aprovechamiento; esto es, este tipo de exámenes se utilizan para tomar decisiones a nivel de un programa de LE o de L2 y, por tanto, deben estar alineados con los contenidos de un determinado curso (Martín Alonso, 2017b). Como resultado, se presenta inminente iniciar el proceso de validación recogiendo evidencias que respalden que el contenido de la prueba oral de clasificación de la PUJ es representativo y relevante para el constructo que se quiere medir. En otras palabras, esta investigación propone examinar que la expresión e interacción orales que mide la escala de evaluación diseñada son acordes con la lengua que se imparte en los programas de ELE de la PUJ.

2.1. La validez de contenido

Entre los diferentes tipos de evidencias que se deben recoger en el proceso de validación, validez de constructo, validez de criterio (concurrente y predictiva) y validez de contenido, se hace énfasis en la pertinencia de iniciar este proceso examinando esta última; por un lado, en la medida que permite relacionar las actuaciones lingüísticas de los estudiantes en el desempeño del examen con el constructo que se quiere medir (Association of Language

Testers in Europe, 2005; Martín Alonso, 2017a). Por otro lado, debido a su relación con la fiabilidad en el sentido de que ambas garantizan que la prueba representa adecuadamente los objetivos y contenidos lingüísticos de la prueba (Association of Language Testers in Europe, 2005; Martín Alonso, 2017a).

En lo referente a su definición, existe unanimidad a la hora de concebirla como el grado en que el contenido de una prueba (tareas, ítems, escalas) es relevante y representativo para el constructo (habilidad lingüística) que se pretende medir. De la anterior definición, se desprenden los tres aspectos esenciales de la validez de contenido que han estado presentes desde mediados del siglo XX, *definición del dominio, representación del dominio* y *relevancia del dominio*.

La definición del dominio consiste en describir las habilidades lingüísticas que evalúa una determinada prueba. Para tal fin, se puede tomar de referencia un modelo teórico o bien los contenidos de un programa, en el caso de los exámenes de clasificación. Esta definición generalmente se recoge en las especificaciones de la prueba.

La representatividad del dominio concierne al grado en que una determinada prueba representa y mide adecuadamente la habilidad lingüística definida. Para ello, normalmente se emplea la técnica del juicio de expertos y se les solicita que analicen el contenido de las tareas a partir de modelos teóricos de habilidad lingüística (Alderson, 1990; Bachman, Davidson y Milanovic, 1996), o sobre la base de su pericia y conocimiento de los programas y estudiantes (Cumming, Grant, Mulcahy-Ernst y Powers, 2004; Wall, Clapham y Alderson, 1994).

La relevancia del dominio se refiere al grado en que cada ítem de una prueba es significativo para el dominio definido en las especificaciones. Se suele investigar pidiendo a un grupo de expertos que califique el grado en que cada ítem es importante para determinados aspectos de las especificaciones (Sireci y Faulkner-Bond, 2014) o para los contenidos de un determinado programa (Cumming y otros, 2004).

Cabe señalar que, a pesar de la aparente simplicidad de este concepto, la validez de contenido se enfrenta a numerosos problemas. Entre estos, se destacan: la complejidad de definir el dominio lingüístico y seleccionar tareas o muestras de lengua que reflejen satisfactoriamente dicho dominio (Bachman, 2002); la dificultad de alcanzar juicios de expertos unánimes (Alderson, Clapham y Wall, 1999); la inexistencia de un procedimiento metodológico que guíe adecuadamente la recogida de este tipo de evidencias (Newman, Lim y Pineda, 2013); y la escasez de investigaciones que han investigado la validez de contenido, fundamentalmente en el ámbito de ELE (Martín Alonso, 2017b).

3. LA FIABILIDAD

Tradicionalmente la validez y la fiabilidad se han concebido como cualidades interdependientes en términos de que la fiabilidad se presenta como un requisito necesario para analizar la validez de una prueba (Fitzpatrick y Clenton, 2010). En otras palabras, un examen que no mide de manera consistente (fiable) no puede medir con precisión (válida) (Alderson, Clapham y Wall, 1999). Además, se presentan como aspectos complementarios para identificar, estimar e interpretar diferentes fuentes de varianza en los resultados de una prueba. Por un lado, la fiabilidad analiza las divergencias observadas en los resultados de

una prueba a causa de errores de medición u otros errores. Por otro lado, la validez identifica los factores que causan los desacuerdos en los resultados de una prueba.

Fruto de esta estrecha relación, esta investigación se centra en la validez de contenido, pero también toma en consideración la fiabilidad entendida como el grado en que las mediciones son consistentes y están libres de errores (Association of Language Testers in Europe, 2005). Es decir, una prueba es fiable cuando arroja los mismos resultados independientemente de que se modifiquen algunas características de la situación de la prueba (por ejemplo, diferentes evaluadores) y cuando sus resultados están libres de errores aleatorios y se puede depender de ellos a la hora de tomar decisiones sobre los candidatos. Es necesario aclarar que, por el término errores, no se hace alusión simplemente a los fallos causados por el proceso de medición sino también a los orginados por factores externos, como los aspectos del método de la prueba (rúbrica, *input*), las características personales (edad, sexo) o bien los debidos a factores aleatorios (cansancio, estado anímico) (Bachman, 1990).

De la anterior definición de fiabilidad se desprende que, si se quiere examinar esta cualidad de la prueba, se deben abordar las posibles causas de error que pueden repercutir en la consistencia de los resultados. Existen varios procedimientos para medir la fiabilidad (paralela, test-restest), pero se destacan la *consistencia interna* y la *fiabilidad evaluadora* (inter e intra-evaluadora).

La consistencia interna evalúa la coherencia entre los elementos internos de una prueba, por ejemplo, la homogeneidad de los ítems. Por su parte, la fiabilidad evaluadora investiga la calidad de los juicios realizados por los calificadores. Por un lado, la fiabilidad intraevaluador examina el grado en que los evaluadores son consistentes consigo mismos (Del Moral, 2015). Por otro lado, la fiabilidad interevaluador calcula la correspondencia entre las calificaciones otorgadas por dos o más evaluadores al mismo candidato sin influirse el uno al otro (Wang, 2009).

Respecto a los procedimientos metodológicos existentes para estimar la consistencia interna y la fiabilidad interevaluador, las investigaciones existentes normalmente pilotan la escala con un número elevado de estudiantes y analizan estadísticamente las calificaciones otorgadas por varios evaluadores (Fulcher, 1997; Wall, Clapham y Alderson, 1994). Entre los coeficientes estadísticos destacan *Alfa de Cronbach*, para la consistencia interna, y el *Kappa de Cohen*, para la fiabilidad interevaluador (Dunsmuir, Kyriacou, Batuwitage, Hinson, Ingram y O´Sullivan, 2015).

4. METODOLOGÍA

4.1. Contexto

Los presentes estudios se enmarcan en una tesis doctoral que se propone diseñar y examinar la validez de contenido y la fiabilidad la prueba oral -tareas y escalas de evaluación- del nuevo examen clasificación del Centro de Lenguas de la PUJ (Martín Alonso, 2017b). Para tal fin, en primer lugar, se diseñaron las tareas de evaluación y se validó su contenido mediante el juicio de expertos y la observación del desempeño lingüístico de los estudiantes durante el pilotaje de la prueba (Martín Alonso, 2017a; Martín Alonso, 2017b). En segundo

lugar, se diseñó y validó el contenido de la escala de evaluación. Finalmente, se calculó la consistencia interna y la fiabilidad interevaluador.

En concreto, en esta investigación se precisa el procedimiento llevado a cabo para examinar la validez de contenido y la fiabilidad de la escala de evaluación oral que servirá para clasificar a los estudiantes en los niveles iniciales e intermedios de ELE de la PUJ.

Esta escala está conformada por seis niveles de dominio lingüístico, presentados de manera ascendente, que se corresponden con los cursos de ELE que se imparten en este centro: ELE 1 (A1.1), ELE 2 (A1.2), ELE 3 (A2.1), ELE 4 (A2.2), ELE 5 (B1.1) y ELE 6 (B1.2). Se optó por una escala de evaluación analítica con los cuatro criterios de calificación que se tienen en cuenta en los exámenes de aprovechamiento que se realizan en la PUJ: alcance, corrección, fluidez y pronunciación. Para la redacción de los descriptores de los criterios de alcance y corrección, se consultaron las funciones lingüísticas y los contenidos lingüísticos detallados en los programas de ELE del Centro (Martín Alonso, 2017b). Para las escalas de pronunciación y fluidez, se tomaron de referencia las escalas de evaluación analítica de los exámenes DELE A1, A2, B1 y B2 y las escalas del MCER del Consejo de Europa (2002). Además, los descriptores se formularon de manera positiva, se evitaron el uso de jerga o palabras valorativas o genéricas y se persiguió la claridad y precisión de los mismos (Consejo de Europa, 2002; Luoma, 2004).

De igual forma, para el diseño de los descriptores de la escala de evaluación oral se tomaron en consideración las cuatro tareas que se diseñaron: Tarea 1 *Calentamiento* donde el candidato tiene que presentarse y dar información personal sobre sí mismo; tarea 2A y 2B *Bogotá y yo* en las que el candidato tiene que describir la ciudad de Bogotá, compararla con su ciudad natal y relatar alguna experiencia que ha vivido; Tarea 3 *Contaminación en Bogotá* donde el alumno tiene que dar su opinión y valoración respecto a la contaminación y a las medidas que existen para combatirla (Martín Alonso, 2017a, 2017b).

4.2. Objetivos

Por medio de tres estudios empíricos se responde a cuatro objetivos específicos: determinar la adecuación de los descriptores de la escala de evaluación oral en lo concerniente a su formulación; evaluar el grado en que los descriptores de la escala de evaluación reflejan el uso de la lengua que producen los candidatos en la realización de las tareas; calcular la consistencia interna y la fiabilidad interevaluadora de la escala.

4.3 Procedimientos

Para determinar la adecuación de los descriptores de la escala de evaluación oral, se empleó la técnica de juicio de expertos, tal y como realizaron Del Moral (2015), Deygers y Van Gorp (2015) y Robles y Rojas (2015). En concreto, se solicitó a los docentes que valoraran cualitativamente la extensión, la enunciación, la independencia y la adecuación de los descriptores de la escala. Además, se les pidió que identificaran el nivel que medía cada descriptor de la escala. Para ello, se diseñó un cuestionario en línea (véase apartado 4.5) y se estableció un plazo de dos meses para que los expertos lo diligenciaran.

Para examinar el grado en que los descriptores de la escala de evaluación reflejan el uso de la lengua que producen los candidatos en la realización de las tareas, se observaron las

actuaciones lingüísticas de alumnos de ELE durante el pilotaje de la escala de evaluación oral. Las entrevistas se llevaron a cabo durante el primer y segundo semestre de 2016 y se grabaron con previo consentimiento del estudiante. Como herramienta de toma de datos, se diseñaron listas de chequeo (véase apartado 4.5) que el investigador principal diligenció tras observar cada grabación tres veces. De igual forma, se compararon sus respuestas con las transcripciones de las entrevistas.

Para examinar la consistencia interna y la fiabilidad interevaluador de la escala, se llevó a cabo un segundo pilotaje de la escala con alumnos de ELE. Las entrevistas se grabaron y se solicitó a dos evaluadores expertos que calificaran a los estudiantes después de visualizar cada muestra dos veces. Las calificaciones otorgadas se analizaron cuantitativamente mediante los coeficientes *Alfa de Cronbach* y *Kappa de Cohen*.

4.4 Participantes

El cuestionario de validación lo diligenciaron seis profesores especialistas en el campo de ELE de la PUJ. Se seleccionaron exclusivamente docentes del contexto donde se administrará la prueba, tal y como realizaron Fulcher (1997) y Wall, Clapham y Alderson (1994), puesto que se trata de una prueba de aprovechamiento y se considera que los validadores deberían ser los futuros usuarios de la prueba. Además, cabe señalar que estos expertos fueron los mismos que participaron en la validación de contenido de las tareas de la prueba oral de clasificación (Martín Alonso, 2017a).

En el primer pilotaje de la escala, se contó con la colaboración de diez alumnos. Todos ellos eran representativos de los futuros candidatos de la prueba oral de clasificación. Para conformar la muestra, se visitaron las clases de niveles iniciales e intermedios que se impartían en ese momento en la PUJ y se habló con los docentes encargados. Además, para garantizar la representatividad de estos participantes, por un lado, se tomaron como base los resultados del análisis del perfil de los candidatos de la prueba de clasificación que se realizó a partir de una muestra de aproximadamente quinientos alumnos, inscritos en la PUJ desde 2010 a 2015 (Martín Alonso, 2017b). Por otro lado, se solicitó a dos expertos que otorgaran a cada uno de estos alumnos, un nivel de dominio lingüístico por cada criterio mediante la escala de evaluación diseñada (alcance, corrección, fluidez y pronunciación).

En el segundo pilotaje de la escala participaron treinta estudiantes de ELE. Para constituir la muestra se realizó el mismo procedimiento que en el primer pilotaje. La única diferencia residió en que, para garantizar la aplicabilidad de la escala, se tomaron en consideración estudiantes de ELE no solo de la PUJ sino de otras universidades colombianas. Por otro lado, dos profesores de la PUJ evaluaron las muestras. Estos se seleccionaron puesto que eran los únicos que en ese momento realizaban los exámenes de clasificación y, además, habían participado en la validación de contenido de las tareas y de la escala.

4.5. Instrumentos

Para recoger las valoraciones de los expertos sobre el contenido de los descriptores, se diseñó un cuestionario en línea con la aplicación Google Drive[2] (Martín Alonso, 2017b). Este cuestionario se desarrolló tomando como referencia el diseñado por Del Moral (2015).

En total, se formularon veintiocho preguntas. De estas, 24 eran tipo Likert y pedían al evaluador que valorara la extensión, enunciación, adecuación e independencia de cada uno de los descriptores de clasificación en una escala de cinco puntos (1=totamente en desacuerdo; 2= en desacuerdo; 3= ni de acuerdo ni en desacuerdo; 4= de acuerdo; 5= totalmente de acuerdo). Las cuatro preguntas restantes eran de emparejamiento donde el experto, según su punto de vista, tenía que relacionar cada descriptor con el nivel que medía. Además, se añadió una casilla en blanco para que los docentes dejaran sus comentarios.

Para recoger las observaciones del primer pilotaje de la escala, se emplearon listas de chequeo, en consonancia con O´ Sullivan, Weir y Saville (2002). En total, se diseñaron cuatro listas de chequeo dicotómicas (Sí/No), una para cada criterio de calificación (alcance, corrección, fluidez y pronunciación). Los ítems que conformaron estas listas fueron las categorías nominales en las que se descompuso cada descriptor de la escala (véase Tablas I y II).

Tabla I. Ejemplo de las categorías para el descriptor A1.2 del criterio de corrección

Descriptor A1.2 de la escala de corrección	**Categorías**
Utiliza pocas construcciones gramaticales como verbos en presente regular e irregular de algunos verbos básicos (estar, haber, encontrarse), forma impersonal con "se", cuantificadores (ningún, algún), preposiciones y adverbios de lugar (detrás, al lado de, en frente de). Comete muchos errores principalmente de conjugación.	-Presente regular e irregular -Impersonal con se -Cuantificadores -Preposiciones y adverbios de lugar -Errores/conjugación

Tabla II. Ejemplo de la lista de chequeo para el descriptor A1.2 del criterio de corrección

Categorías	**SÍ**	**NO**
Presente regular e irregular	√	√
Impersonal con se	√	√
Cuantificadores	√	X

[2] Cuestionario disponible en: https://docs.google.com/forms/d/e/1FAIpQLSciAZ_0FMPsVE3MqwtmTzH8UHn945L7_szQcUY3padZs8N_Pw/viewform

5. RESULTADOS Y DISCUSIÓN

A continuación se presentan, en primer lugar, los resultados de las valoraciones que realizaron los expertos sobre la formulación de los descriptores en lo que respecta a su extensión, enunciación, adecuación, independencia y nivel (véase apartado 5.1). En segundo lugar, se muestra el grado en que los descriptores de la escala de evaluación reflejaban el uso de la lengua que produjeron los candidatos durante el pilotaje de la prueba (véase apartado 5.2). En tercer lugar, se exponen los resultados de los análisis estadísticos que se llevaron a cabo para estimar la consistencia interna y la fiabilidad interevaluador de la escala de evaluación oral (veáse apartado 5.3).

5.1. Resultados de la validación con expertos

5.1.1. *Extensión, enunciación, adecuación e independencia*

Para analizar las valoraciones –veinticuatro preguntas tipo Likert- que realizaron los expertos sobre la formulación de los descriptores de la escala en el cuestionario en línea se calculó la Moda (Mo) y la Media (Me).

Como se aprecia en los Gráficos 1, 2 y 3, la mayoría de los evaluadores expresó que estaba de acuerdo (Mo: 4) con la extensión, la enunciación y la adecuación de los descriptores de la escala de evaluación. Por el contrario, tal y como se observa en el Gráfico 4, manifestaron que estaban ni de acuerdo ni en desacuerdo (Mo:3) con la independencia de los descriptores de las escalas de corrección y de alcance y en desacuerdo con los de las escalas de pronunciación e independencia (Mo:2).

Con respecto a la Media, en líneas generales los docentes señalaron que estaban de acuerdo (Me: 4) con la extensión, enunciación y adecuación de los descriptores de la escala de evaluación (véase Gráficos 1, 2 y 3). No obstante, se destacan los resultados un poco menos satisfactorios en lo que respecta a la adecuación y la extensión de los descriptores de la escala de fluidez (véase Gráficos 2 y 3). En lo concerniente a la independencia de la escala de evaluación, en la mayoría de los casos los expertos expresaron que estaban ni de acuerdo ni en desacuerdo (Me:3), siendo la Me más desfavorable en los criterios de fluidez y pronunciación (véase Gráfico 4).

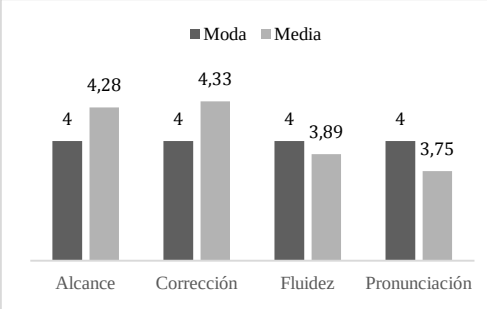

Gráfico 1. Resultados de la enunciación

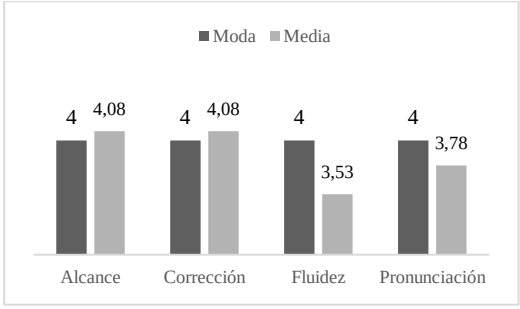

Gráfico 2. Resultados de la adecuación.

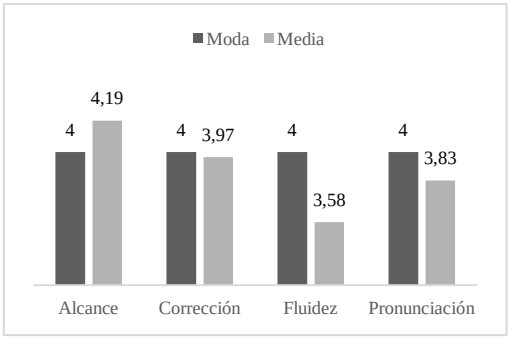

Gráfico 3. Resultados de la extensión

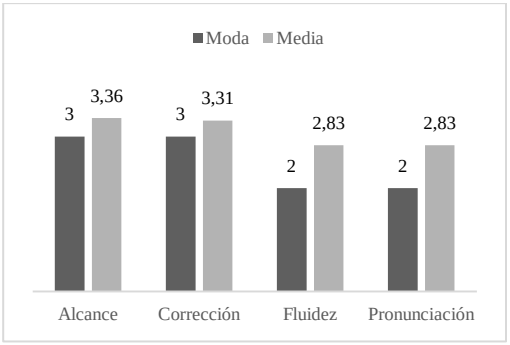

Gráfico 4. Resultados de la independencia

Los resultados registrados en la Media de los descriptores de la escala de evaluación, principalmente en los criterios de fluidez y pronunciación, se atribuyen a las discrepancias que se dieron entre los docentes (Martín Alonso, 2017b). Los comentarios que algunos dejaron por escrito en la casilla en blanco del cuestionario, permitieron comprender algunas de estas divergencias y valoraciones desfavorables.

Por ejemplo, respecto a la extensión, un docente explicó que los descriptores del criterio de corrección le parecían demasiado extensos y manifestó su preferencia por descriptores más cortos y concretos. Por el contrario, dos evaluadores señalaron que los descriptores de la escala de pronunciación les parecían demasiado breves y expresaron que se sentían más cómodos si había más información y ejemplos en el descriptor. Estas apreciaciones, por un lado, llevarían a pensar que existe un desequilibrio en lo que respecta al número de palabras de los descriptores de ambas escalas. Por otro lado, se podría inferir la dificultad de encontrar un equilibrio entre la cantidad de información necesaria para caracterizar el nivel de dominio lingüístico del estudiante y la aplicabilidad.

En cuanto a la enunciación de los descriptores de la escala de pronunciación, un experto argumentó que la diferencia entre los niveles adyacentes era muy sutil y dependía de adverbios cuantificadores. Las observaciones de este docente llevarían a pensar que estos descriptores no estaban bien formulados debido a que se considera, junto con el MCER, que las diferencias entre los descriptores de niveles adyacentes no deberían depender de términos como *mucho*, *poco* o *bastante*. No obstante, cabe señalar que la mayoría de los profesores (el 66,7%) valoró positivamente la enunciación de los descriptores de esta escala, lo cual llevaría a pensar que bien la gradación de los niveles no depende exclusivamente de estos términos o que estos participantes perciben la vaguedad y ambigüedad como una ventaja en términos de que se pueden aplicar en cualquier situación de prueba (Fulcher, 2003).

En lo referente a la independencia de los descriptores, tres expertos explicaron que estos no podían interpretarse de manera aislada y debían consultar los otros descriptores para determinar el nivel. Estas opiniones pondrían de manifiesto la complejidad de crear descriptores completamente independientes y la necesidad de revisar la formulación de los mismos.

5.1.2. *Nivel de dominio lingüístico*

Para analizar las respuestas a las preguntas de emparejamiento del cuestionario de validación, se calculó el porcentaje total de descriptores cuyo nivel lograron identificar los expertos.

Como se puede apreciar en el Gráfico 5, los aciertos de los docentes fueron mayores en las escalas de corrección y fluidez que en las escalas de alcance y pronunciación.

Llama la atención que en ningún caso registró el 100% de aciertos. No obstante, cabe señalar que sí hubo docentes que consiguieron identificar el nivel de todos los descriptores de la escala (el 50% de los evaluadores en la escala de alcance y pronunciación; el 83,3% en la escala de corrección; el 66,7% en la escala de fluidez).

Otro aspecto en el que habría que detenerse es el elevado número de aciertos en las escalas de pronunciación y fluidez y la cantidad de errores que cometieron en los descriptores de la escala de alcance. Estos resultados parecen contradecirse con las valoraciones que otorgaron algunos docentes en lo concerniente a la extensión, enunciación, adecuación e independencia de los descriptores de estas escalas (véase apartado 5.1.1).

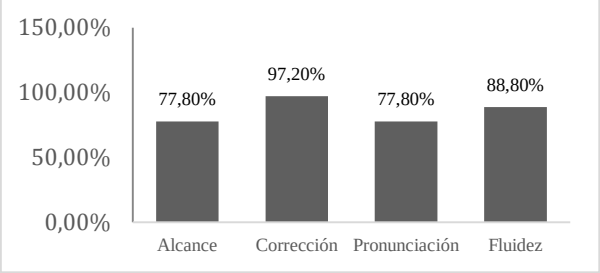

Gráfico 5. Resultados del nivel de los descriptores

Una última observación que debería considerarse concierne a las respuestas que dieron tres expertos en las escalas de alcance y de corrección puesto que otorgaron el mismo nivel a dos descriptores diferentes (véase Tabla III). Estos resultados llevarían a pensar que hubo un problema en el diseño del cuestionario y que sería conveniente modificar el formato de las preguntas para evitar que se repitan los niveles.

Tabla III. Resultados del nivel de las escalas de alcance y corrección

	Alcance		**Corrección**
Descriptores	Ev #1	Ev #2	Ev #1
A1.1	A1.1	A1.1	A1.1
A1.2	A1.2	**A1.1**	A1.2
A2.1	**A1.2**	A1.2	**A1.2**
A2.2	A2.1	A2.2	A2.2
B1.1	B1.1	**A2.2**	B1.1

En definitiva, los resultados recogidos en el cuestionario de validación resultan un tanto contradictorios y vislumbran problemas en la formulación de los descriptores, principalmente en lo que respecta a las escalas de fluidez y pronunciación. A raíz de estos resultados, se pone en entredicho la validez de contenido de la escala y se estima necesario seguir recogiendo más evidencias.

5.2. Resultados de la validación con muestras de alumnos

Para analizar los datos recogidos en las listas de chequeo, se crearon las siguientes categorías:

- Se corresponde (C): la categoría del descriptor se corresponde con la lengua que manifestaron los estudiantes durante la realización de las tareas de evaluación;
- Se corresponde parcialmente (CP): la categoría del descriptor no se corresponde con el discurso de la mayoría de los alumnos en el desempeño de las tareas de evaluación;
- No se corresponde (NC): la categoría del descriptor no se corresponde con las actuaciones lingüísticas de los aprendientes en la realización de las tareas de evaluación;
- No aplica (NA): no se pudo analizar la correspondencia entre la categoría del descriptor y la lengua que manifestaron los estudiantes debido a problemas en la enunciación de la misma.

En los Gráficos 6, 7 y 8 se presentan los porcentajes de las categorías de los descriptores que se observaron en el primer pilotaje de la escala.

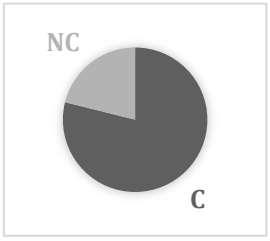

Gráfico 6. Resultados de alcance

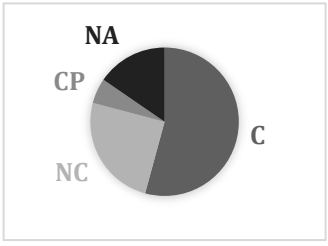

Gráfico 7. Resultados de corrección

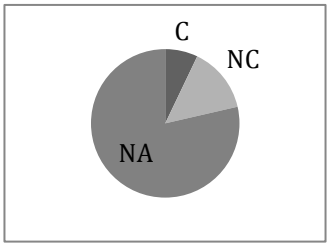

Gráfico 8. Resultados de fluidez

El primer aspecto en el que habría que detenerse es que no se analizó la correspondencia entre la escala de pronunciación y las actuaciones lingüísticas de los alumnos. Lo anterior debido a que los dos docentes, que tenían como labor calificar a los estudiantes para otorgarles un nivel de dominio lingüístico según los cuatro criterios de la escala (véase apartado 4.4), manifestaron la imposibilidad de clasificar a los estudiantes con esa escala debido a la ambigüedad de los descriptores y a que estos no se correspondían con las actuaciones lingüísticas de los alumnos piloto. Las opiniones de estos expertos corroborarían las valoraciones negativas recogidas en el cuestionario de validación de los

descriptores (véase apartado 5.1). Por tanto, se concluye que el contenido de la escala de pronunciación no es válido para clasificar a los estudiantes en los niveles iniciales e intermedios de ELE de la PUJ y se constata la complejidad de definir y nivelar la pronunciación (Llisterri, 2003).

En lo que respecta a la escala de alcance, como se puede apreciar en el Gráfico 6, los resultados son bastantes alentadores puesto que se observaron el 78,9% de las categorías de los descriptores y solamente en el 21,1% de las categorías no se encontró correspondencia. Al respecto de estas últimas se estima necesario revisar su pertinencia.

Por el contrario, en la escala de corrección los resultados resultan menos esperanzadores puesto que el 25% de las categorías de los descriptores no se correspondieron (NC) con el discurso de los alumnos. Estos resultados se pudieron atribuir a que se incluyeron demasiados exponentes gramaticales en los descriptores. Por lo tanto, se confirmarían las valoraciones que realizaron algunos expertos en el cuestionario sobre el desbalance que percibían en la extensión de los descriptores. Además, pusieron en evidencia la complejidad de encontrar un equilibrio entre la concreción y la aplicabilidad, en tanto que cuanto más específica sea la información incluida en los descriptores más difícil resulta encontrar muestras de alumnos que reflejen la lengua que se describe en ellos (Luoma, 2004). Como resultado, se percibió la necesidad de revisar la pertinencia de algunas categorías de esta escala y equiparar el número de palabras que contenían los descriptores de las escalas.

Por otro lado, se señala que el 15.5% no se pudieron analizar (NA) debido a que había problemas en su enunciación. Se destaca el caso de la categoría *errores,* presente en todos los descriptores, donde no se pudo asociar los tipos de errores que cometía el alumno con su nivel de dominio lingüístico. Estos resultados coincidirían con lo expuesto por Campillos Llanos (2012) quien afirma que:

> si bien la existencia de distintas etapas de adquisición hace que los errores se sigan produciendo en algunas estructuras (especialmente, los pronombres clíticos) y lo que se espera es la producción de errores hasta que el hablante haya interiorizado el funcionamiento del sistema. Aún así, lo cierto es que algunos puntos (como el artículo o las preposiciones) siguen originando incorrecciones incluso en el *buen aprendiz* de nivel avanzado. (p.361)

Como resultado de la complejidad de asociar ciertos errores con el nivel de dominio lingüístico en ELE, se optó por no incluirlos en los descriptores de la escala pero sí tomarlos en consideración en la formación de evaluadores, tal y como recomienda Fulcher (2003).

En lo concerniente a la escala de fluidez, como se puede apreciar en el Gráfico 8, los resultados registrados resultan bastante desalentadores: C (7,1%), NC (14,3%), NA (78,6%). Destacan las múltiples categorías que no se pudieron aplicar (NA) debido a problemas en su formulación. Es decir, estas estaban presentes en varias bandas de la escala y su formulación dependía de cuantificadores como *raramente, a veces* o *muchas veces*. Estas observaciones constataron las valoraciones que realizaron algunos expertos sobre la enunciación e independencia de los descriptores de esta escala (véase apartado 5.1.1). De igual forma, confirmarían la complejidad de definir y nivelar este criterio de

calificación (Horche y Marco, 2008). Como resultado, se percibió la necesidad de reformular los descriptores.

Un último aspecto que fue determinante en la investigación fueron las estrategias de comunicación que se observaron en los alumnos de niveles iniciales, tales como las preguntas para solicitar ayuda al interlocutor, las inferencias de su lengua materna para la creación de nuevas palabras o el uso de la lengua extranjera cuando no conocían el término en ELE. Estos resultados llevarían a pensar en la idoneidad de evaluar las estrategias de interacción desde niveles principiantes.

En definitiva, la validación de la escala mediante la observación de muestras de alumnos, puso en evidencia los problemas que se habían vislumbrado en la validación con expertos y constató que el contenido de la escala de evaluación diseñada no era válido. A la luz de estos resultados, se revisó la adecuación, extensión, enunciación e independencia de los descriptores y se diseñó una nueva versión de la escala de evaluación con cuatro criterios de calificación (alcance, corrección, fluidez y pronunciación e interacción).

Se estimó pertinente aunar los criterios de fluidez y pronunciación, tal y como aparecen en las escalas del examen DELE niveles B1, B2, C1 y C2, debido a los problemas encontrados en las escalas de pronunciación y fluidez y a la complejidad de evaluarlos de manera aislada. En lo que respecta al criterio de interacción, se decidió integrarlo debido a los resultados del pilotaje. En este punto, cabe señalar que las escalas del examen DELE evalúan este aspecto, dentro de los criterios de coherencia y fluidez, a partir del nivel B1. No obstante, debido a las características de las tareas de la prueba oral de clasificación, la propia naturaleza de la lengua oral y a las observaciones realizadas durante el pilotaje de la escala, se consideró que estas se deberían evaluar desde el nivel A1.

5.3. Resultados de la fiabilidad

Los resultados obtenidos se presentan en dos apartados. En primer lugar, se exponen los valores que obtuvo el coeficiente *Alfa de Cronbach* para calcular la consistencia interna (véase apartado 5.3.1). En segundo lugar, se muestran los valores que registró el coeficiente de *Kappa de Cohen* a la hora de medir la fiabilidad interevaluador (véase apartado 5.3.2).

5.3.1. *Resultados de la consistencia interna*

Para examinar la consistencia interna de la nueva escala de evaluación, se empleó el coeficiente *Alfa de Cronbach* a las calificaciones otorgadas a los estudiantes por parte de los dos evaluadores. Los valores de este coeficiente fueron superiores a 0.90 en la consistencia interna de la escala (*Alfa de Cronbach* $_{(EV1)}$=0.978 y *Alfa de Cronbach* $_{(EV2)}$ =0.982). Los valores en las correlaciones inter-ítem oscilaron entre 0.847 (corrección y fluidez) y 0.953 (corrección y alcance) en el caso del primer evaluador (véase Tabla IV), y entre 0.892 (corrección e interacción) y 0.965 (fluidez e interacción) en el segundo evaluador (véase Tabla V).

Los valores que adquiere este coeficiente fluctúan entre 0-1 (Gwet, 2014). El valor mínimo se sitúa en .60, mientras que valores entre .70 y .90 indican que el instrumento posee una buena consistencia interna (González Alonso y Pazmiño, 2015). Por lo tanto, se

puede afirmar que la escala de evaluación analítica diseñada para los propósitos del presente estudio tiene una consistencia interna muy satisfactoria.

Tabla IV. Resultados de las correlaciones inter-ítem EV_1

	Alcance *EV1*	Corrección *EV1*	Fluidez *EV1*	Interacción *EV1*
Alcance *EV1*	1.000	.953	.891	.951
Corrección *EV1*	.953	1.000	.847	.913
Fluidez y Pronunciación *EV1*	.891	.847	1.000	.941
Interacción *EV1*	.951	.913	.941	1.000

Tabla V. Resultados de las correlaciones inter-ítem EV_2

	Alcance *EV2*	Corrección *EV2*	Fluidez *EV2*	Interacción *EV2*
Alcance EV1	1.000	.925	.960	.949
Corrección EV1	.925	1.000	.894	.892
Fluidez y Pronunciación EV1	.960	.894	1.000	.965
Interacción EV1	.949	.892	.965	1.000

5.3.2. *Resultados de la fiabilidad interevaluador*

Para calcular la fiabilidad interevaluador, se analizaron las calificaciones otorgadas por los dos expertos mediante el coeficiente *Kappa de Cohen*. Los resultados obtenidos oscilaron entre 0. 309 ($k_{ALCANCE)}$ y 0. 211($k_{FLUIDEZ)}$ (véase Tablas VI, VII, VIII y IX).

Tabla VI. Resultados de la fiabilidad interevaluador en la escala de alcance

	Valor	Error tip.asint.	T.aprox.	Sig. aprox
Acuerdo entre evaluadores (K)	.309	.101	4.577	.000
Nº de casos válidos	30			

a. No asumiendo hipótesis nula

b. Utilizando error típico asintótico asumiendo hipótesis nula

Tabla VII. Resultados de la fiabilidad interevaluador en la escala de corrección

	Valor	Error tip.asint.	T.aprox.	Sig. aprox
Acuerdo entre evaluadores (K)	.327	.102	4.305	.000
Nº de casos válidos	30			

a. No asumiendo hipótesis nula

b. Utilizando error típico asintótico asumiendo hipótesis nula

Tabla VIII. Resultados de la fiabilidad interevaluador en la escala de fluidez

	Valor	Error tip.asint.	T.aprox.	Sig. aprox
Acuerdo entre evaluadores (K)	.211	.099	2.801	.000
Nº de casos válidos	30			

a. No asumiendo hipótesis nula

b. Utilizando error típico asintótico asumiendo hipótesis nula

Tabla IX. Resultados de la fiabilidad interevaluador en la escala de interacción

	Valor	Error tip.asint.	T.aprox.	Sig. aprox
Acuerdo entre evaluadores (K)	.295	.103	3.964	.000
Nº de casos válidos	30			

a. No asumiendo hipótesis nula

b. Utilizando error típico asintótico asumiendo hipótesis nula

A partir de estos resultados, y teniendo en cuenta las directrices de Landis y Koch (1977, citado en McHugh, 2012), se puede afirmar que el acuerdo entre los evaluadores es aceptable. Aunque estas cifras pueden parecer desalentadoras si se comparan con los resultados obtenidos por Fulcher (1997) o Dunsmuir et al. (2015) -lograron grados de fiabilidad interevaluador muy altos (p=0.87 y $k=0.62 < 0.88$, respectivamente)-, esta diferencia se podría atribuir al hecho de que los investigadores anteriores formaron a los evaluadores antes de analizar el grado de acuerdo entre los mismos. En este punto, se aclara que en esta investigación se coincide con Fulcher (2003) en que no se debe formar a los evaluadores antes de la validación de una prueba debido a que podría sesgar los resultados en términos de que los profesores podrían coincidir por la formación que han recibido y no por la calidad de los descriptores.

Fruto de lo anterior podríamos inferir que estos resultados son muy favorables y, junto con los resultados obtenidos por Bresciani, Oakleaf, Kolkhorst, Nebeker, Barlow, Duncan y Hickmott (2009), consideramos que pondrían en entredicho la aseveración que es necesario formar a los evaluadores para obtener grados de acuerdo razonables entre expertos (Weigle, 1998).

6. CONCLUSIONES

La presente investigación surge en el marco de la prueba oral de clasificación en ELE de la PUJ y se propone, como objetivos principales examinar la validez de contenido y la fiabilidad de la escala de evaluación oral. Además, se formularon los siguientes objetivos específicos: valorar la adecuación de los descriptores de la escala en lo concerniente a su formulación; evaluar el grado en que los descriptores de la escala reflejan el uso de la lengua que producen los candidatos; calcular la fiabilidad interevaluador y la consistencia interna de la escala.

Los resultados obtenidos tras los análisis cualitativos y cuantitativos efectuados en los estudios empíricos, permitieron llegar a una serie de conclusiones que se detallan a continuación.

En líneas generales, en lo que respecta a la validez y la fiabilidad, los resultados registrados en los análisis cuantitativos y cualitativos permitieron concluir que el contenido de la escala de evaluación de la prueba oral de clasificación de la PUJ es válido y fiable, y que los cambios realizados durante el proceso de validación fueron pertinentes.

En lo referente a la formulación de los descriptores de la escala de evaluación, las apreciaciones de los expertos pusieron en entredicho la adecuación de los mismos, fundamentalmente de los criterios de fluidez y pronunciación. Fruto de lo anterior, se enfatiza en la importancia de contar con docentes expertos para validar el contenido de pruebas lingüísticas. Se hace especial hincapié en que, en la validación de exámenes con fines de clasificación, se seleccionen profesores que imparten las clases donde se desea implementar este tipo de pruebas. Lo anterior debido a que estos son los que mejor conocen las características y necesidades lingüísticas y comunicativas de los estudiantes que se presentarán al examen y, por ende, son los más adecuados para determinar si el contenido de la prueba es representativo de los programas del centro y relevante para clasificar a los estudiantes en las clases que allí se imparten.

Con relación al grado en que los descriptores de la escala de evaluación reflejaban el uso de la lengua que producían los candidatos en la realización de las tareas, las observaciones realizadas en el primer pilotaje evidenciaron que el uso de la lengua que se describía en los descriptores de las escalas de pronunciación y fluidez no se correspondía con la lengua que emplearon los alumnos durante el pilotaje. Estos hallazgos confirmaron las valoraciones que habían realizado los expertos en el cuestionario y permitieron alcanzar dos conclusiones. Primero, se confirma la necesidad de pilotar la prueba antes de su administración y observar las actuaciones lingüísticas de los alumnos para validar el contenido de un determinado examen. Segundo, se constata la importancia de diseñar y validar una escala de evaluación combinando métodos intuitivos (juicio de expertos) con empíricos (pilotaje). Lo anterior, en el sentido que las actuaciones lingüísticas pueden corroborar o contrarrestar las valoraciones realizadas por los expertos. Tercero, ratificarían que los descriptores que se diseñan exclusivamente a partir de métodos intuitivos pueden acarrear problemas en su aplicación.

En relación a la consistencia interna, los resultados permitieron concluir que los ítems que conforman la nueva versión de la escala de evaluación miden el mismo constructo, esto es, la expresión e interacción orales definidas en los programas de ELE de la PUJ. Estos resultados son muy satisfactorios si se tiene en cuenta que la escala de evaluación se encuentra en su proceso de validación y que el estudio empírico realizado es un primer acercamiento a la fiabilidad. Estos hallazgos deberían enriquecerse analizando la consistencia interna una vez implementada la prueba o con futuros pilotajes.

Con respecto a la fiabilidad interevaluador, los análisis estadísticos indicaron que es es aceptable. Estos resultados son muy alentadores si se toma en consideración que la escala se encuentra en la fase de validación y que los calificadores no habían recibido formación previa. Por consiguiente, estos resultados permitirían concluir, junto con Bresciani et al. (2009), que es posible alcanzar grados de acuerdo aceptables entre los expertos sin formarlos previamente. De esta manera, se pondría en entredicho que es necesario formar previamente a los evaluadores (Weigle, 1998).

Al respecto de las listas de chequeo, los resultados constataron su idoneidad para recoger las observaciones realizadas al discurso producido por los alumnos. Por lo anterior se concluye, junto con O´Sullivan, Weir y Saville (2002), que estas herramientas se presentan como alternativa a la realización de análisis de discursos. Principalmente, se recomienda su uso en aquellos contextos donde no sea necesario caracterizar la lengua que producen los estudiantes y donde se puedan seguir recogiendo otras evidencias de validez. Sin embargo,

se recomienda que dos expertos apliquen esta herramienta para no tener que contrastar los resultados de la observación con las transcripciones de los alumnos.

De los párrafos precedentes se desprende la necesidad de que el proceso de validación se inicie antes de la administración de la prueba, concretamente con la validez de contenido, debido a la dificultad de diseñar escalas de evaluación que sean representativas y relevantes para el dominio lingüístico definido. En este punto, se subraya la pertinencia de la metodología propuesta, combinar las valoraciones de los expertos con las observaciones realizadas a las actuaciones lingüísticas de los estudiantes, con el fin de profundizar en el proceso de validación; fundamentalmente, para evitar descriptores ambiguos y generales que después no se puedan aplicar en la calificación de muestras reales de alumnos. Finalmente, de los hallazgos se desprende que los esfuerzos destinados para corroborar la validez de contenido de la prueba no han sido en vano puesto que se logró un grado de acuerdo moderado entre los expertos.

Sin embargo, estas conclusiones se deben interpretar con cautela considerando algunas limitaciones de los estudios. Entre estas se destacan la muestra limitada de estudiantes y evaluadores que participaron en los pilotajes. Por tanto, se considera oportuno realizar futuros pilotajes de la prueba y análisis cualitativos que analicen, por ejemplo, cómo están interpretando los evaluadores de los descriptores de la escala o si efectivamente es posible alcanzar grados de acuerdo razonables entre expertos sin que estos hayan recibido formación previa.

Finalmente, se advierte que en el proceso de validación se deben recoger diferentes tipos de evidencias que sustenten la adecuación de las interpretaciones y decisiones que tomamos a partir de los resultados de una prueba. Por tanto, se reconoce que estos estudios representan el inicio del proceso de validación de la prueba oral de clasificación de la PUJ pero que se debe seguir indagando al respecto. No obstante, "wihout a clear idea of match between intended content and actual content, any comprehensive investigation of the construct validity of a test is built on sand" (O´Sullivan, Weir y Saville, 2002:46). En otras palabras, se insiste en comenzar cualquier proceso de validación analizando la representatividad y relevancia del contenido de exámenes. De lo contrario, será como construir un castillo de arena.

REFERENCIAS

Alderson, Charles, J. (1990). Testing Reading Comprehension Skills (Part One). *Reading in a foreign language*, 6 (2), 425-438.

Alderson, Charles J.; Clapham, Carolina, y Wall, Dianne (1999). *Exámenes de idiomas: Elaboración y evaluación.* Barcelona, España: Edinumen S.L.

Association of Language Testers in Europe (2005). *Materials for the guidance of test item writers*. Disponible en http://www.alte.org/attachments/files/item_writer_guidelines.pdf

Bachman, Lyle F. (1990). *Fundamental considerations in language testing*. England, United Kingdom: Oxford University Press.

Bachman, Lyle F. (2002). Some reflections on task-based language performance assessment. *Language Tes*ting, 19(4), 453-479.

Bachman, Lyle F. (2005). Building and Supporting a Case for Test Use. *Language Assessment Quaterly*, 2(1), 1-34.

Bachman, Lyle F.; Davidson, Fred, y Milanovic, Michael (1996). The use of test method characteristics in the content analysis and design of EFL proficiency tests. *13th Annual Language Testing Research Colloquium*, 13 (2), 125-150.

Bachman, Lyle F. y Palmer, Adrian S. (2013). *Language Assessment in Practice.* Oxford, England: Oxford University Press.

Bordón, Teresa (2015). La evaluación de segundas lenguas (L2). Balance y perspectivas. *Revista Internacional de Lenguas Extranjeras*, 4, 9-30.

Bresciani, Marilee J.; Oakleaf, Megan; Kolkhorst, Fred; Nebeker, Camille; Barlow, Jessica; Duncan, Kristian, y Hickmott, Jessica (2009). Examining Design and Inter-Rater Reliability of a Rubric Measuring Research Quality across Multiple Disciplines. *Practical Assessment, Research&Evaluation*, 14(12), 2-7.

Campillos Llanos, L. (2012). La expresión oral en español lengua extranjera: interlengua y análisis de errores basado en corpus. Tesis Doctoral. Madrid, España: Universidad Autónoma de Madrid.

Consejo de Europa (2002). *Marco Europeo de Referencias para las lenguas enseñanza, aprendizaje, evaluación.* Madrid, España: MEC y Anaya.

Cumming, Alister; Grant, Leslie; Mulcahy-Ernt, Patricia, y Powers, Donald E. (2004). A teacher-verification study of speaking and writing prototype tasks for a new TOEFL. *Language Testing*, 21(2), 107-145.

Del Moral, Franciso (2015). El proceso de validación de una escala de descriptores para la evaluación de la expresión e interacción orales de ELE y su influencia en la fiabilidad de la prueba. Tesis Doctoral en Lingüística Aplicada a la Enseñanza de Español Lengua Extranjera. Madrid, España: Universidad Antonio de Nebrija.

D´ Este, Claudia (2012). New views of validity in language testing. EL.LE, 1(1), 61-76.

Deygers, Bart y Van Gorp, Koen (2015). Determining the scoring validity of a co-constructed CEFR-based rating scale. *Language Testing*, 32(4), 521-541.

Dunsmuir, Sandra; Kyriacou, Maria; Batuwitage, Su; Hinson, Emily; Ingram, Victoria, y O'Sullivan, Siobhan (2015). An evaluation of the Writing Assessment Measure (WAM) for children´s narrative writing. *Assessing Writing*, 23, 1-18.

Fitzpatrick, Tess y Clenton, Jon (2010). The challenge of validation: Assessing the performance of a test of productive vocabulary. *Language Testing*, 27(4), 537-554.

Fulcher, Glenn (1997). An English language placement test: issues in reliability and validity. *Language Testing*, 14(2), 113-139.

Fulcher, Glenn (2003). *Testing second language speaking*. Harlow, England: Pearson/Longman.

González Alonso, Jorge A. y Pazmiño, Mauro (2015). Cálculo e interpretación de Alfa de Cronbach para el caso de validación de la consistencia interna de un cuestionario con dos escalas tipo Likert. *Revista Publicando*, 2(1), 62-77.

Gwet, Kilem Li (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters.* Gaithersburg, United States: Advanced Analytics.

Horche, Raquel y Marco, Miren J. (2008). El concepto de fluidez en la expresión oral. Disponible en http://cvc.cervantes.es/ensenanza/biblioteca_ele/publicaciones_centros/PDF/rio_2008/37_horche-marco.pdf.

Kane, Michael T. (1992). An Argument-based Approach to Validation. *Psychological bulletin*, 112(3), 527–535.

Lado, Robert (1964). *Language teaching a scientific approach*. New York, United States: McGraw-Hill.

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

Llisterri, Joaquim (2003). La evaluación de la pronunciación en la enseñanza del español como segunda lengua. *Perspectivas teóricas y metodológicas: Lengua de acogida, educación intercultural y contextos inclusivos, 547-561*.

Martín Alonso, Begoña (2017a). Validación del contenido de una prueba oral de clasificación. *Revista Nebrija de Lingüística Aplicada*, 22.

Martín Alonso, Begoña (2017b). Diseño y validación de una prueba oral de clasificación en ELE. Tesis Doctoral en Lingüística Aplicada a la Enseñanza de Español Lengua Extranjera. Madrid, España: Universidad Antonio de Nebrija.

Mendoza Ramos, Arturo (2015). La validez en los exámenes de alto impacto. Un enfoque desde la lógica argumentativa. *Perfiles Educativos*, 37(149), 169-186.

McHugh, Mary L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3), 276-282.

Newman, Isadore; Lim, Janine, y Pineda, Fernanda (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(3), 243-260.

O´Sullivan, Barry; Weir, Cyril J., y Saville, Nick (2002). Using observation checklists to validate speaking-test tasks. *Language Testing*, 19(1), 33-56.

Robles, Pilar y Rojas, Manuela del Carmen (2015). La validación por juicio de expertos: dos investigaciones cualitativas en Lingüística aplicada. *Revista Nebrija de Lingüística Aplicada*, 18.

Sireci, Stephen G. y Faulkner-Bond, Molly (2014). Validity evidence based on test content. *Psicothema*, 26(1), 100-107.

Van der Walt, Johann y Steyn, H.S. (2008). The validation of language tests. *Stellenbosch Papers in Linguistics*, 38, 191-204.

Wall, Dianne; Clapham, Caroline, y Alderson, Charles. (1994). Evaluating a placement test. *Language Testing*, 11(3), 321-344.

Wang, Ping (2009). The inter-reliability in scoring composition. *English Language Testing*, *2*(3), 39-43.

Weigle, Sara C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15(2), 263-287.