

Competencia comunicativa oral en español en contextos de migración: diseño y validación de un examen multinivel¹

Spanish oral communicative competence in migration contexts: design and validation of a multilevel exam

Clara Ureña Tormo

UNIVERSITAT POLITÈCNICA DE VALÈNCIA
ESPAÑA
curetor@idm.upv.es
<https://orcid.org/0000-0002-7540-2931>

María Eugenia Flores

IPEHCS-CONICET
mariaeugeniafloresmeyer@gmail.com
<https://orcid.org/0000-0001-9237-6352>

Susana Martín Leralta

UNIVERSIDAD NEBRIJA
ESPAÑA
smartin@nebrija.es
<https://orcid.org/0000-0002-1816-1967>

Recibido: 6-4-2022 / **Aceptado:** 12-6-2023

DOI: 10.4151/S0718-0934202401150932

Resumen

La evaluación es una parte fundamental del proceso de enseñanza-aprendizaje de una lengua extranjera o lengua adicional. En el ámbito del español como lengua de migración, existe la necesidad de contar con una prueba que, ajustándose a las particularidades de los aprendices migrantes y refugiados, permita evaluar su nivel de competencia comunicativa oral en español. Este artículo describe el diseño y la validación empírica de un examen multinivel que se ha implementado con 30 informantes y en el que han participado 40 especialistas en enseñanza de segundas lenguas y evaluación de competencias orales. Tras el proceso de validación, los resultados reflejan que tanto la prueba como la escala de evaluación son instrumentos válidos y fiables, al aportar valores estadísticamente significativos. Esta prueba, ya validada con anterioridad, prevé estar accesible para profesores y voluntarios que enseñan español en contextos de migración y refugio.

Palabras clave: destrezas orales, competencia comunicativa, español como lengua de migrantes y refugiados, evaluación, examen de nivel

Abstract

Assessment is a key part in the process of teaching and learning a foreign or additional language. In the field of Spanish as a language of migration, there is a need for a level test which fits the specific needs of migrant and refugee learners and allows teachers to assess their oral skills in Spanish. This paper describes the design and empirical validation of a multilevel test in which 30 participants and 40 experts in foreign language assessment have participated. Results reveal that both the test and the rubric

are valid and reliable assessment tools as shown by statistically significant data. This test is expected to be available for teachers and volunteers who teach Spanish to migrants and refugees.

Keywords: oral skills, communicative competence, Spanish language of migrants and refugees, assessment, level test

INTRODUCCIÓN

La evaluación es una parte fundamental del proceso de enseñanza-aprendizaje de una lengua adicional². En el ámbito específico de la enseñanza del español (ELE/L2), existen numerosas publicaciones sobre la evaluación de la lengua en contextos generales (Antón, 2019; Bordón, 2015; Bordón & Liskin-Gasparro, 2007; Pastor Cestero & Roca Marín, 2007). No obstante, el presente artículo se enmarca en las investigaciones que, con un carácter más específico, se centran en la evaluación de la competencia comunicativa oral en español de la población migrante y refugiada (Baralo, 2012; Baralo & Martín Leralta, 2011; Baralo et al., 2016; Villalba, 2012).

A pesar de los avances, en la actualidad se sigue sin disponer de una herramienta de evaluación que, reuniendo los criterios de validez, fiabilidad y viabilidad, se adapte a la realidad lingüística y de adquisición de la lengua de los migrantes y refugiados, y, en definitiva, permita determinar su capacidad para comunicarse oralmente en español. Con el objetivo de dar respuesta a esta necesidad, en este artículo se presenta el proceso de diseño y validación empírica del examen multinivel Prueba Oral de Español Multinivel para Migrantes (POEM).

La adquisición de una lengua adicional en aprendientes migrantes y refugiados conlleva la activación de los mismos mecanismos y procesos cognitivos que en cualquier otro perfil de estudiantes (Villalba, 2018). Asimismo, los procedimientos de evaluación utilizados con este tipo de alumnado no deben ser diferentes a los adoptados con otros aprendientes en cuanto a los aspectos técnicos. Sin embargo, es fundamental tener en cuenta una serie de particularidades a la hora de planificar su enseñanza-aprendizaje y, más concretamente, de cara a diseñar un examen de nivel. Villalba (2012) señala la necesidad de prestar atención a cuatro elementos clave: el distinto nivel de competencia que poseen en cada una de las destrezas comunicativas, su grado de dominio de lectoescritura, sus conocimientos generales y académicos, y sus necesidades comunicativas. Estos aspectos se tienen en cuenta en el diseño de la prueba que aquí se presenta, tal como se indica a continuación.

En cuanto al nivel de dominio de las distintas destrezas y el conocimiento de la lectoescritura, es frecuente que, en la adquisición de las lenguas de migración, el desarrollo de las destrezas comunicativas sea irregular y que no todas ellas se adquieran con el mismo nivel de dominio (Baralo, 2012). Uno de los factores que contribuye a esta circunstancia es que buena parte del grupo meta está compuesto por

neolectores en español, de aquí se desprende el primer motivo para desarrollar un examen centrado en la competencia oral.

En relación con los conocimientos generales y académicos, el informe emitido por la Association of Language Testing in Europe y el Council of Europe (2018) registra un bajo nivel de escolarización reglada en el alumnado migrante. Las dificultades de acceso a la educación reglada justifican, por tanto, que sea imprescindible relegar a un segundo plano las habilidades de la expresión y la comprensión escritas para centrarse exclusivamente en las destrezas orales. Por su parte, las necesidades comunicativas del alumnado determinan el contenido de las pruebas, de ahí que convenga incluir actividades que reflejen situaciones reales o verosímiles de la vida de los estudiantes con el fin de lograr un aprendizaje significativo.

Teniendo en cuenta los elementos señalados, el examen POEM evalúa exclusivamente las competencias de la expresión e interacción orales, las cuales, sobre todo la segunda, llevan implícito un dominio de comprensión oral. Además, la prueba incluye tareas comunicativas sobre situaciones relacionadas con el trabajo y la vivienda que pueden suceder en la vida real de los estudiantes migrantes y que, por consiguiente, responden a sus necesidades comunicativas.

Cabe puntualizar que el objetivo de esta prueba no es reemplazar pruebas existentes, sino proporcionar una nueva herramienta fiable, eficaz y específica para determinar el nivel comunicativo oral en español de personas en contextos de migración y refugio que no pueden acceder a otro tipo de certificaciones por razones de distinta naturaleza. En concreto, en lo referente a la parte socioeconómica, este examen se plantea de acceso abierto para cubrir las necesidades de voluntarios y profesores de español a inmigrantes en contextos reales como ONG y asociaciones sociales. Además, siguiendo la recomendación de Villalba (2012), se trata de un examen gratuito que se desarrolla al margen de criterios políticos y otros aspectos relacionados con la obtención de la nacionalidad y la ciudadanía.

El artículo se estructura de manera que, primero, exponemos un breve repaso de un conjunto de exámenes de certificación lingüística y de nivel existentes en el marco de la enseñanza del español en contextos generales y específicos de la migración (apartado 1). Seguidamente, describimos las tres fases que componen el proceso de elaboración de la prueba y su rúbrica de evaluación, correspondientes al diseño (apartado 2), la validación (apartado 3) y la implementación (apartado 4). En cada apartado, se especifican los aspectos metodológicos y los resultados propios de cada fase. Por último, terminamos con las conclusiones principales sobre el trabajo realizado y el listado de referencias bibliográficas consultadas.

1. Exámenes de certificación lingüística y exámenes de nivel

Como parte del proceso de enseñanza-aprendizaje de una lengua, en lo referente a los instrumentos de evaluación, se distinguen, por una parte, los exámenes de certificación lingüística, responsables de acreditar el grado de competencia que los estudiantes tienen del idioma extranjero mediante una certificación lingüística correspondiente a un determinado nivel; y, por otra parte, los exámenes de nivel o — en términos de Bordón (2015)— exámenes de clasificación, dirigidos a identificar su grado de conocimiento de la lengua para, precisamente, poder clasificarlos en grupos en función de su nivel de dominio.

En el campo de la enseñanza del español como lengua de migración y refugio se cuenta con exámenes de certificación lingüística, entre los que se destacan la Prueba de Eficacia Comunicativa Oral en Lengua Española (PECOLE) (Martín Leralta, 2013) y el examen LETRA (Lengua Española para Trabajadores Inmigrantes), que reemplaza al anterior DILE (Diploma Inicial de Lengua Española para Trabajadores Inmigrantes), promovidos por la Consejería de Inmigración y Cooperación de la Comunidad de Madrid y la Fundación Antonio de Nebrija (Baralo, 2012). Estos diplomas, antecedentes directos del examen POEM, nacieron: “ante el requerimiento de disponer de una herramienta que facilite la integración lingüística de los inmigrantes de la Comunidad de Madrid en el ámbito profesional o en vías de inserción laboral” (Mavrou & Martín Leralta, 2018, p. 225).

Además, se cuenta con los exámenes del Instituto Cervantes para conseguir la nacionalidad española, que comprenden la obtención del diploma de español DELE de nivel A2 o superior y la prueba de conocimientos constitucionales y socioculturales de España (CCSE).

En relación con los exámenes de nivel, cabe destacar que el procedimiento de nivelar a los estudiantes es absolutamente necesario para organizar los cursos de lengua y ofrecer una atención personalizada a cada estudiante en función de los requerimientos que exige su nivel de dominio de la lengua. En el ámbito de la enseñanza del español general, no ciñéndose de manera específica a contextos de migración y refugio, el Servicio Internacional de Evaluación de la Lengua Española (SIELE) administra un examen multinivel que certifica el grado de dominio de español a través de cuatro pruebas: comprensión lectora, comprensión auditiva, expresión e interacción escritas y expresión e interacción orales. Los candidatos reciben la puntuación obtenida y su equivalencia con los niveles del *Marco Común Europeo de Referencia para las Lenguas* (MCER). El candidato puede realizar el examen completo y examinarse de las cuatro pruebas mencionadas o, alternativamente, tienen la posibilidad de seleccionar una modalidad independiente que comprende solo una o dos de las cuatro pruebas.

Además del SIELE, actualmente también existen varios modelos de examen de nivel disponibles en línea con acceso abierto. Entre ellos, la prueba de nivel de la aplicación AVE Global del Instituto Cervantes que se compone de tres partes con preguntas sobre vocabulario y estructuras gramaticales, una tarea de comprensión lectora y una de comprensión auditiva. Cuando el estudiante termina la prueba, no recibe información sobre el número de respuestas correctas e incorrectas ni de la puntuación obtenida, sino que automáticamente recibe información sobre su nivel de español según el MCER.

También se encuentra disponible en línea la prueba de nivel de español de Oxford House, formada por 50 preguntas que evalúan exclusivamente contenidos de tipo gramatical. Al final de la prueba, se muestra el número de respuestas correctas y el nivel del MCER al que equivale esta puntuación. Por último, la prueba de español de Cervantes se compone de 65 cuestiones sobre contenidos léxicos y gramaticales que se distribuyen en seis bloques correspondiente a los seis niveles del MCER: A1, A2, B1, B2, C1 y C2. En estos tres últimos exámenes de nivel, las preguntas se organizan por orden de dificultad ascendente y pueden realizarse de manera gratuita y autónoma a través de internet.

En los párrafos anteriores, se ha comprobado que tenemos a nuestra disposición exámenes de certificación específicos para inmigrantes como el PECOLE y el LETRA, así como de exámenes de nivel de español general, entre ellos, el SIELE y los exámenes de AVE Global, Oxford House y la prueba de español de Cervantes. A pesar de la existencia de estas pruebas, identificamos ciertos vacíos que justifican la necesidad de disponer de una nueva que permita nivelar el grado de competencia comunicativa oral en español y que esté dirigida específicamente a migrantes y refugiados.

Concretamente, los diplomas PECOLE y LETRA permiten certificar únicamente un grado de dominio A1 y A2 en español, y su ámbito de aplicación es diferente al del POEM. En el caso de la PECOLE, se trata de una prueba para comprobar la eficacia comunicativa oral de nivel inicial (A1) de los inmigrantes inmersos en procesos relacionados con trámites de vivienda y familia en la Comunidad de Madrid. El LETRA, por su parte, es un examen de certificación de nivel inicial para inmigrantes en edad laboral, orientado a medir su competencia comunicativa en las cuatro destrezas de uso de la lengua en el ámbito profesional. Su última convocatoria tuvo lugar en 2016 y no se conoce fecha aún de la próxima.

Por otra parte, los exámenes de nivel de español general no contemplan de manera específica las motivaciones y las necesidades de la población migrante (personas en proceso de alfabetización, coste económico de los exámenes, etc.) y, en consecuencia, su aplicación a este contexto resulta limitada. Además, el SIELE tiene un coste económico que puede suponer un impedimento de acceso para determinada

población migrante y refugiada. Asimismo, los exámenes de AVE Global, Oxford House y la prueba de español de Cervantes conceden una importancia destacada a los conocimientos gramaticales y se centran sobre todo en el plano escrito, dos aspectos que no reflejan la realidad de aprendizaje de los estudiantes migrantes y refugiados.

El escenario expuesto justifica la necesidad de cuestionarse la eficacia y la funcionalidad de utilizar las herramientas de evaluación mencionadas y, a partir de ahí, se plantea el examen de clasificación multinivel POEM, cuyo proceso de elaboración y validación se describe a continuación. Este proceso comprende tres fases: el diseño del examen y de la escala de evaluación (ver 2.1 y 2.2); la validación del examen y de la escala de manera independiente (ver 3.1 y 3.2); y la implementación del examen y la escala conjuntamente (ver 4.1 y 4.2).

2. Fase de diseño

2.1 *Diseño de la prueba*³

En la elaboración de la prueba han participado las autoras del artículo, en colaboración con un comité de especialistas en diseño de materiales y enseñanza de ELE, adquisición de español como lengua de migración y evaluación de destrezas orales. Para diseñar las tareas que incluye el examen, se ha tenido en cuenta la propuesta de Bordón (2007) sobre los dos tipos de ejercicios más representativos para evaluar la interacción oral: la entrevista y la conversación. Asimismo, hemos seguido las indicaciones de Luoma (2004) respecto a los exámenes orales individuales para cada candidato [*one to one tests*], en los que puede plantearse una entrevista guiada y estructurada, la cual funciona como un calentamiento inicial que da paso a la interacción principal, en la que pueden plantearse diversas tareas (p. ej., describir una imagen, hablar sobre un tema seleccionado anteriormente, participar en un juego de rol, etc.).

A partir de los planteamientos descritos, la prueba POEM se compone de dos tareas: la primera consiste en una entrevista formada por seis bloques de preguntas y la segunda propone el desarrollo de una conversación sobre un tema elegido por el candidato en la que debe representar un papel determinado. Por su condición de examen multinivel, las tareas del POEM pueden realizarse, al menos parcialmente, por estudiantes de español de todos los niveles (A1-C2), teniendo en cuenta que el modo en que cada candidato desarrolla la tarea variará, entre otros aspectos, según su dominio del idioma. El examen tiene una duración de 30 minutos, aproximadamente, aunque la actuación de cada candidato determina la duración exacta en cada caso.

Más concretamente, la entrevista personal de la primera tarea aborda los siguientes bloques temáticos: 1. Información personal; 2. El día a día; 3. Relaciones sociales; 4. El trabajo; 5. Ciudades y costumbres; 6. El día que llegaste a Madrid. Estos seis bloques de contenidos se han seleccionado teniendo en cuenta los temas más comunes en las

programaciones de los cursos de español, en exámenes previos consultados (ver 3.1) y en las nociones generales y específicas del *Plan Curricular del Instituto Cervantes* (Instituto Cervantes, 2006). El objetivo de esta tarea es que el candidato responda a las preguntas que se le plantean en cada bloque, utilizando los contenidos y las estructuras lingüísticas que conoce de acuerdo con su nivel de español. Por ejemplo, a la pregunta del bloque 3. Relaciones personales: “Por lo general, ¿te gusta relacionarte con la gente o prefieres estar solo?”, un candidato con un nivel inicial puede responder indicando únicamente su preferencia: “Estar con gente” o “Estar solo”. Por el contrario, se espera que un candidato de nivel avanzado sea capaz de justificar su respuesta, aportando información sobre sus preferencias en las relaciones sociales, por ejemplo: “Prefiero relacionarme con amigos y compañeros porque soy una persona muy sociable con una vida social activa”.

En la segunda tarea, se presentan dos láminas con imágenes, de las cuales el candidato debe elegir una para mantener una conversación sobre un tema relacionado con ella. Una vez elegida la lámina, el evaluador plantea una situación contextualizada sobre el tema seleccionado: la vivienda o el trabajo. Una vez que se le ha expuesto, el examinador explica que el objetivo de la tarea es que el candidato haga preguntas para obtener más información sobre el tema, dar su opinión al respecto, ofrecer consejos y llegar a un acuerdo con el evaluador sobre la mejor solución al problema. El candidato dispone de dos minutos para preparar sus aportaciones. Pasado el tiempo de preparación, este debe interactuar con el evaluador, procurando que el intercambio comunicativo progrese mediante la interacción continuada.

Las dos tareas van acompañadas de instrucciones específicas para el examinador, así como ejemplos de posibles intervenciones. Principalmente, las indicaciones orientan al examinador sobre cómo debe adaptar su actuación en función del nivel del candidato, sin desviarse, en ningún caso, de la estructura de la prueba. Por ejemplo, en la tarea 1, si el candidato se muestra incapaz de responder alguna pregunta, el evaluador debe repetir o reformular la cuestión para facilitar su comprensión. Si, aun proporcionando esta ayuda, el candidato no responde, el examinador pasa a la siguiente pregunta. En el caso contrario, si el candidato se desenvuelve cómodamente respondiendo las preguntas, el evaluador debe fomentar que el candidato amplíe sus intervenciones para obtener más información sobre su nivel de competencia en español.

En la tarea 2, si el nivel del candidato manifiesta un nivel inicial, las instrucciones deben ser muy concisas y sus intervenciones deben limitarse a dar consejos y aportar su opinión sobre la situación. Exactamente, una vez que han pasado los dos minutos de preparación, el examinador lee el planteamiento de la situación y de manera directa pregunta al candidato: “¿Qué consejos me darías?” A continuación, el examinador puede plantear otras posibles soluciones a la cuestión planteada y directamente preguntar al candidato: “¿Qué opinas?, ¿te parece una buena opción?”.

Además, las instrucciones subrayan la importancia de que, durante toda la prueba, el examinador adapte el contenido de sus intervenciones y el ritmo de habla al nivel de competencia lingüística que muestre el candidato. En este sentido, es fundamental tener en cuenta que obtener muestras de lengua adecuadas que permitan evaluar el nivel de los candidatos no depende solo del buen diseño del examen, sino de la actuación eficaz de un examinador formado y competente que es capaz de extraer muestras de lengua oral susceptibles de ser evaluadas (Bordón, 2007). Es de suma importancia, así pues, contar con un evaluador experimentado, bien entrenado, que posea las habilidades necesarias para conseguir las mejores muestras, al margen de las características del examen, la personalidad y el nivel de los candidatos. Como apunta Luoma (2004) a partir de estudios precedentes, el hecho de que el entrevistador tenga un considerable poder sobre los examinandos cuando la prueba evaluadora es una entrevista, constituye una de las principales debilidades de este formato de examen. No obstante, como indica la autora, su ventaja radica en la posibilidad de ajustarse al nivel de competencia de cada candidato y a incrementar el grado de control sobre lo que sucede en la interacción, cualidades que creemos propicias para los fines del presente estudio, a saber, poder medir de manera válida y fiable los diferentes niveles de dominio lingüístico.

2.2 Diseño de la escala de evaluación⁴

La fase de elaboración comprende, además del diseño de la prueba, el diseño de una escala de evaluación pensada específicamente para este examen. Tal como indica Luoma (2004), el foco en la confección de exámenes orales se pone a menudo en las tareas, aunque la validez de las puntuaciones depende en igual medida de los criterios de puntuación y de la relación entre estos y las tareas. Por lo tanto, teniendo en cuenta que es necesario asegurarse de que el instrumento de evaluación que se elige es el adecuado para una determinada prueba (Council of Europe & Language Policy Division, 2009), se ha optado por utilizar una escala, de manera análoga a como se procede en la evaluación de las destrezas productivas en los exámenes de certificación lingüística. De acuerdo con Baralo (2012), el uso de estas escalas y, en concreto, de las especificaciones que incluyen los descriptores se vuelve especialmente relevante en este tipo de exámenes en que: “la calificación depende en parte del propio evaluador, que actúa de manera subjetiva frente al candidato y a las variadas condiciones y factores individuales y contextuales que pueden influir” (p. 28).

En el diseño de la escala seguimos el enfoque propuesto por Bachman y Palmer (2010) para el desarrollo de escalas analíticas de habilidad lingüística. Estas se elaboran en función de la manera en la que se define el constructo que va a ser medido y se referencian a un criterio a partir del cual se establecen los distintos niveles de habilidad lingüística, donde el más bajo corresponde a la ausencia de evidencia y el más alto, a la evidencia de maestría en la habilidad. En nuestra escala, el constructo que medimos se identifica con el fundamento teórico acerca de la noción ‘expresión oral’ y del tipo de

pruebas para evaluarla. En concreto, POEM se trata de una prueba oral individual en la que necesitamos medir la capacidad comunicativa examinando varias dimensiones que comprenden, por ejemplo, la coherencia y la cohesión discursivas, la corrección gramatical, la amplitud léxica, la pronunciación, etc. Por su parte, el criterio designa los distintos niveles de dominio que el alumnado puede mostrar en relación con las dimensiones anteriores, de forma que se determinan nueve posibles niveles para cada dimensión.

Además de lo anterior, también nos hemos basado en las escalas de evaluación del MCER correspondientes a las pruebas de producción e interacción orales de los seis niveles (Council of Europe, 2018). Igualmente, se han consultado las escalas de los exámenes DELE de los seis niveles y se han tenido en cuenta los trabajos de Bordón (2000, 2007), las especificaciones del examen LETRA (Baralo & Estaire, 2011) y las programaciones de cursos de ELE a inmigrantes.

El resultado es una escala que contempla cinco aspectos específicos de la competencia comunicativa del estudiante, además de su actuación global. Se valoran, así pues, seis criterios. Primero, se evalúa la pertinencia comunicativa, que mide el grado en que el candidato responde de forma pertinente con la información requerida. En segundo lugar, el grado de coherencia y cohesión en la organización del discurso y la estructuración de las ideas que contiene sus intervenciones. En tercer lugar, el grado de fluidez en la exposición de las ideas y la pronunciación, entonación y el acento. A continuación, se mide, por una parte, el nivel de corrección gramatical y, por otra parte, el alcance léxico. Por último, se descarta la posibilidad de utilizar una escala holística y, en su lugar, se opta por incluir un criterio que evalúe la actuación global del candidato.

Cada criterio se puntúa a partir de 9 descriptores que indican el grado de dominio que el candidato muestra del criterio en cuestión, siendo 1 el grado de dominio más bajo y 9 el más alto. La puntuación mínima que puede obtener un candidato en el examen es de 6 y la puntuación máxima es de 54.

2.2.1 Equivalencia entre la puntuación del examen y los niveles del MCER

La escala de evaluación que se ha diseñado, a diferencia de las escalas que se utilizan en otros exámenes como el DELE, opta por asignarle a cada banda puntos, en lugar de niveles. Esta decisión se ha tomado con el objetivo de evitar que la percepción del evaluador sobre la actuación del candidato se vea influenciada por el hecho de que cada descriptor lleve asignado de antemano un nivel. Se trata de una medida encaminada a evitar la posible inconsistencia producida por diferentes interpretaciones de la escala de evaluación o del estándar de rigor (Bachman & Palmer, 2010). Además, tal como señala Bordón (2007), “la presencia de un evaluador siempre implica un cierto grado de subjetividad” (p. 102), por eso, es imprescindible contar con herramientas que permitan asignar una calificación lo más objetiva posible.

En cualquier caso, una vez que el candidato ha terminado el examen, es necesario establecer la equivalencia correspondiente entre los puntos que ha obtenido y su nivel de competencia en español. En este sentido, el MCER constituye una de las herramientas más utilizadas como sistema de estándares de referencia para comparar certificaciones. Siguiendo a Llorián González (2015), el procedimiento para relacionar los resultados de las pruebas con los descriptores de las escalas del MCER es un proceso complejo que necesita apoyarse sobre unas bases metodológicas sólidas. En nuestro caso, nos hemos basado en las indicaciones del manual *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (Council of Europe & Language Policy Division, 2009), en el que se aborda este tema específicamente.

Así, para comenzar, hemos establecido la nota de corte (*cut-off score*) que marca el límite entre dos niveles de una escala continua, a partir del proceso de *standard setting* descrito en la obra citada. En concreto, hemos determinado cinco notas de corte correspondientes a los niveles A1-A2; A2-B1; B1-B2; B2-C1; C1-C2. Tras ello, para llevar a cabo el proceso de equivalencia propiamente dicho, hemos utilizado el método *Item-descriptor Matching Method* y lo hemos adaptado a las especificaciones de nuestro examen siguiendo un proceso compuesto por dos fases. En la primera, en lugar de asignar un nivel a cada ítem o pregunta, tal y como sugiere el método, hemos asignado un nivel a cada banda de descriptores. Esto es, hemos establecido una correspondencia entre los descriptores de las nueve bandas y los seis niveles del MCER, como muestra la Tabla 1.

Tabla 1. Correspondencia entre las bandas de descriptores y los niveles del MCER

Descriptores de las bandas	Niveles MCER
1	A1
2	
3	A2
4	
5	B1
6	
7	B2
8	C1
9	C2

La correspondencia entre los descriptores y los niveles del MCER no es unívoca, puesto que esta escala incluye tres bandas más que niveles hay en el MCER. Así, los descriptores de las bandas puntuadas con 1 y 2 corresponden al nivel A1; los de las bandas 3 y 4 al nivel A2; los de las bandas 5 y 6 al nivel B1; los de la banda 7 al nivel B2; la banda 8 al nivel C1; y la banda 9 al nivel C2.

Frecuentemente, la actuación de un candidato no es uniforme y, por lo tanto, no recibe la misma puntuación en todos los criterios que se evalúan. Por ejemplo, es

posible que el candidato se sitúe en la banda 3 en cuanto a la fluidez, mientras que reciba un 5 en corrección gramatical y alcance léxico, y un 4 en pertinencia, coherencia e impresión global. En ese caso, fijar la correspondencia entre la puntuación obtenida en el examen y el nivel en el MCER se vuelve más compleja. Para facilitar esta tarea, partiendo de las decisiones tomadas en la primera fase (esto es, la correspondencia entre las 9 bandas de descriptores y los 6 niveles del MCER), la segunda fase del proceso ha consistido en establecer seis franjas de puntos, delimitadas por un límite inferior y un límite superior de puntos para cada nivel. De acuerdo con el manual citado anteriormente, “determining the cut scores or *setting the (performance) standards* is usually a group decision. The group that makes such decisions is normally called a *panel*” (Council of Europe & Language Policy Division, 2009, p. 58). En nuestro caso, el punto de corte (*cut-off score*) para cada nivel se han establecido a partir de la decisión consensuada de las autoras de este estudio, que configuran el denominado *panel*, y ha sido verificada por un comité de especialistas mediante un cuestionario de validación de la escala de evaluación (ver 3.2, 3.2.1, 3.2.2). La siguiente tabla muestra los resultados de esta decisión.

Tabla 2. Correspondencia entre las puntuaciones del examen y los niveles del MCER

Puntos obtenidos	Niveles MCER
6 - 16 puntos	A1
17 - 27 puntos	A2
28 - 37 puntos	B1
38 - 45 puntos	B2
46- 51 puntos	C1
52 - 54 puntos	C2

Para ilustrar lo expuesto: las bandas 5 y 6 se corresponden con el nivel B1, de manera que un candidato que se sitúa en estas bandas puede obtener una puntuación total que oscila entre los 30 puntos (seis criterios puntuados con 5 puntos) y los 36 puntos (seis criterios puntuados con 6 puntos). Para determinar la nota de corte de cada nivel, se han ampliado ligeramente estas puntuaciones en su límite inferior y en su límite superior, de forma que se dan cabida a todos los niveles. En el caso del nivel B1, el límite inferior se sitúa en los 28 puntos y el límite superior, en los 37. Así, un candidato con dos descriptores en la banda del 4 (8 puntos) y cuatro en la banda del 5 (20 puntos) se situaría en el límite inferior del B1 (28 puntos); mientras que un candidato con cinco descriptores en la banda del 6 (30 puntos) y uno en la banda del 7 (7 puntos) se situaría en el límite superior del nivel B1 (37 puntos). Por su parte, un candidato que recibe tres puntuaciones en la banda del 5 (15 puntos) y tres en la banda del 6 (18 puntos) se situaría en el límite intermedio (33 puntos). Se observa, así pues, que, para establecer los límites inferior y superior de un nivel, el B1 en este caso, se comprende la posible asignación de algún criterio en una banda correspondiente a un

nivel inferior (la banda 4 = nivel A2) o a un nivel superior (la banda 7 = nivel C1) del nivel que se asigna finalmente: el B1.

3. Fase de validación

3.1 Validación de la prueba

El proceso de validación se aplica, por una parte, a la prueba y, por otra parte, a la escala de evaluación. En relación con la validación de la prueba, se ha verificado la ‘validez de contenido’ con el fin de garantizar que el examen incluye el espectro de contenidos que se consideran representativos de la competencia oral de los candidatos en todos los niveles. La principal dificultad a la hora de validar esta dimensión viene dada por el hecho de no contar con un currículo específico para la enseñanza del español a migrantes, cuya inexistencia en la actualidad constituye, tal como señala Villalba (2018), “una de las mayores carencias que se plantean en la enseñanza del español L2 a inmigrantes” (p. 469). Ante esta situación se han tomado las siguientes decisiones: por una parte, se han tenido en cuenta las directrices generales recogidas en el MCER y el PCIC para todos los niveles de enseñanza, la primera por tratarse de la obra de referencia para la enseñanza de lenguas y la segunda por centrarse en la enseñanza del español específicamente. Por otra parte, con el objetivo de diseñar un material que refleje la realidad de los migrantes y refugiados tal como se estudia en Baralo (2012), ha sido fundamental consultar tanto la guía del profesor del examen LETRA (Baralo & Martín Leralta, 2010), diseñado exclusivamente para el contexto migrante, como las programaciones de cursos facilitadas por las ONG que mantienen convenio con la Cátedra de investigación en la que se enmarca este estudio.

De esta manera, para decidir el contenido de las tareas que componen el examen ‘validez de contenido’, se han definido las principales funciones comunicativas que deciden trabajarse y se han seleccionado los exponentes lingüísticos necesarios para su consecución, los cuales se corresponden fundamentalmente con contenidos gramaticales y léxicos. Recuérdese que se trata de un examen multinivel que, por sus características, permite evaluar todos los niveles de dominio de la lengua.

También se ha comprobado la ‘validez de constructo’, es decir, el grado en que las preguntas del examen permiten obtener la información necesaria para establecer el nivel de español de los candidatos. Para diseñar las dos tareas del examen POEM (una entrevista y una conversación), se ha tenido en cuenta las características de las tareas orales incluidas en exámenes de certificación lingüística en español, como el DELE, el CELU y el LETRA, pero también en exámenes de referencia para la certificación lingüística en inglés, concretamente, los exámenes Cambridge y Pearson. En este sentido, Luoma (2004) señala que la validez de las puntuaciones de una prueba oral se fundamenta en el propósito que las puntuaciones intentan satisfacer y, por tanto, debe examinar el uso de la lengua en función del objetivo que se persigue, ya sea certificar la habilidad de un profesional para el desempeño de su profesión en la nueva lengua, ya

sea ofrecer *feedback* al alumnado tras una actividad oral para ayudarlos a seguir avanzando. El proceso de validación de la prueba se completa con su implementación con un grupo de 30 candidatos, cuyos procedimientos y resultados se exponen en el apartado 4.1.3.

3.2 Validación de la escala de evaluación

La validación del instrumento de evaluación se ha realizado en dos fases. Primero, se han tenido en cuenta las rúbricas incluidas en obras de referencia y en exámenes de nivel anteriores, como se ha señalado en el apartado 2.2. Segundo, se ha utilizado el cuestionario de validación con expertos que se describe a continuación. Mediante estos dos procedimientos metodológicos, se ha verificado la ‘validez del contenido’ y del ‘constructo’ de la escala, es decir, se ha comprobado que los criterios de la escala y los descriptores incluidos en cada criterio reúnen todos los aspectos que deben medirse y, por consiguiente, la escala permite obtener la información necesaria para determinar el nivel de competencia de los candidatos. Además, se ha confirmado la ‘validez de criterio’, es decir, el grado en que los puntajes asignados a cada descriptor por parte del comité de expertos son válidos al compararlos con la escala de referencia. Y, por último, se ha verificado la ‘fiabilidad’ de la escala, esto es, si hay consistencia, coherencia y homogeneidad en la asignación de descriptores con los puntos en cada criterio.

3.2.1 Cuestionario de validación con expertos

El cuestionario para validar la escala de evaluación se ha distribuido mediante la plataforma *QuestionPro*⁵ y ha sido respondido por un comité de diez especialistas: seis expertos en evaluación de segundas lenguas, dos en destrezas orales y dos en enseñanza de español como lengua de migración.

El cuestionario se compone de siete páginas en total, las seis primeras están destinadas a valorar cada uno de los criterios de la escala: pertinencia (primera página), coherencia y cohesión (segunda página), fluidez y pronunciación (tercera página), corrección (cuarta página), alcance (quinta página) e impresión global (sexta página). En cada página hay tres preguntas. La primera pregunta consiste en relacionar los nueve descriptores de cada criterio con los puntos que se le asignan. En la segunda pregunta debe indicarse el grado de acuerdo respecto a unas afirmaciones sobre la formulación de los descriptores (su grado de claridad expositiva, su extensión y su independencia) a partir de una escala de Likert que ofrece cinco posibilidades: 1. Totalmente en desacuerdo, 2. En desacuerdo, 3. Ni de acuerdo ni en desacuerdo, 4. De acuerdo y 5. Totalmente de acuerdo. Finalmente, la tercera cuestión ofrece la posibilidad de añadir comentarios y observaciones sobre el criterio y los descriptores analizados.

Por su parte, la séptima página aborda la correspondencia entre los nueve descriptores de nuestra escala y los niveles del MCER. La primera pregunta de esta parte se centra en el criterio que mide la impresión global con el objetivo de relacionar los nueve descriptores de este criterio con los seis niveles de competencia del MCER: A1, A2, B1, B2, C1 y C2. Las tres preguntas siguientes muestran como ejemplo la valoración que han recibido tres candidatos en cada uno de los seis criterios. A partir de ahí, debe asignarse un nivel del MCER a cada candidato.

3.2.2 Resultados del cuestionario de validación con expertos

La primera pregunta del cuestionario, en la que los expertos asignan un puntaje a cada uno de los nueve descriptores establecidos por cada criterio, permite comprobar el grado de consistencia entre ellos y verificar, así, el grado de fiabilidad de la rúbrica. Para ello, se ha utilizado la prueba estadística del Alfa de Cronbach y el coeficiente de correlación intraclase, siguiendo las recomendaciones de Robles Garrote y Rojas (2015) y Guzmán-Cedillo et al. (2017) para medir la consistencia interna y la fiabilidad de un instrumento. Estas pruebas se han aplicado a los puntos que cada experto asigna a cada descriptor en todos los criterios. Los resultados muestran un grado de fiabilidad estadísticamente significativo para todos los criterios con un $p < .001$. En la Tabla 3 se recogen los resultados específicos para cada criterio.

Tabla 3. Resultados sobre la fiabilidad de la escala de evaluación

	Pertinencia	Coherencia	Pronunciación	Corrección	Alcance	Global
Alfa de Cronbach	.991	.994	.987	.993	.995	.995
CC Intraclase	.991	.991	.985	.990	.994	.994
Sig.	.001	.001	.001	.001	.001	.001

Tras comprobar que los datos siguen una distribución normal según las pruebas de normalidad de Shapiro-Wilk, los puntos que los expertos asignan a los descriptores se han comparado con los valores reales de la escala con el fin de comprobar la validez de criterio mediante la prueba CC de Pearson. Como muestra la Tabla 4, los resultados exponen que la correlación entre los puntajes de los expertos y los establecidos en la rúbrica es significativa en todos los casos $p < .001$, con lo que queda verificada la validez de criterio.

Tabla 4. Resultados sobre la validez de criterio de la escala de evaluación

	Pertinencia	Coherencia	Pronunciación	Corrección	Alcance	Global
Correlación de Pearson	0.997	0.996	0.994	0.984	0.962	0.995
Sig.	.001	.001	.001	.001	.001	.001

Teniendo en cuenta que el valor de las pruebas estadísticas aplicadas es superior a 0.9, podemos establecer que el instrumento de evaluación ha sido validado con un alto grado de validez y fiabilidad.

Por otra parte, las respuestas a la segunda y la tercera pregunta del cuestionario permiten comprobar la validez de contenido y de constructo. Empezamos con la segunda pregunta, dirigida a valorar el nivel de acuerdo/desacuerdo de los expertos respecto a distintas afirmaciones sobre la escala. En relación con el primer ítem, sobre el nivel de adecuación que tiene la extensión de los descriptores, los resultados muestran un 100% de acuerdo o total acuerdo en los criterios de pertinencia, coherencia, pronunciación, corrección e impresión global. Sin embargo, la valoración sobre el alcance léxico mostró un 80% de acuerdo o total acuerdo y un 20% ni acuerdo ni desacuerdo.

En cuanto al segundo ítem, sobre si la enunciación de los descriptores es clara (lengua comprensible y sintaxis sencilla), los resultados son similares a los anteriores. Existe un 100% de acuerdo o total acuerdo en los criterios de pertinencia, coherencia, pronunciación, corrección e impresión global. Por su parte, en alcance léxico se observó un 90% de acuerdo o total acuerdo y un 10% ni acuerdo ni desacuerdo.

Por lo que respecta al tercer ítem, también encontramos una gran consistencia entre los evaluadores, quienes están de acuerdo con que los descriptores son independientes y se entienden por sí mismos. Para los criterios de pertinencia e impresión global, existe un 80% de acuerdo o total acuerdo y un 20% de desacuerdo; para el de coherencia, un 70% de acuerdo o total acuerdo, hay un 20% que no está de acuerdo ni en desacuerdo, y un 10% de desacuerdo; para los de pronunciación, corrección y alcance, se obtiene un 80 de acuerdo o total acuerdo, un 10% que no está de acuerdo ni en desacuerdo y un 10% de desacuerdo.

Como muestran los datos, hay casos específicos en los que los evaluadores se muestran en desacuerdo con algún criterio. Aun así, entre sus comentarios al respecto reflejados en la tercera pregunta del cuestionario señalan, por ejemplo, que “es inevitable que algunos descriptores dependan de otros, porque no hablamos de un nivel concreto de lengua”. Otros comentarios por parte de los evaluadores apuntan que algunos términos utilizados en la enunciación de los descriptores resultan poco claros. Concretamente, un experto se refiere a la “dificultad para distinguir *bastante amplio* vs. *relativamente amplio*”. En esta misma línea, otro experto sugiere que “quizá la alusión a la *extensión apropiada* se podría perfilar”, pero que, en general, los descriptores “son claros y específicos”. Asimismo, los expertos detectan alguna errata en los descriptores; proponen revisar la formulación de algún descriptor; sustituir el verbo *conocer* de los descriptores en el criterio de corrección; y evitar hacer alusiones al concepto del *acento extranjero* en el criterio de la pronunciación porque les parece un poco problemático ya que debemos preguntarnos antes: “¿Acento con respecto a qué

variedad del español?”. Estas apreciaciones se han tenido en cuenta en la modificación de la rúbrica, tal y como queda reflejado en el apartado 4.2.2.

4. Fase de implementación

4.1. Implementación de la prueba y la rúbrica en su conjunto

4.1.1 Contexto y participantes

Con el objetivo de comprobar la validez y la fiabilidad de la prueba y la rúbrica en su conjunto, el examen de nivel se ha pilotado con un total de 30 candidatos de doce nacionalidades distintas procedentes de países de África, Asia, Europa y América Central y del Sur. Participaron seis mujeres y veinticuatro hombres, con una edad media de 31 años, solicitantes de asilo, refugiados y migrantes. Los participantes tienen un nivel sociocultural y económico dispar. Asimismo, la mayoría se encuentran en situación de desempleo, aunque un número reducido de participantes está en activo y son profesores inmigrantes procedentes de Francia y Jordania. En el momento de realizar el examen, el tiempo de residencia en España de los participantes es heterogéneo, pues comprende desde 1 año hasta los 15, lo que refleja la realidad de la población migrante que aprende español y podría ser usuaria potencial del POEM. La implementación de la prueba se realizó entre los meses de diciembre de 2020 y marzo de 2021, y se llevó a cabo en todos los casos por la misma examinadora, de cara a evitar la variabilidad propia de los exámenes orales realizados por diferentes evaluadores (Karavas & Delieza, 2009). Entre sus funciones, la examinadora administraba el examen, controlaba los tiempos y registraba mediante grabaciones de voz el desarrollo de la prueba. En ningún caso la examinadora evaluó al candidato durante el examen. La realización de los exámenes ha tenido lugar en espacios de la Universidad Nebrija en Madrid, concretamente, en las aulas del Campus Joaquín María López y en el *Living Lab* instalado por convenio en el Centro de Acogida Temporal de Cruz Roja “CAT – San Blas”. Un número reducido de exámenes se han realizado de forma telemática mediante la plataforma para videoconferencias *Microsoft Teams* a causa de la situación de emergencia sanitaria causada por la Covid-19. Tanto los candidatos como la examinadora tenían el micrófono y el vídeo encendidos, y se encontraban en espacios sin ruidos para minimizar el impacto de la mediación tecnológica.

4.1.2 Procedimiento

Una vez concluido el proceso de implementación del examen, se solicitó su validación a un comité de 30 especialistas en enseñanza de ELE y evaluación de competencia oral en ELE de la Universidad Nebrija, la KU Leuven, la Universidad de Jaén y la Universidad de Siena. Se trata de uno de los tres tipos de análisis de la

fiabilidad que identifica Louma (2004), concretamente, la fiabilidad interevaluador, que supone que diferentes evaluadores puntúan las realizaciones de los candidatos de maneras similares. El procedimiento de validación comprende, por una parte, la evaluación de una selección de muestras orales de examen en forma de grabaciones de audio y, por otra parte, un cuestionario con preguntas sobre el examen. En la primera parte han participado un total de 30 especialistas, cada uno de los cuales ha evaluado tres muestras de examen, de forma que cada candidato cuenta con tres evaluaciones distintas. De los 30 expertos, 14 han respondido el cuestionario sobre el examen. Todos los especialistas recibieron previamente una formación por parte de las autoras de este artículo con información sobre el contenido y la estructura del examen, el manejo de la escala de evaluación y el cuestionario con preguntas sobre la prueba. Para realizar la validación, cada experto recibió las grabaciones en formato mp3 correspondientes a las muestras de examen que debía evaluar; la ficha del examen con las tareas y las instrucciones para el examinador; la escala de evaluación; y el enlace al cuestionario online.

El cuestionario, distribuido mediante la herramienta *Google Forms*, se compone de 11 preguntas que inciden en los siguientes aspectos: adecuación de la duración de cada tarea y del examen completo; adecuación del contenido de las preguntas; adecuación de las tareas a todos los niveles de competencia; claridad de las instrucciones que recibe el examinador para administrar la prueba y el candidato para realizar las tareas; y obtención de muestras de lengua suficientes para determinar el nivel de los candidatos a partir de las tareas propuestas. El cuestionario incluye preguntas cerradas con dos opciones de respuesta posible: ‘sí’ y ‘no’. Cada pregunta se acompaña de un espacio para aportar comentarios y observaciones. Los evaluadores tuvieron un plazo de 3 semanas para validar la prueba.

El proceso descrito, además de corroborar la validez de contenido y de constructo de la prueba, que ya se había comprobado parcialmente según lo expuesto en el apartado 3.1, permite verificar su fiabilidad, es decir, el grado en el que esta herramienta es capaz de medir el nivel de español del grupo meta sin error. Para ello, se comprueba si hay consistencia, coherencia y homogeneidad en los resultados obtenidos en tres evaluaciones diferentes de un mismo candidato.

4.1.3 Resultados de la implementación

En primer lugar, se han aplicado nuevamente las pruebas estadísticas de Alfa de Cronbach y CC intraclase a la puntuación total que cada evaluador le ha dado a cada candidato. De esta manera, se ha medido el grado de consistencia interna entre las puntuaciones que los tres evaluadores correspondientes le dan a un mismo candidato. El nivel de consistencia es muy alto, con un valor de .969 en el Alfa de Cronbach y en el CC Intraclase, que resulta estadísticamente significativo ($p < .001$).

En segundo lugar, se ha tenido en cuenta que un candidato puede obtener la misma puntuación total por los tres evaluadores (p. ej.: 54), pero ello no implica que haya acuerdo en la puntuación asignada en todos los criterios. Es decir, todos los evaluadores pueden dar la misma puntuación total a un candidato calificando de manera dispar los diferentes criterios. Por consiguiente, se ha medido también el grado de acuerdo por criterios. Los resultados arrojan nuevamente un alto nivel de consistencia en todos los criterios que se muestra significativo en todos los casos ($p < .001$). La tabla 5 recoge los valores del Alfa de Cronbach y el CC Intraclase obtenidos para cada criterio. La última columna de la tabla también refleja los valores correspondientes a la puntuación total de los candidatos, comentada anteriormente.

Tabla 5. Resultados de la validación de la prueba y la escala de evaluación en su conjunto

	Pertinencia	Coherencia	Pronunciación	Corrección	Alcance	Global	TOTAL
Alfa de Cronbach	.917	.932	.913	.941	.916	.948	.962
CC Intraclase	.919	.933	.912	.941	.918	.949	.962
Sig.	.001	.001	.001	.001	.001	.001	.001

Con estos resultados, podemos concluir que el examen y la rúbrica son instrumentos de evaluación fiables para establecer el nivel de competencia oral en español de los candidatos evaluados.

Por otra parte, los resultados del cuestionario que completaron los evaluadores arrojan los siguientes datos. En primer lugar, el cuestionario hacía referencia a la duración de cada una de las dos tareas y de la prueba en general. Respecto a la primera tarea, 11 evaluadores (78,57%) consideran que la duración es adecuada; en cuanto a la segunda tarea, 13 evaluadores (92,85%) consideran que es adecuada, y finalmente 11 (78,57%) consideran que la duración total de la prueba es adecuada. Cabe destacar que todos los evaluadores que no consideraban adecuada la duración hicieron referencia a la longitud de la primera tarea y recomendaron acortar la primera parte para que ambas tareas quedaran más equilibradas.

En segundo lugar, se preguntó sobre la adecuación del examen para evaluar los diferentes niveles de competencia en español. Con ello, se pretendía conocer la opinión de los expertos sobre si candidatos con cualquier nivel podrían realizar en mayor o menor medida las tareas y si la prueba en general lograría obtener suficientes muestras de lengua para evaluar todos los niveles abarcados (A1-C2). En este punto, el 100% de los evaluadores consideran que la prueba genera suficientes muestras de lengua para evaluar cualquier nivel; sin embargo, solo 12 evaluadores (85,71%) consideran que la tarea 2 se adecua a todos los niveles, y 9 (64,28%) así lo estiman respecto a la tarea 1. En este caso, los expertos recomendaron incluir preguntas más elaboradas que elicitaran estructuras gramaticales más complejas para poder distinguir mejor los niveles de dominio C y, a su vez, dar la posibilidad de que los candidatos

avanzados profundicen en temas que excedan lo laboral o personal, para producir más vocabulario de especialidad.

En tercer lugar, se les preguntó a los evaluadores si añadirían, modificarían o eliminarían algo en cada una de las tareas. En la primera tarea, 5 evaluadores (35,71%) no realizarían ninguna modificación, mientras que 9 de ellos propusieron quitar o agregar alguna pregunta en particular, las cuales fueron consideradas en las modificaciones (ver apartado 4.2.1). Por su parte, en la segunda tarea, 9 evaluadores (64,28%) no cambiarían nada, y los otros 5 propusieron agregar más temas para que el candidato pueda elegir entre más opciones además del trabajo o la vivienda.

Finalmente, el cuestionario hacía referencia a las instrucciones para el candidato (dadas en la tarea 2), las instrucciones para el examinador que vaya a implementar la prueba, y la accesibilidad de la rúbrica para el evaluador. Respecto a las dos últimas, 13 expertos (92,85%) consideraron que tanto las instrucciones como la claridad de la rúbrica son adecuadas. Por su parte, solo 4 sugirieron cambios en las instrucciones que se dan al candidato durante la segunda tarea, particularmente, estos cambios se referían a dar instrucciones más simples y hacer preguntas más directas a los candidatos con nivel A.

4.2 Modificaciones tras la implementación

El proceso de implementación del examen ha sido fundamental para verificar los aspectos positivos de la prueba y de la escala de evaluación, así como para identificar las modificaciones que eran necesarias implementar antes de obtener la versión final de la prueba y de la escala.

4.2.1 Modificaciones aplicadas al examen

La primera tarea del examen sufre una serie de modificaciones que aplican al contenido de las preguntas y la extensión de la prueba. A continuación, se resumen los principales cambios:

- En el bloque 1. Información personal, se sustituye la pregunta *¿Qué es lo que más te gusta de estudiar español?, ¿y lo que menos?* por *¿Piensas que es importante estudiar español?, ¿por qué?*, en tanto que la primera provocaba problemas de comprensión.
- Por el mismo motivo que el cambio anterior, en el bloque 3. Relaciones sociales, las preguntas: *Por lo general, ¿te gusta relacionarte con la gente o prefieres estar solo?* y *Cuando te relacionas con la gente, ¿qué cosas te gustan y te molestan en una persona?* se cambian por *¿Qué es lo que más te gusta de estar con amigos, compañeros, familia...?, ¿te relacionas con españoles normalmente?, ¿qué ventajas y desventajas tiene relacionarse con españoles?*

- Se añaden las siguientes preguntas: *¿Qué hubieras hecho si en tu país te hubieran ofrecido el trabajo de tus sueños o la situación político-social hubiera sido favorable?* (Bloque 4. El trabajo) y *Si pudieras cambiar alguna de las costumbres de tu país o de España, ¿qué cambiarías?, ¿por qué?* (Bloque 5. Ciudades y costumbres), con el objetivo de activar la producción de construcciones más complejas.
- Se elimina el bloque 6. El día que llegaste a Madrid, tras observar que las preguntas tratan un tema delicado que a menudo producía situaciones incómodas durante la administración del examen. Este bloque iba dirigido especialmente a activar el uso de los tiempos pasados, de forma que, tras eliminarlo, se adapta la segunda pregunta del bloque 4 de esta manera: *Y en tu país, ¿en qué trabajabas?, ¿cómo era tu trabajo allí?*. Al prescindir de un bloque, la duración de esta prueba disminuye.

En la tarea 2 no se aplican cambios como tal tras la validación; sin embargo, se modifica la parte dedicada a las instrucciones que recibe el examinador, incorporando pautas más específicas sobre cómo actuar en función de su percepción respecto al nivel del candidato. En los niveles iniciales, el examinador se limita a pedirle al candidato que lo ayude con la situación, incentivándolo a interactuar de manera simple, preguntándole: “¿qué puedo hacer?, ¿me ayudas?” En los niveles intermedios, en primer lugar, el evaluador le pide al candidato que le haga preguntas para obtener más información sobre la situación descrita; en segundo lugar, le pide que dé consejos (p. ej.: “¿qué crees que debo hacer?”, “¿qué harías tú en mi lugar?”) En los niveles avanzados, el examinador debe aumentar el grado de complejidad de las tareas anteriores, por ejemplo, aportando contraargumentos, fomentando la negociación e invitando al acuerdo.

Por último, con la intención de que la implementación del examen sea práctica y eficaz, y siguiendo las recomendaciones obtenidas a través del cuestionario sobre la duración de la prueba, se opta por gestionar su administración de forma que no supere los 20 minutos.

4.2.2 Modificaciones aplicadas a la escala

Tras el proceso de validación de la escala de evaluación, se han realizado un conjunto de cambios relativos a uno de los criterios evaluados: fluidez y pronunciación, así como la redacción de algunos descriptores que incluían términos difíciles de interpretar y resultaban confusos. En el primer caso, se ha optado por valorar la fluidez y la pronunciación como criterios independientes, de manera que, por un lado, se evalúa la pronunciación, que incluye aspectos sobre la claridad en la dicción, la adecuación de la entonación y los matices entonativos, y, por otro lado, se evalúa la fluidez con la que se comunica el candidato como parte de la impresión global que produce su actuación. En el segundo caso, en cuanto a la redacción de los descriptores, se ha aclarado el uso de algunos términos con el objetivo de facilitar la

distinción entre descriptores y, en definitiva, simplificar el manejo de la herramienta. Por ejemplo, en el criterio sobre el grado de coherencia y cohesión del discurso, se elimina la distinción entre *un número suficiente de conectores* (banda 6) y *número adecuado de conectores* (banda 7), y se sustituye por *un número suficiente de conectores* (banda 6) y *conectores adecuados y variados* (banda 7). Asimismo, en el criterio relativo a la impresión global, los descriptores: *El candidato se comunica sin grandes dificultades. Su actuación lingüística general es limitada, pero comprensible* y *El candidato se comunica sin dificultades. Su discurso es comprensible, aunque comete errores e imprecisiones*, situados en las bandas 4 y 5 respectivamente, se sustituyen por: *El candidato se comunica sin grandes dificultades. Su actuación lingüística general es limitada, se observan algunas pausas y titubeos, pero su discurso es comprensible* (banda 4) y *El candidato se comunica sin dificultades. Habla de forma continuada y las pausas para planificar su discurso son ocasionales. Su discurso es comprensible, aunque comete errores léxicos y gramaticales* (banda 5).

Lo anterior también ha contribuido a que la formulación de los descriptores y las diferencias entre los que se encuentran en bandas de puntos contiguas sea más sistemática. Asimismo, se han eliminado las referencias al acento extranjero por la complejidad que implica su tratamiento, ya que a menudo un acento marcado no se considera un descriptor de nivel si no dificulta la comprensión (Munro & Derwing, 1995; Levis, 2006).

CONCLUSIONES

La evaluación de la competencia lingüística comunicativa debe basarse en una teoría de la medición y en una teoría del uso de la lengua. De acuerdo con Fulcher (2015), la forma en que definimos el constructo de expresión oral y la forma en que diseñamos los descriptores de la escala de valoración están interconectados, con lo que las decisiones de diseño deben basarse en el propósito de las pruebas y en los marcos teóricos pertinentes. En el caso que nos ocupa, se ha llevado a cabo el diseño y validación de una prueba consistente en tareas de interacción oral uno a uno, concebida para incrementar la significatividad y la susceptibilidad de generalizar [*generalizability*] los resultados de la evaluación, tal como los describen Bachman y Palmer (2010), a saber, el grado en el que un registro de evaluación proporciona información sobre la habilidad que se está midiendo y la transmite en términos comprensibles, y el grado de correspondencia entre una tarea de evaluación de la lengua y una tarea de uso de la lengua en las características de la propia tarea y en los componentes de la habilidad lingüística implicados.

Concretamente, POEM surge ante la necesidad de disponer de un examen multinivel que permita evaluar la competencia comunicativa oral en español en un contexto de migración y refugio. El proceso de diseño y validación empírica de este examen se ha desarrollado en tres fases. La primera corresponde al diseño preliminar de la prueba y la escala de evaluación; la segunda, a la validación de la prueba y de la

escala de evaluación por juicio de expertos; y la tercera, a la implementación del examen, con la participación de 30 estudiantes de español inmigrantes, 1 examinadora y 30 evaluadores. Mediante este proceso, se ha comprobado que tanto la prueba como la escala de evaluación son válidas en relación con el contenido y el constructo. Asimismo, se ha verificado que la escala es fiable y muestra validez de criterio en todos los aspectos que evalúa. Finalmente, la implementación del examen en su conjunto ha permitido establecer que estamos ante una herramienta válida y fiable para evaluar el nivel de español oral de migrantes y refugiados.

El alcance del presente estudio ha de contemplarse en dos direcciones. Por una parte, contribuye a ampliar el estado de la cuestión sobre evaluación de lengua de migración. Por otra, ofrece un recurso válido y fiable de aplicabilidad en el proceso de enseñanza, aprendizaje y evaluación de la nueva lengua, a disposición de investigadores, profesores y voluntarios dedicados a atender a la población desplazada, con el ánimo de contribuir a una mejora de este proceso. De acuerdo con Budría y Martínez Ibarreta (2021): “el conocimiento del español es un salvoconducto que pone en valor las capacidades y conocimientos que el inmigrante trae consigo, reforzando su posición laboral y su situación económica” (p. 97).

Confiamos en que un instrumento como la POEM ayude a evidenciar la competencia lingüística comunicativa de los trabajadores migrantes en la lengua de la población de acogida, en aras de favorecer su desarrollo académico y profesional y garantizar el ejercicio de sus derechos.

REFERENCIAS BIBLIOGRÁFICAS

Antón, M. (2019). *Métodos de evaluación de ELE*. Arco Libros.

Association of Language Testing in Europe & Council of Europe (2018). *Linguistic Integration of Adult Migrants: Requirements and Learning Opportunities. Report on the 2018 Council of Europe and ALTE survey on language and knowledge of society policies for migrants*. <https://rm.coe.int/linguistic-integration-of-adult-migrants-requirements-and-learning-opp/16809b93cb>

Bachman, L. y Palmer, A. (2010). *Language Assessment in Practice*. Oxford University Press.

Baralo, M. (2012). Diseño de certificación lingüística comunicativa: ¿cómo evaluar la expresión e interacción escritas y orales? *Revista Internacional de Lenguas Extranjeras*, 1, 11-30.

Baralo, M. y Estaire, S. (2011). Variables socioculturales y comunicativas en el diseño curricular de una certificación de español para trabajadores inmigrantes. *Lengua y Migración*, 3(2), 5-41.

- Baralo, M. y Martín Leralta, S. (Coords.) (2010). *Guía del profesor del examen LETRA*. http://www.diplomalettra.com/examen_letra/pdf/LETRA_guia_para_el_profesor.pdf
- Baralo, M., Martín Leralta, S. y Pascual, I. (2016). La certificación de las competencias comunicativas del español para inmigrantes. *Porta Linguarum*, 25, 105-117. <https://doi.org/10.30827/Digibug.53892>
- Bordón, T. (2000). La evaluación de la expresión oral en el aula de ELE. *Carabela*, 47, 151-175.
- Bordón, T. (2007). [Las destrezas orales en la enseñanza del español L2-LE](https://cvc.cervantes.es/ensenanza/biblioteca_ele/asele/pdf/17/17_0097.pdf). En E. Balmaseda Maestu (Coord.), *XVII Congreso Internacional de la Asociación del Español como lengua extranjera (ASELE): logroño 27-30 de septiembre de 2006* (Vol. 1, pp. 97-112). https://cvc.cervantes.es/ensenanza/biblioteca_ele/asele/pdf/17/17_0097.pdf
- Bordón, T. (2015). La evaluación de segundas lenguas (L2). Balance y perspectivas. *Revista Internacional de Lenguas Extranjeras*, 4, 9-30.
- Bordón, T. y Liskin-Gasparro, J. (2007). Evaluación. En M. Lacorte (Ed.), *Lingüística Aplicada del Español* (pp. 211-251). Arco Libros.
- Budría, S. y Martínez de Iberreta, C. (2021). El español como elemento integrador en el mercado laboral. En J. L. García Delgado (Dir.), *El Español, Lengua Internacional: Proyección y Economía* (pp. 85-99). Thomson Reuters Aranzadi.
- Council of Europe (2018). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors*. Press Syndicate of the University of Cambridge. <https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989>
- Council of Europe & Language Policy Division (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR) A Manual*. <https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680667a2d>
- Fulcher, G. (2015). Research Timeline: Assessing Second Language Speaking. *Language Teaching*, 48, 198-216.
- Guzmán-Cedillo, Y. I., Lima-Villeda, D. N. y Meza-Cano, J. M. (2017). Diseño y confiabilidad de una rúbrica para evaluar infografías didácticas. *Enseñanza & Teaching*, 35(2), 17-36.

- Instituto Cervantes (2006). *Plan Curricular del Instituto Cervantes. Niveles de referencia para el español*. https://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular/
- Judd, E. L., Tan, L. y Walberg, H. J. (2001). Teaching Additional Languages. *Educational Practices Series*, 6, 1-30.
- Karavas, E. y Delieza, X. (2009). On Site Observation of KPG Oral Examiners: Implications for Oral Examiner Training and Evaluation. *Apples: Journal of Applied Language Studies*, 3(1), 51-77.
- Levis, J. M. (2006). Pronunciation and the Assessment of Spoken Language. En R. Hughes (Ed.), *Spoken English, TESOL and Applied Linguistics: Challenges for Theory and Practice* (pp. 245-270). Palgrave Macmillan.
- Llorián González, S. (2015). Exámenes certificativos de ELE referenciados a sistemas externos como el del MCER. *Revista Internacional de Lenguas Extranjeras = International Journal of Foreign Languages*, 4, 97-122.
- Luoma, S. (2004). *Assessing Speaking*. Cambridge University Press.
- Martín Leralta, S. (2011). Certificación lingüística de nivel inicial para inmigrantes en contexto laboral: ejemplo de una prueba del examen DILE. *Lengua y migración/Language and Migration*, 3(1), 89-104. <https://ebuah.uah.es/dspace/handle/10017/10823>
- Martín Leralta, S. (2013). Diseño y formación de evaluadores de una prueba de eficacia comunicativa oral en lengua española (PECOLE). En A. Llanes (Coord.), *La Lingüística Aplicada en la Era de la Globalización* (pp. 186-191). Ediciones de la Universidad de Lleida.
- Mavrou, I. y Martín Leralta, S. (2018). Factores determinantes de la competencia comunicativa en español de trabajadores inmigrantes residentes en Madrid. *Círculo de Lingüística Aplicada a la Comunicación*, 74, 221-246. <https://doi.org/10.5209/CLAC.60521>
- Munro, M. J. y Derwing, T. M. (1995). Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners. *Language Learning*, 45(1), 73-97.
- Pastor Cesteros, S. y Roca Marín, S. (Coords.) (2007). *La evaluación en el aprendizaje y la enseñanza del español como lengua extranjera / segunda lengua*. XVIII Congreso Internacional de la Asociación para la Enseñanza del Español como lengua Extranjera (ASELE): Alicante, 19-22 de septiembre de 2007. Universitat d'Alacant.

- Robles Garrote, P. y Rojas, M. D. C. (2015). La validación por juicio de expertos: dos investigaciones cualitativas en Lingüística aplicada. *Revista Nebrija de Lingüística Aplicada*, 18.
- Villalba, F. (2012). La evaluación en el aprendizaje de lenguas por migrantes. En I. Ballano (Ed.), *Evaluar la competencia lingüística del alumnado inmigrante* (pp. 109-126). Universidad de Deusto.
- Villalba, F. (2018). Inmigración y enseñanza a adultos. En J. Muñoz-Basols, E. Gironzetti y M. Lacorte (Eds.). *The Routledge Handbook of Spanish Language Teaching* (pp. 463-476). Routledge.

NOTAS

¹ Este trabajo está financiado con los fondos de la Cátedra Global Nebrija-Santander del español como lengua de migrantes y refugiados. Además, se enmarca en los proyectos de investigación *EMILLA: emoción, memoria, identidad lingüística y aculturación emocional: su influencia en el aprendizaje de español como lengua de migración*, como parte del subproyecto EMOidentiCOM (FFI2017-83166-C2-1-R) e *INMIGRA3-CM: La población migrante en la Comunidad de Madrid: factores lingüísticos, comunicativos, culturales y sociales del proceso de integración y recursos lingüísticos de intervención* (H2019/HUM5772).

² Siguiendo a Judd et al. (2001), en este trabajo se utiliza el término 'lengua adicional' frente a otras posibles denominaciones como 'segunda lengua' y 'lengua extranjera' por adaptarse mejor al contexto lingüístico de los alumnos meta.

³ Versión final de la prueba disponible en: https://osf.io/tq6rv/?view_only=0a40de9df2ce4cde95ad1883e94590b6

⁴ Versión final de la escala de evaluación disponible en: https://osf.io/tq6rv/?view_only=0a40de9df2ce4cde95ad1883e94590b6

⁵ Disponible en: <https://www.questionpro.com/t/AQz73ZjH4m>