

Exámenes certificativos de ELE referenciados a sistemas externos como el del *MCER*

Susana Llorián González
Universidad Nebrija
Universidad Complutense de Madrid
slloriang@lys.jazztel.es

Resumen: La cultura de los exámenes de lenguas referenciados a sistemas de estándares externos se está imponiendo progresivamente en el panorama internacional. El impacto sin precedentes del *Marco común europeo de referencia (MCER)* ha erigido a este documento en el principal sistema con el que se relacionan los resultados de los exámenes en todo el mundo. Este fenómeno lleva aparejado el reto de la argumentación de la validez y el de la superación de los problemas que presenta el *MCER* para este tipo de uso, lo cual implica la profesionalización y el compromiso ético de los agentes implicados en el desarrollo y la administración de exámenes.

Palabras clave: Sistemas de referencia externos, exámenes vinculados al *MCER*, exámenes referenciados a estándares externos.

Abstract: The culture of aligning tests to external standard is being increasingly imposed in worldwide language certification exams. The unprecedented impact of the Common European Framework (CEF) has placed this document in the primary standard system that the test results relate. This phenomenon entails the challenge of the validity argument and overcoming the implicit problems of using CEF as a reference in testing. This involves professional and ethical commitment of those involved in the development and management testing systems.

Keywords: Standard setting, standard systems, linking exams to external standard systems, aligning exams to external standard system

0. Introducción

La desaparición progresiva de las barreras transnacionales ha incrementado el interés de las organizaciones y de los ciudadanos a título individual por poder transmitir en cualquier lugar, de forma transparente y compatible, el significado de las calificaciones de los exámenes que miden la capacidad de comunicarse en una lengua extranjera. Por consiguiente, los principales exámenes certificativos de idiomas en el mundo recurren cada vez más a sistemas de estándares comparables entre sí. Entre todos, sobresale el uso masivo, dentro y fuera de los límites del continente europeo, del *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza y evaluación* (en adelante, *Marco de referencia*, *Marco* o *MCER*). Con ello, no solo se ven superados los confines continentales en los que se inscriben los destinatarios del *Marco*, sino que el fenómeno del reconocimiento mutuo de las certificaciones referenciadas a la escala de los niveles del *MCER*, ya generalizado en todo el mundo, eclipsa el propósito principal de un documento —más político que técnico— que se ha desarrollado para promover los valores sobre los que se sustenta el Consejo de Europa: la ciudadanía democrática, la cooperación y el entendimiento mutuo, a través del plurilingüismo.

El uso del *MCER* como sistema de estándares de referencia para la comparabilidad de certificaciones no está exento de problemas: a la multifuncionalidad del documento, se suma el grado de generalidad con el que está formulado, que se supone que lo hace apto para aplicarse a diferentes contextos. Esto implica que se pueden relacionar entre sí resultados de exámenes con propósitos diversos, de diferentes lenguas, asimilando las notas a uno o varios niveles de la escala del *MCER*.

Las escalas de los niveles comunes de referencia pretenden facilitar la descripción del nivel de dominio lingüístico alcanzado en los diplomas y certificados existentes y de esa forma permitir la comparación entre sistemas.

MCER, 2002: 9.2.9

La comparabilidad podría, sin embargo, quedar en entredicho si la comunidad evaluadora no se compromete a alcanzar un consenso acerca de los procedimientos que se deben aplicar para relacionar los resultados de los exámenes con los descriptores de las escalas del *MCER*. A pesar de los esfuerzos del Consejo de Europa y de los expertos para evitar usos espurios de las escalas del *Marco* para fines certificativos, los interrogantes siguientes siguen abiertos: ¿Qué procedimientos de vinculación garantizan realmente que los resultados de los exámenes se corresponden, en efecto, con los niveles del *MCER*? ¿Qué cantidad de procedimientos se requieren en virtud del perfil de examen y de los recursos disponibles?

¿Quién aplica los procedimientos? ¿Con qué frecuencia, grado de intensidad, de rigor...? ¿Quién supervisa y verifica la aplicación y sus resultados? La falta de respuestas contundentes a estas preguntas puede originar fallos, que, en ocasiones, afectan a la convergencia de las certificaciones que, en teoría, debieran ser equivalentes. ¿Debería, por ejemplo, una universidad, que declara que los alumnos de ingreso precisan un nivel B2, aceptar un certificado equivalente de español de un examen desarrollado con fines generales? ¿Con qué garantías se cuenta para ello?

Es necesario recordar, además, que son varios los autores (Alderson, 2007; Alderson *et al*, 2004; Fulcher, 2004, 2010; Weir, 2005; Fulcher y Davidson, 2007; Davidson y Fulcher, 2007, entre otros) que dan cuenta de las numerosas deficiencias que presenta la formulación de algunos de los descriptores de las escalas del MCER. También, que los propios autores del *Marco* reconocen que los descriptores de algunas de las escalas, en especial, las de los niveles C1 y C2 no han sido adecuadamente calibrados (North, 2007). Pese a la difusión posterior de bancos de descriptores adicionales (Consejo de Europa, 2007) para estos niveles a través de las páginas electrónicas de la organización, este tipo de problemas sigue constituyendo un foco de amenazas para la fiabilidad de los resultados de los exámenes y, por consiguiente, para la validez.

La asimilación de los resultados de los exámenes a uno o varios de los niveles de la escala del MCER precisa del despliegue de un aparato argumentativo con el que justificar esa suerte de declaración. En el paradigma actual de la evaluación mediante exámenes con fines certificativos, este tipo de demostraciones necesita la aportación de evidencias que proporcionen bases teóricas y empíricas a los argumentos.

El logro de las exigencias garantistas y la superación de los problemas y dificultades, más arriba comentados, requiere de la colaboración entre los miembros de la comunidad evaluadora, de la provisión de herramientas y de recursos que medien entre el MCER y los instrumentos de evaluación y, sobre todo, de la profesionalización y la responsabilidad de los agentes involucrados en el desarrollo de los exámenes, además de su compromiso, en especial, en el caso de exámenes de alto perfil. En el contexto de la certificación lingüística del español, es inevitable formularse las preguntas siguientes: ¿Está capacitado el colectivo de profesionales de Español como Lengua Extranjera (ELE, en lo sucesivo) para asumir este reto? ¿Qué competencias debería desarrollar? ¿Proporcionan los programas de formación de ELE suficientes conocimientos, destrezas en el uso de procedimientos y herramientas, y conciencia de las responsabilidades que todo esto implica?

Este artículo, de tono y contenido divulgativos, tras aclarar los principales conceptos involucrados en la cuestión central que trata, ‘estándar’ y ‘validez’, se detiene brevemente en una descripción muy sucinta de los procedimientos propuestos por el Consejo de Europa para demostrar que los resultados de los exámenes se corresponden con un sistema de niveles de referencia como el del *MCER*. Realiza, además, un recorrido por los principales códigos deontológicos y de buenas prácticas que imperan en el ámbito internacional, y en el de los países de habla hispana en particular, en lo referente a la vinculación de los exámenes a sistemas de estándares internos o externos. Finalmente, analiza la situación general que presentan los principales sistemas de evaluación certificativa del español respecto a todas estas cuestiones.

1. Sistemas de referencia y concepto de estándar

En el área de la evaluación certificativa y, en particular, en el ámbito de la referenciación de los resultados de los exámenes a sistemas externos, el concepto de ‘estándar’ se emplea de forma profusa y con diversas acepciones, por lo que conviene comenzar delimitándolo. Consideramos, en primer lugar, la idea de *estándar* que se refiere a un modelo de actuación lingüística (de recepción o producción), descrito en términos cualitativos y cuantitativos, que se requiere en un contexto determinado —por ejemplo, el mínimo grado de dominio que se precisa para cursar un programa de estudios en una universidad— y con el que se comparan las actuaciones de los candidatos de un examen determinado. Los estándares suelen presentarse en escalas de descriptores del rendimiento, que muestran grados del desarrollo de una capacidad o de aspectos parciales de esta.

La evaluación basada en estándares cuenta con amplia tradición en los países anglosajones, a diferencia de lo que sucede en España. Fulcher (2010: 226) considera que este modelo, que se ajusta al paradigma de la evaluación referenciada a un criterio, es altamente ventajoso, en especial, en el caso de exámenes de alto perfil que se emplean en los procesos de selección de personal, puesto que preserva la transparencia de las decisiones. Contribuye además a garantizar que los candidatos declarados aptos cumplen determinadas condiciones, cuyas consecuencias pueden ser de importancia crucial en materia, por ejemplo, de seguridad cuando se trata de profesionales de la aviación. Este autor se hace eco asimismo de cómo en los Estados Unidos los exámenes del sistema educativo, referenciados a estándares, se erigen en los instrumentos mediante los cuales se mide su funcionamiento y, por consiguiente, se vela por el futuro del país. Los estándares contribuyen significativamente a la mejora del sistema (Fulcher, 2010: 227) y al progreso de las sociedades democráticas.

Mager (1962), en su texto clásico, de signo claramente conductista, y ampliamente difundido acerca de la formulación de los objetivos de instrucción, señala la necesidad de que el texto del objetivo contenga los tres elementos que lo hacen mensurable y, por lo tanto, válido para la evaluación: una acción, expresada en términos de conducta observable (*can do statement*), para cuya formulación resultan útiles, aunque no imprescindibles, las taxonomías de Bloom; unas condiciones en las que se realiza la mencionada acción y un criterio, es decir, el nivel en el que se requiere que la acción sea realizada. Van Ek (1896, citado en Green, 2012: 42) se hace eco de la necesidad de que los objetivos de instrucción reflejen estos tres elementos, que sería cinco para Brown (1995), en los documentos de la serie *T-Level* del Consejo de Europa. Green (2012: 42) advierte de que la fuerte influencia que el modelo de Mager ha venido ejerciendo en el ámbito del aprendizaje, la enseñanza y la evaluación de lenguas se debe al impacto que tiene en los Estados Unidos la corriente que se denominó *Proficiency movement*. Esta tiene origen en las escalas de descriptores del dominio lingüístico desarrolladas en 1986 por el American Council on the Teaching of Foreign Languages (ACTFL), en colaboración con el Educational Testing Service (ETS). Desde la difusión de este modelo, las escalas de descriptores que funcionan a modo de estándares, y que se enuncian siguiendo la propuesta de Mager, comienzan a utilizarse para los mismos usos en el que lo hacen actualmente las del MCER en todo el mundo. De ahí, derivan otras escalas que se emplearán como base para formular objetivos de cursos o como sistemas externos para exámenes cuyos resultados se referencian a un criterio como, por ejemplo, los siguientes: ASLPR (Australian Second Language Proficiency Ratings), desde 1996; el detalladísimo y elaborado sistema conocido como Canadian Language Benchmarks (CLB), desarrollado por el gobierno canadiense en el marco de la integración lingüística de las poblaciones migrantes, actualizado recientemente, en el año 2014; los *can do statements* de ALTE (Association of Language Testers in Europe); los descriptores de dominio de la ICAO (*International Civil Aviation Organization*) para pilotos y controladores aéreos, LPRs (Language Proficiency Requirements for Pilots and Air Traffic Controllers; los estándares de dominio de lenguas (2000) de la OTAN (Organización del Tratado del Atlántico Norte), etc.

Las escalas del ACFL/ETS —que, según Green (2012: 46), fueron utilizadas por North, junto con la mayor parte de las anteriores, en el desarrollo del MCER— a diferencia de las del Marco, son de naturaleza holística, es decir, emplean descriptores únicos, extensos en su redacción, cuyo contenido se refiere a diversas facetas del dominio de la lengua. Para situarse en el nivel, el candidato tiene que demostrar que está en posesión o dominio de todos los atributos que

muestra el descriptor. Lo mismo sucede con el sistema de ASLPR. El *MCER*, frente a estos sistemas, es abierto e ilimitado, de naturaleza ilustrativa y no definitiva (Green, 2012: 44). Su aplicación al ámbito de la evaluación, sobre todo, certificativa, plantea, entre otros, el problema de hasta qué punto debe considerarse la respuesta (de comprensión o producción) del candidato más o menos exhaustiva, más o menos ceñida al descriptor para poder declarar que este está en un determinado nivel de la escala. Abre además el dilema acerca de qué descriptores pueden y deben considerarse en cada prueba, cómo deben ponderarse y si hay escalas más importantes que otras para determinar el nivel en cuestión; por último, cuántos descriptores deben tenerse en cuenta en un examen para que un candidato sea, finalmente, valorado a partir de la escala global (Green, 2010: 7; 2012: 44).

El capítulo 9 del *MCER* se dedica íntegramente a los usos del documento destinados a la evaluación. En el epígrafe 9.1, se indica que el *Marco* puede emplearse: «1. Para especificar el contenido de las pruebas y de los exámenes: *lo que se evalúa*; 2. Para establecer los criterios con los que se determina la consecución de un objetivo de aprendizaje: *cómo se interpreta la actuación*; 3. Para describir los niveles de dominio lingüístico en pruebas y exámenes existentes, permitiendo así realizar comparaciones entre distintos sistemas de certificados: *cómo se pueden realizar las comparaciones*» (*MCER*, 2002: 9.1). Con arreglo a esta declaración, está claro que el *MCER* opera respecto a los exámenes como sistema de referencia único al proporcionar la descripción detallada de la habilidad lingüística comunicativa, en diferentes estadios de su desarrollo, a partir de las categorías del llamado *enfoque centrado en la acción*: competencias, actividades comunicativas de la lengua y estrategias. Para determinar el grado de desarrollo de esta capacidad, es preciso contar con instrumentos de medida precisos y fiables, es decir, exámenes, capaces de extraer muestras de actuación suficientemente representativas como para erigirse en las evidencias de del grado de desarrollo de la habilidad que subyace a esa actuación.

La armonización deseada en los resultados de los diferentes sistemas de exámenes requiere que estas muestras se puedan comparar con unos estándares de contenido, que es, en definitiva, lo que se ofrece en el *Marco común europeo de referencia* en forma de descripciones de la actuación y del uso de las competencias y estrategias que subyace a esta. De esta forma, cada uno de los exámenes que se desarrollan en los diferentes contextos podrán seleccionar, agrupar, resumir, etc., los aspectos relevantes para cada contexto de los estándares de contenido que proporciona el *MCER* en sus 52 escalas. El uso de exámenes para evaluar las producciones de los candidatos y valorarlas en relación con un estándar absoluto

o un estándar de contenido se corresponde con lo que se conoce tradicionalmente como evaluación referenciada a un criterio (Fulcher, 2010: 225).

Desde esta perspectiva, cada uno de estos exámenes tendrá que ser capaz de situar a los candidatos en los diferentes puntos de esta escala de niveles comunes de referencia del MCER. Esto se realiza mediante un procedimiento denominado fijación de puntos de corte. Fijar uno o más puntos de corte significa establecer el límite entre dos categorías (Cizek y Bunch, 2007: 5), de manera que los candidatos a los exámenes quedan distribuidos con arreglo a estas. Cuando estos autores hablan de categorías se refieren a lo que tradicionalmente se entiende como las notas de un examen, que se transmiten bien en términos cuantitativos o cualitativos, a través de etiquetas (*apto*, *no apto*; *suspense*, *notable*, *sobresaliente*, etc.). Gracias a los puntos de corte, mediante el examen, se podrá discriminar a los candidatos que están por encima o por debajo de ese punto y que determinará, bien el *apto*, bien cada uno de los grados que se establecen por encima o por debajo de éste. El *apto* puede o no coincidir con uno de los estándares de contenido, es decir, puede corresponderse con uno de los niveles comunes de referencia (A1, A2, B1, etc.) o situarse a cierta distancia de estos. El número de puntos de corte y las categorías resultantes vienen determinados por el propósito del examen. La fijación de puntos de corte supone determinar cuál es la actuación mínimamente aceptable para que un candidato alcance determinado grado de dominio dentro de la escala. Esto es lo que se denomina *estándares de dominio*.

Como se puede inferir, este tipo de operación no se puede realizar de forma meramente intuitiva, sino que debe hacerse mediante la aplicación rigurosa de procedimientos de tipo cualitativo y cuantitativo, no exentos a veces de dificultad. El usuario de los sistemas de exámenes, sobre todo, cuando son certificativos, está en el derecho de conocer el rigor con el que han sido aplicados estos procedimientos.

En definitiva, el uso de un sistema de referencia externo como el del MCER precisa de la concienciación y de la capacitación de los miembros de la comunidad evaluadora respecto a la necesidad de alcanzar un entendimiento compartido de los estándares de contenido y de los estándares de dominio. Es necesaria, además, la profesionalización y especialización de los integrantes del colectivo.

2. Construcción de argumentos de validez

En los Estados Unidos de América, desde el año 1966, las tres organizaciones siguientes trabajan en el marco de un acuerdo de colaboración para elaborar los denominados *Standards for Educational and Psychological Testing* (estándares para la evaluación educativa y psicológica): American Educational Research Associa-

tion (AERA), American Psychological Association (APA) y National Council on Measurement in Education (NCME). He aquí una tercera acepción de la idea de *estándar*, que difiere de las presentadas en el epígrafe anterior. Se trata, en este caso, de una serie de líneas directrices y recomendaciones que garantizan el uso de buenas prácticas en el desarrollo, la interpretación y el uso de instrumentos de medición de constructos psicológicos y de evaluación educativa. Los estándares AERA, APA NCME se siguen y se respetan no solo en el país norteamericano, sino en todo el mundo. Se actualizan cada 15 años aproximadamente y la versión más reciente vio la luz en julio del año 2014. Esta última edición no presenta cambios sustanciales respecto a la anterior en lo que se refiere a la idea de validez: «grado en el que la evidencia y la teoría soportan la interpretación de las puntuaciones de los test, para el propósito con el que estos son utilizados» (AERA, APA, NCME, 2014: 13). Esta concepción de la validez se incardina en el paradigma de lo que se conoce como *modern validity theory* (teoría moderna de la validez), que es deudora de la visión que aporta Samuel Messick (1989). Según este autor,

Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and *appropriateness of inferences and actions* based on test scores and other modes of assessment.

Messick, S. (1989: 13).

La adopción de la teoría moderna de la validez comporta una serie de cambios de perspectiva. En primer lugar, la validez deja de verse como una característica del examen en sí mismo y pasa a entenderse como una propiedad de los usos e interpretaciones de las puntuaciones de los exámenes. De este modo, las notas de un examen pueden servir, es decir, ser válidas para determinados usos y dejar de serlo para otros. Una calificación de *apto* en un examen generalista de español puede servir para saber, siempre que se demuestre, si un candidato se encuentra, por ejemplo, por encima de la línea que separa el nivel B1 del nivel B2. El *apto* perdería su sentido cuando se trata de determinar, verbi gracia, si el candidato es el mejor para ocupar un puesto de trabajo en la función pública. Alderson (2007) argumenta por extenso acerca de la necesidad de investigación que avale aseveraciones, muy frecuentes en la publicidad de las organizaciones certificadoras públicas o privadas, como que un B2 es suficiente para cursar estudios universitarios o que para integrarse en la comunidad lingüística y cultural de una nación se precisa al menos un nivel A2. Esta visión de la validez enfatiza la dimensión social del impacto de los exámenes, en la medida en que los concibe como los instrumentos necesarios para arrojar resultados, a partir de los cuales se

tomarán decisiones en relación con los candidatos, algunas de ellas de gran calado (estatuto de ciudadanía, acceso a programas de estudios, puestos de trabajo, etc.) (Milanovic *et alii*, 2011: 16).

La validez se concibe desde la perspectiva de esta nueva óptica como un concepto unitario (Perron y Gillespie, 2015: 39). Lo que se entendía con anterioridad como tipos de validez se ve ahora como facetas de este concepto o fuentes de evidencia, que es preciso recoger a lo largo de las diferentes fases del desarrollo del examen y en distintos momentos de su ciclo de vida. Estas evidencias constituyen las bases científicas sobre las que se apoyan las interpretaciones de las notas de los exámenes.

La declaración de que las notas de un examen se corresponden con uno o con varios de los niveles de la escala del *MCER* requiere la construcción de un argumento de validez y, por consiguiente, la recogida de evidencias procedentes de diferentes fuentes. Fulcher (2010: 230) advierte del peligro que entraña el uso de sistemas de referencia externos como el del *MCER* cuando se pretende alinear a este todo tipo de exámenes. Su razonamiento lógico se basa en la idea de que es necesario que haya diferentes exámenes para los diversos propósitos. Las evidentes variaciones de un examen a otro se manifestarían en la estructura y en el contenido. Si todos los exámenes se pueden alinear con las escalas del *MCER* y sus resultados, es decir, lo que pueden hacer los candidatos se pueden interpretar en los términos de las escalas del *MCER*, la mencionada diversidad queda anulada, al igual que el propósito del examen. Este planteamiento se opone radicalmente a la idea de validez, tal y como ha sido descrita más arriba. Sobre esta base, lanza una fuerte crítica contra los exámenes generalistas, desarrollados sin propósito alguno, que declaran estar vinculados al *MCER* y que buscan reconocimiento en ámbitos tan dispares como la adquisición del estatuto de ciudadanía o el acceso a programas de educación superior. Por su lado, Milanovic y Weir (2010: x) argumentan en la dirección contraria y defienden la flexibilidad del *MCER*, gracias a la cual se pueden comparar múltiples sistemas entre sí.

A los problemas que apunta Fulcher (2010: 244–248), se suman las críticas que recibe la formulación de los descriptores ilustrativos del *MCER*, mencionadas más arriba y recogidas por Llorián, S. (2011: 15,16), que dificultan el uso de las escalas del *MCER* para la evaluación. Todo ello precisa el uso de procedimientos que medien entre los sistemas de exámenes y el *MCER*: procesos de validación cuyas evidencias permitan la construcción de un argumento de validez, banco de ítems, auditorías periódicas, tanto externas como internas, etc. Precisamente con la idea de guiar la construcción de un argumento de validez a lo largo de diversos procesos, el Consejo de Europa impulsa, en el año 2003, a partir

de una iniciativa del gobierno finlandés, muy sensible a todas estas cuestiones, el desarrollo del *Manual para relacionar exámenes con el Marco de referencia (MRE o Manual, de aquí en adelante)*. La versión definitiva se difunde a partir de 2009, una vez pilotado el mencionado borrador. De su contenido, nos ocupamos en el epígrafe siguiente.

3. El Manual para relacionar exámenes con el MCER del Consejo de Europa

3.1 El MRE: descripción general

El MRE se ha convertido en uno de los principales componentes de la *caja de herramientas (toolkit)* que el Consejo de Europa pone a disposición de los usuarios del *Marco de referencia* cuando se emplea con fines de evaluación. Se acompaña de un suplemento que contiene información técnica (*Reference Supplement to the Manual for Relating Examinations to the CEFR*, Takala, 2009) y de un documento que facilita información adicional (*Further Material on Maintaining Standards across Languages, Contexts and Administrations by exploiting Teacher Judgment and IRT Scaling*, North y Jones, 2009). Siguiendo la tradición del Consejo de Europa, se presenta con una declaración acerca de lo que contiene el documento frente a lo que no contiene para erradicar falsas expectativas. Por consiguiente, según sus propios autores, el MRE aporta procedimientos y herramientas para vincular los exámenes al MCER obteniendo evidencias de diversas fuentes (contenido, validación psicométrica, etc.). No es un instrumento destinado a guiar el desarrollo de exámenes. Con este fin, en el año 2011, el Consejo de Europa encarga la redacción de un segundo manual a ALTE (Association of Language Testers in Europe): Milanovic et al (2011). *A Manual for Language test Development and Examining for Use with the CEFR*, Estrasburgo, Consejo de Europa.

3.2 Procesos de vinculación descritos en el MRE

Los procedimientos del MRE se agrupan en cuatro procesos: familiarización, especificación, estandarización y fijación de puntos de corte, validación, que se describen sucintamente en los epígrafes sucesivos. Según el propio *Manual*, estos procesos se han recogido de la bibliografía especializada y dispuesto en el documento de manera que puedan ser aplicados a sistemas de exámenes de diferente alcance, cuenten o no con abundancia de recursos.

Las indicaciones que contiene el MRE son de carácter procedimental y práctico. Orientan al usuario, por un lado, acerca de los objetivos de las actividades y

de su fundamento; por otro lado, facilitan la implementación y puesta en marcha, a partir de una descripción muy detallada de la gestión de tiempos, espacios y personas, así como de los pasos que se deben seguir y de los recursos que se han de disponer en la preparación y en el desarrollo.

Cada organización certificadora, en virtud de sus características y recursos, seleccionará las circunstancias en que se implementarán las herramientas y procedimientos que proporciona, es decir, determinará lo siguiente: qué procedimientos entre los descritos necesita y está en condiciones de aplicar; quién o cómo configurará los paneles de jueces expertos que se requieren (número de personas, perfil, filiación institucional interna o externa, funciones, etc.), qué responsables se designarán...; cuándo, o sea, con qué frecuencia realizará análisis, convocará paneles y elaborará informes...; dónde (dentro o fuera de la propia organización, etc.); cómo programará las actividades involucradas en los procesos (fases del desarrollo de un examen, momentos del ciclo de vida de un examen...); cuánto, dicho de otro modo, en qué grado de intensidad y de frecuencia operará; para qué, o lo que es lo mismo, para rendir cuentas ante auditores externos, para reflexionar internamente acerca de la calidad del examen, etc.

3.2.1 *Procesos de familiarización*

En la representación gráfica que ofrece el *MRE* (*MRE*: 15) de los procesos de vinculación de los exámenes al *MCER*, el de familiarización se proyecta de un modo transversal sobre el resto. Esto quiere decir que los paneles de jueces deben realizar las actividades de familiarización antes de las sesiones en que se ponen en marcha el resto de los procesos, especialmente, los de especificación y estandarización. El conjunto de las propuestas tiene el objetivo de alcanzar un conocimiento profundo, y a la vez intuitivo, de los descriptores ilustrativos de las escalas del *MCER* que se emplearán en el resto de los procesos.

Las actividades que se proponen durante el desarrollo de las sesiones responden a la siguiente tipología: autoevaluación con las escalas destinadas a este fin; clasificación de textos recortados y de descriptores ilustrativos en los niveles, las categorías y las escalas correspondientes; relación de los descriptores ilustrativos con muestras concretas de actuación.

3.2.2 *Procesos de especificación*

Consisten en relacionar el examen con el *MCER* desde la perspectiva de la cobertura del contenido (*MRE*: 26). Es preciso recordar aquí que los procesos de vinculación pueden aplicarse a exámenes existentes, que se relacionan con el documento *a posteriori* o bien a exámenes que se han desarrollado a partir del

MCER. Conviene tener presente asimismo que la relación de los exámenes con el *Marco* no significa que estos den cuenta de todos los aspectos que se contemplan en los textos descriptivos y en las escalas. Por ejemplo, se podría relacionar con el *MCER* el contenido de un examen, destinado a constatar el grado de desarrollo de las competencias en español de aspirantes a operadores telefónicos. Lo lógico sería que no incluyera pruebas de expresión e interacción escritas y que se consideraran escalas de comprensión auditiva y de expresión e interacción orales. Podría darse el caso, por otro lado, que un examen abarque varios niveles, en virtud de las necesidades y características de las poblaciones candidatas. Por ejemplo, en un contexto determinado, pudiera ser que la prueba de comprensión lectura se situara en un nivel B2+ y el resto de las pruebas, en el nivel B2.

El capítulo 4 del *MRE*, en que se describen los procesos de especificación, proporciona procedimientos detallados y herramientas para llevar a cabo lo anteriormente descrito. Las herramientas mencionadas son una serie de formularios que se proporcionan en los apéndices del documento y unas parrillas para describir el contenido. Los procedimientos consisten fundamentalmente en cumplimentar los formularios y en usar los descriptores del *MCER* relevantes para cada una de las pruebas y relacionarlos con su contenido. La cumplimentación de los formularios se realiza a partir del análisis detallado de la documentación de examen y de un número significativo de formas empleadas en varias administraciones. Los paneles de jueces, internos o externos a la organización que promueve el examen, actúan, durante los seminarios convocados con este fin, a modo de auditores. Las sesiones de especificación, en la mayor parte de los casos, no se convocan para abarcar la totalidad del proceso, sino aspectos relativos a la descripción global o a pruebas concretas. Los formularios de especificación son de cuatro tipos. Se prestan, en consecuencia, a utilizarse en diferentes clases de auditoría, interna o externa, destinada, bien a la descripción general del examen o a la especificación del contenido.

El primer tipo de formulario, comprende la *Sección A2* de los apéndices (formularios 1 a 8) y se emplea para la descripción general del examen. Estos instrumentos están diseñados para funcionar como listas de comprobación de la calidad de los procesos que se ponen en marcha a lo largo de todas las fases de su ciclo de vida: planificación, redacción y pilotaje, calificación, fijación de puntos de corte, comunicación de resultados, análisis de los datos. El último formulario (*A8*), recoge la declaración acerca de la estimación inicial del nivel o niveles de la escala del *MCER* con los que se relaciona el examen. Uno de los cometidos de esta serie consiste en infundir en los responsables del examen auditado la toma de conciencia de que, si el sistema de exámenes no aporta evidencias de validez,

recogidas a lo largo de todo el ciclo de su desarrollo, no se puede relacionar con un sistema de referencia externo. Los aspectos que recogen los formularios son los que se consideran habitualmente en los códigos o directrices de buena práctica y en los sistemas de aseguramiento de calidad, a los que nos referiremos más abajo.

El resto de los formularios (del A9 al A24) se destinan a la especificación propiamente dicha. Se organizan en tres áreas: especificación de las actividades comunicativas de la lengua, especificación de las competencias comunicativas de la lengua y especificación del resultado general del análisis y trazado del perfil del examen. Como se ha indicado, cada contexto de examen organiza las sesiones en virtud de sus necesidades y de los recursos disponibles en la organización. Los formularios presentan un formato fijo para analizar las pruebas. En ellos, se solicita del panel que registre los aspectos de los capítulos 4 y 5 del MCER (situaciones, temas, textos, etc.) que se reflejan en las pruebas para, tras el análisis, determinar, a partir de las escalas relevantes, el nivel en que se sitúan las pruebas.

Pese al detalle de las descripciones del MRE, hay cuestiones que quedan sujetas a la ambigüedad la selección de los formularios. No está clara la cuestión de los mínimos o los criterios que se pueden emplear. Es asimismo algo confuso el procedimiento que se emplea para cumplimentarlos: ¿se ha de hacer a partir de las especificaciones, de las formas de examen, de ambos?, ¿rellena un formulario cada uno de los jueces?, ¿se rellena un solo formulario a partir del consenso?...

Además de los formularios, este haz de procesos cuenta con otro tipo de herramienta: las parrillas de análisis del contenido, mediante las cuales el panel de jueces puede, tras peinar sistemáticamente los diversos aspectos de las tareas contemplados en los descriptores de las escalas, estimar el nivel que se requiere por parte del candidato para resolverlas adecuadamente. Las parrillas analizan, por un lado, el material de entrada y, por otro, el reactivo propiamente dicho. Son de dos tipos: las que han diseñado para las tareas de evaluación indirecta y las que analizan las tareas de respuesta abierta. Las primeras proceden del *Dutch Construct Project* y las segundas se encargaron a la organización ALTE.

Las evidencias de validez derivadas de estos procesos pueden recogerse en sesiones independientes. Una alternativa consiste en convocar paneles que las realicen como paso previo a las actividades de estandarización, que se describen en el epígrafe siguiente.

3.2.3 Procesos de estandarización y fijación de puntos de corte

Los capítulos 5 y 6 del *Manual* se ocupan respectivamente de los procesos de estandarización (anclaje o *benchmarking* de las pruebas de examen en los niveles del

MCER) y de los procedimientos de fijación de puntos de corte o *standard setting procedures*. Estos últimos, descritos en el capítulo 6, constituyen las fuentes de evidencia más importantes.

La estandarización comprende dos subprocesos: la formación de los panelistas con muestras de actuación y tareas previamente estandarizadas, y el anclaje del examen o sistema de exámenes sobre el que se está trabajando. Se presentan por separado las actividades y los recursos que se precisan para trabajar con las pruebas de evaluación directa y para las pruebas de evaluación de aspectos formales del sistema (generalmente, gramática y vocabulario) y de comprensión de lectura y auditiva.

Las muestras de actuación (expresión e interacción orales y escritas) y las tareas (textos de entrada y reactivos) previamente estandarizadas se utilizan para formar a los paneles de jueces en la primera fase de los subprocesos de estandarización. El consejo de Europa pone a disposición de los usuarios, en su *caja de herramientas*, diversas fuentes de muestras orales y escritas a través de cederrones, además de recursos en línea cedidos por varias organizaciones. Es necesario apuntar que, para el español, el material es insuficiente. Para la expresión y la interacción orales se cuenta únicamente con las muestras facilitadas por el CIEP, organismo dependiente del Ministerio de Educación de la República de Francia, resultante de la celebración, en el año 2008 de un seminario interlingüístico en la localidad de Sevrès (Francia). Se compone de una serie de actuaciones grabadas en soporte vídeo, que se acompañan del informe de calificación de los paneles de jueces moderados que participaron en la valoración de actuaciones en diversas lenguas. Pese al indiscutible valor de este material, es necesario apuntar que ha sido objeto de críticas, puesto que las muestras corresponden únicamente a informantes de edades adolescentes (estudiantes de enseñanza secundaria). Se ha cuestionado bastante la validez de los informes de las muestras de los niveles C1 y C2. No existen muestras escritas estandarizadas disponibles, frente a lo que sucede para las lenguas europeas que cuentan con mayor número de hablantes: inglés, francés, alemán, portugués e italiano. Ninguna organización relacionada con la enseñanza, el aprendizaje y la evaluación de nuestra importantísima lengua se ha ocupado de estandarizar muestras escritas y ponerlas a disposición de la comunidad. El Consejo de Europa pone a disposición de los usuarios un cederón, en el se encuentran dos tareas de comprensión de lectura, con buenos datos psicométricos y de calibración, para los niveles inferiores de la escala, cedidos por DIALANG. No existe más material estandarizado a disposición pública.

Los procedimientos de formación propuestos en el *Manual* para las pruebas de calificación subjetiva consisten en que los jueces califiquen las muestras es-

tandarizadas siguiendo unas dinámicas claramente descritas. Se pide finalmente que se contraste los juicios con la respuesta correcta procedente de los informes resultantes de la estandarización previa. Respecto a las pruebas de calificación objetiva, el procedimiento es similar: en este caso, los paneles deben estimar cuál es el nivel mínimo que debe haber el candidato para procesar los textos de entrada y para responder correctamente a los reactivos asociados a este. Una vez que el panel ha sido debidamente entrenado, estará ya en condiciones de aplicar el procedimiento a las muestras y tareas del examen objeto de vinculación. Este procedimiento recibe duras críticas por parte de Fulcher (2010: 245-246), debido a que considera que el hecho de facilitar respuestas correctas impide que los juicios de los paneles sean objetivos y, por consiguiente, válidos y reales.

El capítulo 6 del *MRE* recoge los procedimientos referidos a la fijación de puntos de corte. Se provee al usuario de información exhaustiva acerca de los métodos descritos en la bibliografía para llevar a cabo esta operación y de los conceptos y procedimientos necesarios para ponerla en marcha. La descripción de los métodos toma como referencia principal la obra de Cizek y Bunch de 2007, con algunos añadidos. Este capítulo no existía en la versión piloto del *MRE* de 2003. Este tipo de información se proporcionaba, de hecho, en la Sección B del *Suplemento* (Council of Europe, 2004, Kadranieva, Section B). Los métodos se clasifican en el *Manual* en los tres grupos siguientes: centrados en los ítems del examen, para lo cual será necesario estimar la actuación mínimamente aceptable que se requiere del candidato para responder a los ítems del examen; métodos centrados en los candidatos, que dependerán de juicios previos que sobre estos habrán hecho los profesores antes del examen; métodos basados en los análisis psicométricos a partir de la Teoría de Respuesta al Ítem (TRI), para los cuales el panel de jueces jugará en los márgenes de una región umbral en la que coinciden sus juicios sobre la dificultad de los ítems con los datos del análisis psicométrico basado en la TRI. El *MRE*, además de describir con detalle los métodos, explica la función de los agentes implicados en la operación, en la que destaca la figura del moderador, encargado de facilitar la información normativa y la de impacto, es decir la del efecto que, en términos estadísticos, supondría tomar una u otra decisión acerca de la nota de corte o de las puntuaciones de las pruebas y del conjunto del examen que se corresponden con un nivel de la escala del *MCER*. El *Manual* deja claro que la decisión referida a los puntos de corte es, en ocasiones, más política que técnica y está fuertemente determinada por los factores del contexto del examen.

Finalmente, el documento se refiere a la fijación de estándares teniendo en consideración las diferentes pruebas de un examen. Trata también la cuestión del

mantenimiento de los estándares a lo largo de las diferentes administraciones y termina apelando al reto de relacionar exámenes de diferentes lenguas aplicados a una misma población mediante la fijación de estándares.

3.2.4 Procesos de validación

El último haz de procedimientos se concentra en el capítulo 7. Este se inicia con unas consideraciones iniciales, referidas a los requisitos que deben cumplir los exámenes para poder garantizar unos mínimos de validez. Se centran en la necesidad de validar el contenido y de realizar ensayos y pilotajes de las pruebas, de forma sistemática, con el fin de poder analizar psicométricamente los datos y de fijar los puntos de corte de forma rigurosa.

El resto del capítulo se organiza en tres grandes bloques, en cada uno de los cuales se describen las acciones y recomendaciones de naturaleza garantista. El primer bloque se centra en aspectos procedimentales, orientados a asegurar que los procedimientos descritos en las sesiones anteriores se aplican con el debido rigor.

El segundo bloque de procedimientos se centra en los procedimientos destinados a garantizar la validez interna durante el proceso de fijación de puntos de corte. Proporciona, sobre todo, técnicas para medir la consistencia de los juicios emitidos por los paneles. El tercer bloque se ocupa de aspectos de la validez externa, es decir de fuentes de evidencia externas al propio examen.

Los procedimientos de validación que se facilitan en esta sección no difieren de los que se pueden encontrar en la bibliografía especializada. Se trata de confirmar que los procedimientos anteriores se han llevado a la práctica correctamente.

4. Cuestiones de ética y de calidad

A comienzos de los años noventa del pasado siglo, empiezan a proliferar asociaciones que se preocupan y ocupan de la aplicación de prácticas garantistas al desarrollo y la administración de los exámenes de lenguas. Comienzan a afianzarse en Europa, importando la tradición norteamericana (Figueras, Puig, 2013: 30). Nos hemos referido ya, más arriba, a las que, en Norteamérica, regulan la medición de constructos y la evaluación educativa (AERA, APA NCME). Estas organizaciones, de naturaleza asociativa, promueven códigos deontológicos de buenas prácticas, en algunos casos, o establecen, en otros, estándares mínimos de aseguramiento de la calidad. Nos encontramos ante una nueva acepción del término *estándar*, que aquí se refiere a los requisitos mínimos que tienen que cumplir los exámenes para garantizar unos mínimos de calidad que, en el ámbito

de la evaluación mediante exámenes, es sinónimo, a la postre, de validez. La más antigua de todas las asociaciones, a escala internacional, es ILTA (International Language Testing Association). En Europa, en dos planos muy diferentes, se encuentran ALTE (Association of Language Testers in Europe) y EALTA (European Association for Language Testing and Assessment). Especial interés para el tema de este artículo reviste SICELE (Sistema Internacional de Certificación del Español como Lengua Extranjera).

La organización ALTE ha actualizado recientemente su sistema de aseguramiento de la calidad. Ha reforzado los procesos de auditoría y ha introducido el denominado *Q-mark*, que es un nuevo marchamo, indicador de la calidad de los exámenes de las organizaciones miembros. El aseguramiento de la calidad se realiza a partir de una lista de 17 estándares mínimos (*minimum standards*). Los estándares se complementan con un código deontológico, destinado a los agentes implicados en el desarrollo y la administración de los exámenes, por un lado, y a los usuarios, por otro. El quinto estándar, recientemente introducido, que se reproduce más abajo, hace clara mención a las evidencias de vinculación de los exámenes a sistemas de referencia externos como el del MCER.

Si se dice que el examen está vinculado a un sistema de referencia externo, por ejemplo el MCER, existen pruebas fehacientes de su alineación a este sistema.

<<http://www.alte.org>>

El texto del estándar es escueto como corresponde a este tipo de formulaciones. No se especifica el grado de exigencia en cuanto a la aplicación de los procedimientos de vinculación y tampoco se menciona el tipo de evidencias que se precisa aportar. Los documentos que se desarrollan para guiar las auditorías, que sí son muy detallados, se detienen en aspectos como la fijación de puntos de corte y la estandarización de tareas.

I. How do you establish your assessment scales and link them to the ALTE Framework / Common European Framework of Reference?

iii. If not, how do you address these issues? For example: how do you calculate the difficulty of the items / tasks used in the sub-tests in relation to the ALTE Framework / Common European Framework of Reference.

<<http://www.alte.org>>

El aseguramiento de la calidad de los exámenes se realiza a través de un proceso de auditorías, en este caso, internas y de pares. ALTE hace públicos los exámenes o sistemas de exámenes que han superado este proceso y que, por consiguiente, pueden exhibir el marchamo *Q-mark*. En la fecha de publicación de este

artículo solamente figura en la lista uno de los exámenes certificativos públicos de español como lengua extranjera, el DELE A2, que se encuentra en pleno proceso. No se sabe hasta la fecha si la ha superado o si está pendiente de aplicar recomendaciones.

El sistema SICELE dispone de página electrónica propia desde el año 2010. Uno de sus apartados se dedica al Sello SICELE, que se describe como el distintivo que la organización otorga a los exámenes o sistemas de exámenes, promovidos por organizaciones universitarias de los países de habla hispana, que hayan superado los procesos de verificación del cumplimiento de los estándares de calidad. La página no facilita información acerca de las características de los procesos de verificación de estándares. Se indica únicamente que los exámenes de las organizaciones miembros deben someterse a estos periódicamente. Puede suponerse que se trate de auditorías, internas o de pares, puesto que se alude a políticas de autorregulación. Hasta la fecha, cinco años después de haber presentado el Sello a través de la página electrónica, ninguno de los exámenes que se realizan en el seno de esta organización ostenta este marchamo de calidad.

Los estándares de calidad del SICELE recuerdan en su formulación y organización a los de la organización ALTE y a la lista de control que esta organización utiliza para guiar las auditorías internas. Cinco de los estándares del SICELE, diseminados por los cuatro apartados en los que estos se articulan, hacen mención a la vinculación con sistemas de referencia externos. Dos de estos cinco estándares se refieren a la necesidad de aportar evidencias de los procesos de vinculación y de justificar y hacer pública este tipo de declaración a través de informes. Sirva el siguiente texto de uno de estos estándares como botón de muestra:

Se elaboran y difunden informes sobre resultados, los procesos de validación del examen y su vinculación con sistemas de referencia externos o internos.

<<http://www.sicele.org>>

Sin lugar a dudas es la asociación EALTA la que más espacio dedica a los procesos de vinculación al MCER en sus *Directrices de buenas prácticas en evaluación*. Esta asociación opera en un plano muy diferente a las dos anteriores, sobre todo, porque sus objetivos no se orientan al aseguramiento de la calidad de los exámenes, sino al fomento de la reflexión en una dimensión formativa. Por esta razón, las *Directrices* están formuladas entre interrogantes y tienen como principal destinatario al profesor de idiomas, que puede ser miembro de EALTA a título individual. Un apartado completo se destina a la vinculación de exámenes con el MCER. Los interrogantes siguen muy de cerca los procesos descritos en el *Manual para relacionar exámenes con el MCER*. Al igual que las anteriores, se

insiste en el registro y en la publicación de las evidencias de los procesos que se han llevado a cabo.

Como se ha visto, las políticas de calidad del conjunto de estas organizaciones se sitúan en el plano de la regulación interna, la colaboración y la práctica reflexiva. No se especifica el grado de rigor y de profundidad, en virtud del perfil del examen, con el que se tienen que desarrollar y documentar los procesos de vinculación. En ningún caso se habla de auditorías externas o de terceros. Se alude a las evidencias, pero no queda claro cuántas ni de qué tipo han de ser.

5. Los exámenes de español vinculados al MCER

Al margen de estas asociaciones, el Comité de Ministros de los estados miembros del Consejo de Europa, en el año 2008, emitió una serie de recomendaciones acerca del uso del *Marco común europeo de referencia (Recommendation CM/Rec (2008)7)*. Respecto a la evaluación, en la número 4.5 y 4.6, se puede leer lo siguiente:

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

4.6. ensure that full information regarding the procedures applied in all tests, examinations and assessment systems leading to officially recognised language qualifications, particularly those used to relate them to the common reference levels (A1-C2) of the CEFR, is published and made freely available and readily accessible to all the interested parties;

Recommendation CM/Rec (2008)7

Lo que cabe preguntarse en este punto es cuáles de los exámenes públicos de ELE declaran que sus resultados se corresponden con uno de los niveles de la escala del MCER y en qué medida estos exámenes o sistemas de exámenes presentan evidencias públicas que den cuenta de los procesos de vinculación empleados y de sus resultados. No albergamos intenciones de exhaustividad respecto sistemas de certificación del español objeto de comentario, por lo que partimos de una serie de criterios de selección: nos limitamos a exámenes públicos, gestionados por organizaciones de países de habla hispana, administrados a gran escala,

realizados al margen de los programas de cursos ofrecidos por las organizaciones que los promueven. La información analizada y comentada es la que se facilita a través de las páginas electrónicas de los sistemas de certificación y se refiere a informes resultantes de los procesos de vinculación de los exámenes al MCER o a publicaciones de trabajos de investigación relacionados con esta cuestión.

En el continente americano, se localiza la siguiente lista de exámenes de ELE que se ajustan a las características mencionadas: el DUCLE (*Diploma Universitario de Competencia en Lengua Española*), de la Universidad Nacional de Rosario (Argentina) el CELU (*Certificado del Español Lengua y Uso*), a cargo de un consorcio de universidades públicas de la República Argentina y, en México, en el seno de la UNAM (Universidad Nacional Autónoma de México), el EPLE (*Examen de Posesión de la Lengua Española*) el CELA (*Certificado de español como Lengua Adicional*) y el EXELEAA (*Examen de Español Lengua Extranjera para el Ámbito Académico*). De todos los anteriores, solamente uno de los certificados, el CELA, en cinco niveles, establece correspondencias con los niveles del MCER (de A2 a C2). No hace públicas evidencias de los procesos de vinculación a través de sus páginas electrónicas. En el continente europeo se localizan los siguientes sistemas de certificación que responden a los criterios más arriba señalados: DELE (*Diploma de Español como Lengua Extranjera*) gestionado por el Instituto Cervantes por mandato del Ministerio de Educación del Gobierno de España; los certificados y diplomas de la Cámara de Comercio, Industria y navegación de Madrid, con fines específicos; los diplomas DIE (*Diploma Internacional de Español*) de la organización FIDESCU, administrados en línea, también con fines específicos; el *Diploma LETRA*, de la Universidad Nebrija, destinado a trabajadores inmigrantes; finalmente, el examen de salida doble ELADE (*Examen en Línea de Acreditación del Español*), B1/B2, del Centro de Lenguas Modernas de la Universidad de Granada, también en línea. Todos ellos, sin excepción, relacionan sus resultados con los niveles de la escala del MCER. Solamente uno, el diploma ELADE, de muy reciente aparición, presenta, a través de sus páginas electrónicas, evidencias públicas de vinculación del examen al MCER, referidas a los datos psicométricos. Es necesario señalar que, a la vista de la información que difunde públicamente, se trata de un sistema que cuenta con escasos recursos. Pese a ello, facilita información sucinta, que no se ajusta al formato del MRE, acerca de los criterios empleados para la selección de los ítems, sobre la base de las tablas facilitadas por North y Schneider (2009) y también acerca de los criterios y resultados empleados para la fijación de puntos de corte. Nos referimos finalmente a las denominadas *pruebas unificadas* de las Escuelas Oficiales de Idiomas, en relación con las cuales, algunas publicaciones como Figueras et al (2011: 113–140) dan cuenta de la relación que los exámenes establecen con el MCER.

La constatación de que las evidencias no se hagan públicas y visibles en la mayor parte de los exámenes y sistemas que responden a los criterios señalados no lleva implícita, ni mucho menos, la conclusión o la sospecha de que estas evidencias no existan. Se trata más bien de dar cuenta de un panorama que difiere radicalmente del de los exámenes que se realizan en el Reino Unido. Estos sistemas suelen reservar un espacio en el que recogen artículos que relatan procesos y resultados de trabajos de investigación destinados a argumentar la validez de diversas facetas del examen, entre las que se encuentra la relación de sus resultados con la escala de niveles del *MCER*. Este sería el caso de Ackermann, K. (2011) o Milanovic (2009) por limitarnos solamente a dos ejemplos entre las innumerables publicaciones disponibles.

Al margen de los países de habla hispana, nos referiremos, para finalizar, a un estudio, promovido y financiado por el Parlamento Europeo, publicado en el año 2013 (Broek, S. van den Ende, I., 2013), acerca de la implementación del *MCER* en los sistemas educativos del continente. En algunos de los países que se incluyen en la muestra seleccionada para el estudio, se imparte español como tercera o cuarta lengua. El capítulo que corresponde a la evaluación contiene una serie de críticas al documento (Broek y van den Ende, 2013: 30): los informantes declaran que el *MCER* presenta dificultades para aplicarse a los exámenes, que resulta difícil evaluar el constructo de las competencias y que no resulta fácil justificar los procesos de vinculación. La mayoría de los países seleccionados implementan el *MCER* en los exámenes. Sin embargo, las relaciones entre los resultados y los niveles del *MCER* carecen de evidencia empírica, al no disponer de estudios de investigación, salvo excepciones como las que se promueven desde instancias gubernamentales holandesas (Broek y van den Ende, 2013: 13). Parece, en resumidas cuentas, que la difusión pública de este tipo de información no es práctica habitual, salvo en el caso de las grandes agencias certificadoras privadas del mercado de la lengua inglesa, las entidades certificadoras europeas

6. Conclusiones y consideraciones finales

Tal y como ha quedado expuesto y argumentado más arriba, la situación general de los exámenes de ELE respecto a los procesos de referenciación a sistemas de estándares, en especial, al del *MCER*, es susceptible de algunas acciones de mejora, en términos muy generales. Esta mejora debería localizarse principalmente en dos áreas: la provisión de recursos a la comunidad evaluadora del español y, sobre todo, la visualización pública de evidencias de los procesos de vinculación de los exámenes al *MCER* u otros sistemas de referencia externos, a través de publicaciones que recojan trabajos de investigación o de las páginas electrónicas

oficiales de los exámenes en que se difundan los datos de validación o los informes de vinculación.

En relación con los recursos, urge la celebración de seminarios de estandarización interinstitucionales, en que participen paneles de jueces expertos de diversa procedencia, con el objetivo de poner a disposición pública muestras de actuación orales y, sobre todo, escritas, con los consiguientes informes de calificación, de diversos perfiles de candidatos, representativos de los destinatarios de los sistemas de exámenes de ELE. Es igualmente acuciante la necesidad de poner a disposición de la comunidad de evaluadores de ELE tareas estandarizadas, acompañadas de buenos datos de calibración. Lo idóneo sería desarrollar un banco común de reactivos, para todos los sistemas de habla hispana. Apuntamos finalmente a la importancia de incrementar por parte de las organizaciones formadoras la intensificación de actividades formativas centradas en el área de la evaluación a través de exámenes, con el fin de que los profesionales de ELE se conciencien de la importancia de su posible contribución y responsabilidad a la calidad de los exámenes.

7. Referencias bibliográficas

- AERA, APA, NCME (2014). *Standards for Educational and Psychological Testing*. Washington: AERA.
- ACKERMANN, K. (2011). PTE General. «Using the Common European Framework of Reference for Languages to rate test taker responses». *PTE Research Notes*. [http://pearsonpte.com/wp-content/uploads/2014/07/Aligning_PTEA_Scores_CEF.pdf]. Consulta, 1 de abril de 2015].
- ALDERSON, J.C. (2007). «The CEFR and the Need for More Research», *The Modern Language Journal* 94 (4): 559–663.
- ALDERSON, J.C. et alii (2004), «The development of specifications for item development and classification within *The CFR*». *Final Report of The Dutch CEF Construct Project* [<http://eprints.lancs.ac.uk/44/>]. Consulta: 1 de abril de 2015].
- ALDERSON, J. C.; FIGUERAS, N.; KUIJPER, H.; NOLD, G.; TAKALA, S. & TARDIEU, C. (2006). «Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of the *Dutch CEFR Construct Project*». *Language Assessment Quarterly* 3(1): 3–30.
- BROEK, S. VAN DEN ENDE, I. (2013). *The Implementation of the Common European Framework for Languages in European Education Systems*. European Parliament. [<http://www.europarl.europa.eu/RegData/etudes/etudes/join/2013/495871/>]

- IPOL-CULT_ET%282013%29495871%28SUM01%29_EN.pdf>. Consulta, 1 de abril de 2015].
- BROWN, J. D. (1995). *The elements of language curriculum*. Massachusetts: Heinle & Heinle Publishers.
- COUNCIL OF EUROPE (2007). *Expanded set of C1 and C2 descriptors*. Collected by Brian North, Eurocentres, to supplement the relatively small number of descriptors at these levels in the CEFR. [<http://www.coe.int/t/dg4/education/elp/elp-reg/Source/Key_reference/C1C2descriptors_EN.pdf>. Consulta, 1 de abril de 2015]
- COUNCIL OF EUROPE (2009). *Relating language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment (CEF)*. [<http://www.coe.int/t/dg4/linguistic/source/manualrevision-proofread-final_en.pdf>. Consulta, 1 de abril de 2014]
- COUNCIL OF EUROPE (2008). *Recommendation CM/Rec (2008)7 of the Committee of Ministers to member states on the use of the Council of Europe's Common European Framework of Reference for Languages (CEFR) and the promotion of plurilingualism*. (Adopted by the Committee of Ministers on 2 July 2008 at the 1031st meeting of the Ministers' Deputies). [<http://www.coe.int/t/dg4/linguistic/cadre_en.asp>. Consulta: 1 de abril de 2015].
- COUNCIL OF EUROPE (2005). *Draft Guide for the Production of RLD*. Version 2. Strasbourg: Language Policy Division.
- COUNCIL OF EUROPE (2004). *Reference Supplement to the Manual for Relating Language examinations to the CEFR. Section B*. [<<http://www.coe.int/t/dg4/linguistic/CEF-refSupp-SectionB.pdf>>. Consulta, 1 de abril de 2014]
- COUNCIL OF EUROPE (2003). *Relating language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment (CEF)*. A Manual, Preliminary Pilot Version. Strasbourg: Language Policy Division.
- CONSEJO DE EUROPA (2002). *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza y evaluación*, Madrid: Anaya, Instituto Cervantes, Ministerio de Educación. [<<http://cvc.cervantes.es/obref/marco>>. Consulta: 1 de abril de 2015].
- CIZEK, G. Y BUNCH, M. (2007). *Standard Setting*. London, Sage.
- DAVIDSON, F. Y FULCHER, G. (2007). «The Common European Framework of Reference (CEFR) and the design of language tests: A matter of effect». *Language Teacher* 40 (3): 231–241.
- GREEN, A. (2012). *Language Functions Revisited*. Cambridge: UCLES/ Cambridge University Press.

- GREEN, A. (2010). «Requirements for Reference Level Descriptions for English». *English Profhilele Journal*, 1. [<http://journals.cambridge.org>]. Consulta, 1 de abril de 2015].
- FIGUERAS, N. y PUIG, f. (2014). *Pautas para la evaluación del español como lengua extranjera*. Madrid: Arco Libros.
- FIGUERAS, N. & NOIJONS, J. (eds.) (2009). *Linking to the CEFR levels: Research perspectives*. Arnhem: Cito/EALTA.
- FULCHER, G. (2010). *Practical Language Testing*. London: Hodder Education
- FULCHER, G. (2004). «Deluded by Artifices? The Common European Framework and Harmonization». *Language Assessment Quarterly* 1(4): 253–266.
- FULCHER, G. (1999). «Assessing English for academic purposes: putting content validity in its place». *Applied linguistics* 20(2): 221–236, OUP.
- FULCHER, G. y DAVIDSON, F. (2007). *Testing and Assessment: an advanced resource book*. London and New York: Routledge.
- LLORIÁN, S. (2011). «Los retos de la certificación del español con fines específicos». *Revista Nebrija, de Lingüística Aplicada a la enseñanza de lenguas*, 12. [<http://www.nebrija.com/revista-linguistica>]. Consulta, 1 de abril de 2015]
- MARTYNIUK, W. (ed.) (2010). *Aligning Tests with CEFR*. Cambridge: Cambridge/UCLES University Press.
- MAGER, R. (1962). *Preparing Instructional Objectives*. Palo Alto, CA: Fearon Publishers.
- MESSICK, S. (1994). «The interplay of evidence consequences in the validation of performance assessments». *Educational Researcher* 23(2): 13–23.
- MILANOVIC, M. et alii (2011). *Manual for Language Test Development and Examining*. ALTE y Language Policy Unit of the Council of Europe.
- MILANOVIC, K. (2009). «Cambridge ESOL and the CEFR». *Cambridge Research Notes*, ISSUE 37. [<http://www.cambridgeenglish.org/images/23156-research-notes-37.pdf>]. Consulta, 1 de abril de 2015]
- MILANOVIC, M. y WEIR, C. (2013). «Series editors'note». En GALACZI y WEIR (eds.) (2013). *Exploring Language Frameworks*. Cambridge: Cambridge/UCLES University Press: ix-xii.
- NORTH, B. (2007). «The CEFR illustrative descriptor scales». *Modern Language Journal*, 91/4: 656–659.
- NORTH, B. y SCHNEIDER (2009). *Further Material on Maintaining Standards across Languages, Contexts and Administrations by exploiting Teacher Judgment and IRT Scaling*. Strasbourg: Council of Europe.

- PERRON, E. GILLESPIE, D.F. (2015). *Key Concepts in Measurement*. Oxford: Oxford University Press.
- VAN EK, J.A. (1986). *Objectives for Foreign Language Learning*. Vol. 1: Strasbourg: Scope.
- WEIR, C.J. (2005). «Limitations of the Common European Framework for developing comparable examinations and tests». *Language Testing* 22: 281–300.
- Takala, (2009). *Reference Supplement to the Manual for Relating Examinations to the CEFR*. Strasbourg: Council of Europe. [http://www.coe.int/t/dg4/linguistic/Source/ManualForewordContSectA_2009_en.pdf]. Consulta, 1 de abril de 2015].

Sistemas de estándares citados

- ACTFL Guidelines (American Council on the Teaching of Foreign Languages Proficiency). [www.actfl.org. Consulta, 1 de abril de 2015].
- ASLPR (Australian Second Language Proficiency Ratings). [<http://www.islpr.org>]. Consulta, 1 de abril de 2015].
- Can do statements de ALTE (Asociation of Language Testers in Europe). [<http://www.cambridgeenglish.org/images/28906-alte-can-do-document.pdf>]. Consulta, 1 de abril de 2015].
- CLB (Canadian Language Benchmarks). [<http://www.language.ca>]. Consulta, 1 de abril de 2015].
- ICAO (*International Civil Aviation Organization*). [www.icao.int]. Consulta: 1 de abril de 2015].
- NATO Language Standards. [<http://www.etsglobal.org>]. Consulta, 1 de abril de 2015].

Sistemas de exámenes certificativos citados

- Certificados y diplomas de la Cámara Oficial de Comercio e Industria de Madrid, IFE (Instituto de Formación Empresarial). [<http://www.camaramadrid.es>]. Consulta: 1 de abril de 2015].
- Certificados CELA (Certificados del Español como Lengua Adicional). [Certificados de Español como Lengua Adicional. Consulta, 15 de abril de 2015].
- Certificados CELU (Certificado de Español Lengua y Uso). [www.celu.edu.ar]. Consulta, 1 de abril de 2015].
- Diplomas DELE (Diplomas de Español como Lengua Extranjera). [<http://www.diplomas.cervantes.es>]. Consulta: 1 de abril de 2015].

- Diplomas del sistema DUCLE (Diploma Universitario de Competencia en Español como Lengua Extranjera), Universidad Nacional de Rosario (Argentina). [<http://www.ducle.unr.edu.ar/examenes.html>]. Consulta: 1 de abril de 2015].
- Diplomas específicos DIE de la organización FIDESCU (Diploma Internacional de Español) de la Fundación. [www.fidescu.org/diploma-d-i-e]. Consulta, 1 de abril de 2015].
- Exámenes EPLE (*Examen de Posesión de la Lengua Española (EPL)*). [www.cepe.unam.mx/eple]. Consulta, 1 de abril de 2015].
- Examen ELADE [http://www.cdm-granada.com/html/c_espanol/esp]. Consulta, 1 de abril de 2015].
- Exámenes EXELEAA (Examen de Español Lengua Extranjera para el Ámbito Académico). [<http://www.cepe.unam.mx/exeleaa>]. Consulta: 1 de abril de 2015].
- Diploma LETRA (Diploma de Lengua Española para Trabajadores Inmigrantes). [http://www.diplomalettra.com/quienes_somos]. Consulta: 1 de abril, de 2015].

Asociaciones de evaluadores citadas

- ALTE (Association of Language Testers in Europe). [<http://www.alte.org>]. Consulta: 1 de abril de 2015]
- EALTA (European Association for Language Testing and Assessment). [<http://www.ealta.eu.org>]. Consulta: 1 de abril de 2015]
- ILTA (International Language Testing Association). [<http://www.iltaonline.com>]. Consulta: 1 de abril de 2015]
- SICELE (Sistema Internacional de Certificación del Español como Lengua Extranjera): [<http://www.sicele.org>]. Consulta: 1 de abril de 2015].