La evaluación de lenguas por ordenador y en línea: el eLADE (examen en línea para acreditación de dominio de español)

Alfonso Martínez Baztán
Centro de Lenguas Modernas, Universidad de Granada

Aurora Biedma Torrecillas
Centro de Lenguas Modernas, Universidad de Granada

Lola Chamorro Guerrero
Centro de Lenguas Modernas, Universidad de Granada

Adolfo Sánchez Cuadrado
Centro de Lenguas Modernas, Universidad de Granada

Sonia Sánchez Molero
Centro de Lenguas Modernas, Universidad de Granada

Resumen

Este artículo va a comentar algunos de los aspectos innovadores de la práctica de la evaluación por ordenador y, en concreto, en línea con sus ventajas y algunos problemas todavía no resueltos, y va a centrarse en la explicación del diseño y uso del nuevo examen de acreditación de español del Centro de Lenguas Modernas de la Universidad de Granada, administrado enteramente en línea a través de la web 2.0. A lo largo del artículo vamos a mostrar algunos aspectos que deben tenerse en cuenta en el diseño de este tipo de exámenes y a discutir los resultados del examen y de su utilización. En él se exponen las circunstancias de la puesta en marcha del *eLADE* en el año 2012, el proceso de diseño y producción en línea, los protocolos de validación y estandarización que se siguen, así como el desarrollo de la cuarta y más reciente convocatoria. El objetivo es tanto contribuir a la difusión del primer examen en línea de dominio de español como promover la máxima transparencia respecto a todos los procesos relacionados con los exámenes, en consonancia con lo establecido en los protocolos internacionales de diseño y validación de exámenes de lenguas.

1. Introducción: la tecnología en la enseñanza de lenguas y la evaluación de lenguas por ordenador

Cada día aparecen nuevas herramientas informáticas ideadas con fines no solo sociales sino educativos y académicos, consecuencia del desarrollo de las nuevas tecnologías y de su demanda y aplicación a ámbitos de aprendizaje. Uno de los campos que tradicionalmente antes ha adaptado los nuevos sistemas y técnicas de enseñanza-aprendizaje suele ser la enseñanza de lenguas pues desde siempre ha hecho uso de métodos audiovisuales incorporados a los métodos de enseñanza casi como algo inelu-

dible; y estos usos se han seguido desarrollando durante este siglo con la irrupción de Internet y todas las posibilidades de acceso a materiales reales en el aula, y rápidamente se han incorporado a la docencia como herramientas indispensables para las buenas clases de aprendizaje y práctica de cualquier lengua extranjera. Como consecuencia de ello, Internet también se ha extendido a la evaluación donde ya venían utilizándose técnicas que solo son posibles debido al desarrollo informático (análisis estadísticos de fiabilidad y validez, de correlación, etc. –en principio, propias de las ciencias y extendidas a las ciencias sociales y a la enseñanza, aprendizaje y evaluación de lenguas–).

Podemos citar el *software SPSS* de IBM (*Statistical Package on the Social Sciences*) que se usa para el diseño de exámenes de lenguas de alto impacto, exámenes que tanto por el coste de realización como por las consecuencias de su uso social, académico o laboral que ocasionan, tienen la enorme responsabilidad de tener que asegurar unos resultados óptimos y un diseño todo lo libre de fallos y subjetividad que se pueda hacer. A este *software* (indispensable para toda institución que se dedique a la evaluación y quiera dotarse de herramientas fiables que den crédito a sus resultados) se le ha venido a sumar recientemente otro que completa los cálculos y soluciona uno de los problemas más grandes ante los que suele encontrarse todo diseñador de exámenes de idiomas: la incertidumbre respecto al verdadero nivel del examen, de sus tareas y de cada uno de sus ítems. Hablamos de un programa como *Winsteps*, el cual permite aplicar la Teoría de Respuesta al Ítem (*IRT* en sus siglas en inglés, *Item Response Theory*) y tratar otros aspectos de difícil solución cuando se depende solamente de las opiniones de expertos y de pilotajes con candidatos voluntarios. Estos aspectos son, por ejemplo, el margen de error de las preguntas, las actividades o las tareas a la hora de medir el conocimiento y nivel de los estudiantes/candidatos, o el grado en que todo lo diseñado, planificado y acordado por los grupos de expertos efectivamente produce los resultados para los que se había pensado. Todos hemos pasado por la experiencia de encontrarnos, al preguntar a un grupo de expertos su opinión respecto a un ítem o sobre qué es lo que verdaderamente estaríamos midiendo con una actividad, con que algunos de ellos opinan que eso sobre lo que se les pregunta es de nivel A2 frente a otros que a lo mismo le asignan un nivel B1 o B2, incluso C1 (Consejo de Europa, 2011). El análisis estadístico de los exámenes mediante la Teoría de Respuesta al Ítem permite asegurarse de ello fehacientemente (para una descripción del uso de estos sistemas, *IRT* y *SPSS*, ver Green, 2013).

Estas y otras herramientas sirven para hacer cálculos y estadísticas sobre ítems, así como para diseñar exámenes en formatos nuevos que permiten el control adaptativo de las preguntas, la autocorrección o el envío inmediato de resultados; permiten incluso corregir y calificar pruebas de respuesta abierta como la expresión escrita y la expresión oral, mediante muestras comparativas, campos léxicos o algoritmos complicados, imposibles de hacer manualmente. Otros desarrollos tecnológicos, basados en en el ordenador y ahora también en Internet, permiten no solo la evaluación de lenguas por ordenador sino la evaluación en línea a través de la web, con notables ventajas para los usuarios y administradores, de las que más adelante hablaremos. Esto último, la evaluación en línea, es una posibilidad cada día más real, también para el español, que debe ser tenida en cuenta porque, como se dice en Douglas (2010: 139), "no puede ser razonablemente rechazada. La tecnología se está usando y va a continuar usándose (...) como

un ineludible aspecto de la evaluación de lenguas" o según Dunkel (1999: 77) "como una evolución natural en la práctica de la evaluación". Incluso, y debido tanto a las posibilidades que se abren en cuanto a formatos, mejoras en los aspectos de autenticidad y motivación y a una suma de características provenientes tanto del propio campo de la evaluación de lenguas como de las tecnologías, (la tecnología) está llamada a provocar una transformación en las prácticas de evaluación (Chalhoub-Deville 2010) o cuando menos notables mejoras en sus sistemas de medición. No obstante, el uso de la tecnología no es la solución a todos los problemas y también conlleva ventajas y desventajas y obliga a seguir estando alerta respecto a los usos y aplicaciones y a si estos producen los efectos previstos o tiene otros nuevos impredecibles. Como dice Alderson, no debemos tener una fe ciega en la tecnología y pensar que ella es la solución a todos los problemas (Alderson 1998). Los nuevos problemas a los que nos vamos enfrentar son de dos tipos, de índole humana y de naturaleza tecnológica (Brown 1997), que deberán tenerse en cuenta y analizarse en cada caso. La tecnología no es *per se* la solución directa a ningún problema sino una herramienta a utilizar para hallar soluciones y producir mejoras y, debido a su facilidad de uso y al creciente manejo de esta entre profesores y estudiantes, hay que seguir recordando la necesidad de usarla con responsabilidad y profesionalidad, dentro de lo que exige una evaluación con garantías, válida, fiable y justa (Martínez Baztán, 2011) pues hace falta saber que "no todos los profesores que crean y suben a Internet exámenes para uso público son conscientes de muchas de las dificultades del diseño y de la evaluación de lenguas, las cuales, no por utilizar ordenadores, se solventan automáticamente. Adoptar, de entre todos los modelos existentes, un examen de lenguas u otro y diseñarlo de una manera o de otra, sigue siendo una tarea que requiere cuidado, incluso si este solo tiene el propósito de usarse para funciones de bajo impacto" (Winke 2008: 358).

El primer uso de ordenadores en la evaluación de lenguas data de 1935 cuando la compañía IBM utilizó el modelo 805 de computador "para calcular notas objetivas en exámenes en Estados Unidos con el fin de facilitar la intensa labor y reducir el costoso trabajo de calificar millones de exámenes que se hacían cada año" (Mubarak Pathan 2012: 31). Han pasado muchos años ya pero la necesidad de evaluar a grandes cantidades de personas de forma económica y eficaz es cada día mayor. Por ello la búsqueda y desarrollo de nuevas herramientas que la faciliten no ha parado y podemos decir que ha sufrido un enorme empujón que la ha hecho más sencilla gracias al desarrollo de la tecnología acaecido a partir de los años 80. Tal es así que a día de hoy nos encontramos con numerosos exámenes para evaluar lenguas y entre ellos (como uno de los más nuevos; de hecho, el primero en español) el *eLADE*. Estos exámenes difieren en formatos e intenciones (los hay progresivos –que presentan las tareas y preguntas moduladas en virtud del nivel de cada respuesta dada por el candidato– o lineales –con todas las preguntas una tras otra diseñadas según el nivel adecuado a sus objetivos–) pero tienen en común el deseo de adaptarse a las necesidades de la sociedad y aspirar a generar la confianza y credibilidad sin la cual el desembolso que suponen sería una ruina. Por citar los más conocidos internacionalmente, por su popularidad y prestigio, encontramos el *TOEFEL iBT* (que mide la habilidad de hablantes no nativos de inglés para desenvolverse en tareas académicas de nivel universitario utilizando el inglés), el *BULATS* (diseñado para evaluar el dominio de inglés para empleadores de negocios, solicitantes de

empleo y candidatos para cursos de lengua de inglés de negocios), el *Basic English Skills Test* (*BEST*) *Plus* (para medir las destrezas de expresión oral y de comprensión auditiva de estudiantes de inglés en los Estados Unidos), el *COMPASS ESL Placement Test* (para evaluar el conocimiento del inglés americano en estudiantes de inglés en escuelas e instituciones americanas de bachillerato), el *Versant English Test* (para medir habilidades de expresión oral en inglés de hablantes no nativos de inglés) o el *Pearson Test of English* (*PTE*) *Academic* (para evaluar el dominio de inglés de estudiantes internacionales en contextos académicos).

Y en español, aunque es un campo novedoso donde no encontramos tantos como en el mundo anglosajón, además del *eLADE* que nos ocupa en este artículo, podemos nombrar (ver Winkel, 2008: 359) el *STAMP* (*Standards-based Measurement of Proficiency*), diseñado por el *Center for Applied Second Language Studies* (*CASLS*) de la Universidad de Oregón, que mide las destrezas de lectura, escritura y expresión oral según la escala de niveles de ACTFL inicial bajo, intermedio bajo e intermedio medio (lo que viene a ser A1, A2 y B1, Martínez Baztán 2008) o el desarrollo que cinco instituciones americanas (ACTFL, Brigham Young University, CAL, Defense Language Institute, y la Universidad de San Diego) están haciendo de un sistema llamado *Computer-Assisted Screening Tool* (*CAST*). A estos se ha venido a añadir recientemente el nuevo examen del Instituto Cervantes hecho en colaboración con la Universidad Autónoma de México (*SIELE*).

Ya hemos advertido –y nos hemos hecho eco de las palabras de Alderson– de que no debemos tener una fe ciega en la tecnología que nos confunda respecto a nuestra responsabilidad de demostrar que nuestros exámenes son verdaderamente válidos y fiables. No por la inevitabilidad del desarrollo tecnológico debemos olvidar que es necesario investigar sus resultados (Brown 1997), así como analizar los problemas que pueda provocar. El uso de la tecnología tiene indudables ventajas pero también puede generar algunas desventajas. Entre las ventajas (según Mubarak Pathan 2012) encontramos que los exámenes administrados por ordenador son más consistentes y fiables porque todos los candidatos reciben exactamente la misma información y debe realizar las mismas acciones sin influencias externas del medio en que las realizan; son más motivadores y facilitan una actitud más positiva; y administrativamente se simplifica la burocracia de envíos de materiales, reenvíos de exámenes una vez realizados, acceso a la información y a ejemplos o muestras de exámenes. Otros autores señalan que el uso adecuado de herramientas multimedia puede mejorar la autenticidad de los materiales y la interacción real con ellos, ya que puede ayudar al uso de situaciones comunicativas de la lengua en muchos aspectos (Douglas 2010: 118) o que las habilidades de audio y vídeo son más válidas y auténticas (Winke 2008).

Entre los problemas que señala la literatura al uso destacan principalmente los relacionados con aspectos de la validación y fiabilidad, especialmente en exámenes progresivos (Brown 1997): la adecuada selección de ítems (problema que no es ajeno a los exámenes tradicionales hechos en papel y que requiere de equipos de expertos y uso del software del que ya hemos hablado); los sistemas para establecer los puntos de corte y de hacer llegar estos de forma inteligible a los usuarios; la comparación

entre los resultados de un examen en papel y otro realizado por ordenador y el grado en que el formato provoca la medición de diferentes constructos (Winke 2008), con la posibilidad de que el conocimiento del uso de la tecnología influya en la obtención de la calificación (Madsen 1991) y –por tanto, nuevamente un problema de validez de constructo–; o la duda de si estos exámenes miden la destreza en el uso de ordenadores o la habilidad lingüística.

Para terminar esta introducción sobre los exámenes por ordenador y sus aspectos más significativos, vamos a señalar algunas de sus principales características y situar el *eLADE* dentro de ellas: a. la direccionalidad permite diseñar exámenes que ofrezcan los ítems de forma lineal, o progresiva, esto es, dependiendo de la respuesta dada a un ítem previo (el *eLADE* es un examen lineal); b. el formato de envío puede ser simplemente por ordenador o a través de la web (*eLADE* se realiza en línea a través de la web); c. la densidad de empleo de las herramientas multimedia, que hacen que el examen utilice un solo medio o los integre en formato multimedia (*eLADE* es multimedia); d. los objetivos de las tareas, bien sean mediante tareas independientes o que integren habilidades (*eLADE* utiliza tareas independientes); e. el sistema de calificación, humano o automático (*eLADE* utiliza ambos según las destrezas); f. el riesgo del examen, de alto, medio o bajo impacto (*eLADE* es de riesgo medio pues sus usos no conllevan consecuencias de alta responsabilidad); g. el propósito, diseñando exámenes relacionados con un currículo o no (*eLADE* es un examen de dominio no relacionado con un currículo sino con niveles de dominio de la lengua); h. los tipos de respuesta, seleccionada o abierta (*eLADE* tiene ambos tipos de respuesta en sus tareas); e i. los tipos de tareas utilizadas (*eLADE* utiliza varios tipos de tareas que se describen más adelante en este artículo).

2. ¿Por qué un nuevo examen de español y en línea? Justificación del *eLADE*

La creciente demanda para acreditar el conocimiento de las lenguas más habladas en el mundo hace necesaria la ampliación y diversificación de las ofertas de exámenes de dominio reconocidos por las instituciones académicas y laborales. Estos exámenes deben adaptarse a diferentes contextos y a las necesidades de los distintos usuarios potenciales, atendiendo a los nuevos intereses tanto de los usuarios como de las instituciones, que además deben velar por su buen uso, validez, operatividad y fiabilidad.

Con el Plan Bolonia y con la reciente implantación del Programa *Erasmus Plus* la demanda de exámenes de acreditación de dominio de español ha aumentado considerablemente, extendiéndose por todo el mundo y haciendo necesarias nuevas herramientas que posibiliten un buen acceso a tests fiables, validados, de sencilla administración y operativos. La Universidad de Granada se cuestionó estos aspectos ante la importante demanda de los estudiantes de los diversos programas que deseaban seguir estudios de Grado o Posgrado y que procedían de muy diversas partes del mundo (Biedma *et alii*, en prensa). La universidad se enfrentaba al problema de la certificación adecuada de dominio suficiente de lengua española para seguir dichos estudios o para conceder becas a los estudiantes de otros países, para lo cual estableció, al igual que el resto de las universidades europeas, unos criterios de nivel que debían acreditarse debidamente. De

ahí que el Vicerrectorado de Relaciones Internacionales de la Universidad de Granada encargara en el año 2010 a su Centro de Lenguas Modernas (CLM) la elaboración de un examen binivel que acreditara B1 o B2 de español, tarea que desarrolló el Equipo de Evaluación y Acreditación del CLM, expertos en evaluación de lenguas, siguiendo los estándares y protocolos más exigentes de diseño y validación de exámenes internacionales y según las directrices del MCER, EALTA, ALTE y los manuales del Consejo de Europa (CoE, 2002; CoE, 2009; CoE, 2011; ILTA, 2000; ALTE, 2001; EALTA, 2004). A partir de 2013 se procedió a realizarlo en línea con el fin de que se pudieran presentar personas desde cualquier parte del mundo sin necesidad de desplazarse, así como de aprovechar las ventajas de este tipo de examinación. Se puede acceder a toda la información en: www.clm-granada.com/html/c_espanol/esp/ex amenes/examen_acreditacion/examen_acreditacion.htm, donde hay también muestras del examen.

El examen posee la certificación de garantía de la Asociación de Centros de Lenguas de Enseñanza Superior (ACLES) así como por CERCLES (2012) y por la CRUE (2014) y a él se pueden sumar cuantas organizaciones e instituciones lo deseen, haciéndose sede examinadora según los requisitos establecidos en la página web: www.clm-granada.com/html/c_espanol/esp/examenes/examen_acreditacion/examen_acreditacion_solicitud_sede.htm.

Al conocer la existencia de un examen en línea con estas garantías de fiabilidad y validez, aumentó la demanda de otras instituciones no universitarias, tales como centros de enseñanza secundaria, otros centros de idiomas, así como personas en general interesadas en acreditar su nivel de lengua por la facilidad que ofrece el formato en línea del examen. Debido a esto, se amplió el uso y las funciones del examen ampliando los contextos de uso y los contenidos tanto para ámbitos académicos como sociales y convirtiendo el test en uno de uso abierto a todo tipo de usuarios.

Su administración y los resultados obtenidos en las cuatro convocatorias realizadas hasta el momento han sido excelentes, mostrando los usuarios sociales y académicos gran interés y manifestando su agrado con este nuevo formato de administración y asesoramiento que les facilita la labor de evaluar, administrar, gestionar y acceder a un certificado de nivel de español válido, fiable, operativo y acreditado. Entre lo que opinan los candidatos cabe destacar que el uso de elementos multimedia para evaluar la habilidad audiovisual es considerado como muy positivo. Asimismo, en relación al uso de audio y vídeo en línea, los usuarios no señalan problemas de operatividad ni audición en comparación con lo que suelen manifestar cuando se realizan pruebas audiovisuales en papel. Otro elemento valorado positivamente es la posibilidad de gestionar el tiempo de que dispone en las tareas de la prueba de comprensión lectora. Igualmente se señala la tranquilidad que produce el hecho de controlar el tiempo restante para cada tarea y prueba, y especialmente –señalan algunos candidatos– la posibilidad de gestionar la pausa al terminar las prueba de CA y CL y antes de comenzar la de EIE.

Finalmente, muchos candidatos valoran el hecho positivo de ser un examen binivel y la posibilidad de guardar el resultado de una nota de nivel mayor que la obtenida en

el certificado, con la posibilidad de presentarse al nivel más alto (B2) sabiendo que ya tiene una parte de esa prueba superada; esto se guarda durante dos convocatorias. El hecho de ser binivel garantiza a casi todos los candidatos la obtención de un certificado, ya sea de B1 o de B2 (salvo a los que se quedan en nivel A2 o menos y no obtienen ninguna certificación).

En cuanto a los administradores de las pruebas y los centros meta de estas, cuando se utiliza como medio de admisión, hasta el momento no han manifestado disconformidades ni comentarios de mejoras relevantes. La Universidad de Granada y el Centro de Lenguas Modernas, principales usuarios de este examen por el momento, están muy satisfechos tanto por su operatividad como por sus resultados. Lo principal ha sido el cambio ofrecido a los estudiantes que antes debían venir a España sin saber si finalmente obtendrían un nivel suficiente para cursar los estudios que les interesaban y ahora ya vienen admitidos y seguros de su certificación. Otro aspecto reseñable ha sido el ahorro en tiempo que supone administrar los exámenes en línea pudiendo acceder a los materiales rápidamente y a sus resultados de forma fiable y rápida. La existencia de este nuevo examen en línea ha cumplimentado adecuadamente y ha satisfecho todas las expectativas respondiendo de forma óptima a una necesidad social y académica que puede beneficiar a todas las universidades españolas.

3. Descripción general del examen

El eLADE es un examen binivel en línea que acredita, por el momento, el dominio del candidato en los niveles B1 o B2. Consta de cuatro pruebas diseñadas para evaluar seis actividades comunicativas de la lengua: la compresión lectora, la comprensión auditiva, la expresión e interacción escritas, y la expresión e interacción orales, siempre definidas de acuerdo al MCER. Tiene una duración de 3 horas y 15 minutos.

En la siguiente tabla se especifican el número de tareas de cada prueba, así como su duración.

Comprensión auditiva	**Comprensión lectora**	**Expresión e Interacción escrita**	**Expresión e Interacción orales**
5 tareas auditivas y/o audiovisuales.	5 tareas de lectura.	2 tareas de escritura.	3 tareas de expresión e interacción orales.
45 minutos.	75 minutos.	60 minutos.	10-15 minutos.
En línea.	En línea.	En línea.	*Skype* o en el CLM.

3.1. Comprensión auditiva

La prueba de comprensión auditiva (CA) consta de 5 tareas auditivas y/o audiovisuales con fragmentos es de entre 2 y 5 minutos que se escuchan dos veces. Al menos dos son de B1 y dos de B2. La prueba, de acuerdo con las normas de ACLES, consta de 25-40 ítems (de 5 a 8 por tarea). Los fragmentos están extraídos de noticias, reporta-

jes, conferencias, cuñas publicitarias, en los que se describe, argumenta, expone o se dan instrucciones. El registro de lengua es estándar (informal, formal y/o académico).

Los tipos de ítems son:

a. Respuesta alternativa o múltiple, de tres o cuatro opciones.
b. Relacionar enunciados con partes de la audición.
c. Respuesta corta (máximo de cinco palabras).

3.2. Comprensión lectora

La prueba de comprensión lectora (CL) incluye 5 tareas con textos que pueden variar de extensión entre 250 y 900 palabras según sean de B1 o B2. Al menos dos son de B1 y dos de B2. La prueba, de acuerdo con las normas de ACLES, consta de entre 25-40 ítems (de 5 a 8 por tarea). Los textos elegidos son tanto del ámbito privado como público, escritos en un registro estándar (informal, formal y/o académico) intentando que respondan a géneros textuales diferentes –correos, artículos, editoriales, reseñas, anuncios, guías, instrucciones, entre otros– a diversas funciones de la lengua y de variedad temática.

Los tipos de ítems son:

a. Respuesta alternativa o múltiple, de tres o cuatro opciones.
b. Respuesta corta (máximo de cinco palabras).
c. Respuesta verdadera (V), falsa (F) con justificación de la respuesta.
d. Reintegración de un fragmento a un texto.
e. Relacionar o emparejar.

3.3. Expresión e interacción escritas

La prueba de expresión e interacción escritas (EIE) consiste en la redacción de 2 textos de entre 200 y 350 palabras para el primero, de nivel B1 y entre 300 y 450 palabras para el segundo, de nivel B2. El primero puede ser en formato de carta o correo electrónico (informal o neutro) o narración, siendo un texto descriptivo, narrativo o expositivo; el segundo, carta o correo electrónico formal, artículo, informe o ensayo, pudiendo ser también argumentativo.

3.4. Expresión e interacción orales

La prueba de expresión e interacción orales (EIO) se compone de tres tareas: las dos primeras diseñadas para evaluar el nivel B1 y la tercera el nivel B2. La tarea 1 consiste en una entrevista, de entre 3-5 minutos, sobre cuestiones personales relacionadas con el candidato. La tarea 2, de entre 4-6 minutos, es una interacción en la que el candidato debe ser capaz de desenvolverse para resolver un conflicto o solicitar ayuda para solucionar un problema, eligiendo al azar entre tres opciones. La tarea 3 es una presentación o monólogo, de entre 4-6 minutos. Para facilitar que pueda acercarse a un tema

de su interés o especialidad, se elige libremente un tema entre tres alternativas y se prepara durante 10 minutos antes de comenzar la sesión de evaluación oral. El candidato debe exponer y argumentar su punto de vista, analizar el asunto y evaluar los pros y contras al respecto. De ella se pasa a una breve discusión con el entrevistador de unos 3-5 minutos de duración.

4. Desarrollo del examen y pilotaje: aspectos técnicos, validación, administración y certificaciones

Todo examen debe someterse a criterios y protocolos estrictos de validación que demuestren que evalúa lo que pretende. Igualmente, un examen administrado por ordenador tiene la responsabilidad de demostrar su fiabilidad y validez, con independencia de la credibilidad y validez aparente que los usuarios le otorguen. El examen *eLADE* se somete durante todo el proceso de planificación y diseño a protocolos y directrices de buenas prácticas para asegurar: a) su adecuación y anclaje al MCER, b) que es justo y fiable, c) que posee el nivel apropiado y d) que evalúa lo que se pretende, esto es, el dominio del español en los niveles B1 o B2 (Alderson 1998; Bachman 1990; Bachman 1996; Bachman, 2004).

Para las pruebas de CA y CL, el diseño y planificación de tareas e ítems es sometido a análisis críticos con expertos, nativos y candidatos similares al candidato meta y se siguen los procesos establecidos para asegurar la validez de contenido y apariencia de las tareas: elección de los textos de partida, audiciones y textos escritos, mapeo en grupo de expertos para identificar la información relevante sobre la que diseñar los ítems, diseño de las tareas y revisión de los ítems. Posteriormente se someten a un pilotaje formal estadístico con *SPSS* e *IRT*. Las pruebas de EIE y EIO se elaboran partiendo de un análisis de contenido y funciones según el contexto de uso de la lengua y las habilidades necesarias en los niveles B1 y B2. Las tareas se confeccionan considerando que su realización sea operativa para los candidatos, poniendo especial cuidado y atención para que no se deslicen elementos que favorezcan o penalicen a ningún tipo de candidato. Todas ellas se pilotan con grupos de candidatos similares al candidato meta, expertos y nativos para comprobar tanto la adecuación de los textos y su nivel, como la claridad y capacidad de motivación de las instrucciones, así como para ver si provocan el lenguaje esperado (escrito u oral) y en calidad y cantidad suficientes como para poder extraer de los datos de discurso obtenidos lo necesario para elaborar una buena y justa calificación, basada en pruebas (no en meros indicios) de existencia de las habilidades, al aplicar las escalas y criterios determinados al efecto. A continuación ofrecemos los más relevantes de estos datos (para ver cada uno de estos conceptos remitimos a la bibliografía sobre evaluación de lenguas: Alderson 1998; Bachman 1990; Bachman 2004; Consejo de Europa 2009; Consejo de Europa 2011; Martínez Baztán 2011).

4.1. Fiabilidad

Las cifras para asegurar que una prueba tiene una buena fiabilidad no deben estar por debajo de .7 (0.7), siendo mejores cuanto más próximas están al 1. En la convocato-

ria de junio de 2015 la fiabilidad fue de .844 para la prueba de comprensión auditiva y de .892 para la prueba de comprensión lectora.

4.2. Discriminación

Una buena discriminación (en datos de *SPSS*) está entre .250 y 1.00, siendo mejor cuanto más alta sea (Green 2013). Los ítems con los que finalmente se componen nuestros exámenes se sitúan en unos índices de entre .250 y .800.

4.3. Características de los niveles B1 y B2 que mide el examen

Las escalas y descriptores utilizados para la confección del examen proceden de las escalas del MCER para cada una de las actividades de la lengua que se miden (comprensión audiovisual y comprensión escrita, expresión e interacción orales y expresión e interacción escritas) tanto para el nivel B1 como para el B2. Remitimos al MCER o –más concretamente– a la página web del *eLADE* (ya citada) para más información y detalle sobre la descripción de los niveles, los tipos de textos que se deben entender o producir, los temas de interacción o monólogo para la expresión oral y escrita, el grado adecuado de coherencia o cohesión en la expresión escrita, o las situaciones ante las que se debe poder desenvolver el candidato.

4.4. Garantía sobre los niveles de los ítems y lo que mide el examen

Para asegurarnos de que el nivel de los ítems, tareas de las pruebas y el examen en su totalidad evalúa el dominio tanto de candidatos de B1 como de B2 nos basamos en los datos del análisis de IRT, un método que permite decidir varias cuestiones importantes para la composición del examen: el nivel de cada ítem, su margen de error de medida, su poder de discriminación, la cantidad de niveles que se pueden evaluar mediante esa batería de ítems, la unidad de la prueba (o su capacidad para evaluar las facetas deseadas, eliminando los ítems que se salgan del modelo general del test), así como el establecimiento de los posibles puntos de corte entre los dos niveles medidos (tarea esta que, en última instancia, se realiza por un grupo de expertos en una sesión de fijación de puntos de corte utilizando su conocimiento y experiencia así como los datos estadísticos). El nivel de los ítems y su relación con los niveles del MCER se establece mediante los dígitos establecidos en North y Jones (2009), donde se considera que los ítems para evaluar el nivel B1 deben tener un "logit" de entre -1.23 y 0.72 (0.72 ya es B2) y para evaluar el B2, entre 0.72 y 2.80 (2.80 ya es C1). Para el diseño final del examen se seleccionan aquellos que tienen el nivel adecuado y un nivel de error de medida aceptable.

Con los resultados del pilotaje y análisis de ítems y tareas de *SPSS* y *IRT* (fiabilidad, discriminación, error de medida, etc.) se reforma y diseña finalmente el examen, seleccionando los ítems que tienen el nivel apropiado para conseguir un test binivel (B1 y B2) suficientemente significativo y con los que se puedan obtener datos con los que elaborar una calificación final justa. Usualmente elaboramos unos 100 ítems de CA y CL y tras el pilotaje se seleccionan entre 25-40 de los mejores para cada destreza,

desechándose los que no poseen el nivel de dificultad adecuado, los que no tienen una capacidad de discriminación óptima, los que tienen un error de medida inaceptable o los que no ayudan a mantener la fiabilidad del conjunto de la prueba. Con ellos se confeccionan las pruebas de CA y CL.

4.5. Correlación entre pruebas

Una vez diseñadas las pruebas con los ítems apropiados, también comprobamos si la confección del examen sirve para evaluar realmente habilidades diferenciadas entre sí (CA, CL, EIE y EIO). Para esto hallamos sus coeficientes de correlación: la medida en que en dos pruebas hay elementos comunes que, por lo tanto, estarían siendo evaluados varias veces y pesarían más en la nota final del examen. En nuestro caso, para considerar como definitivas las pruebas de CA, CL, EIE y EIO solo aceptamos –siguiendo la norma– una correlación menor de 0.7, lo que es prueba de que las pruebas miden habilidades diferentes. Por ejemplo, la correlación de la convocatoria de junio de 2015 entre las pruebas de CA y CL fue de .685.

4.6. Establecimiento de puntos de corte

En las pruebas de CA y CL, los puntos de corte, es decir, el número de ítems que se deben responder correctamente para superar el examen en cada nivel, se establecieron siguiendo el método Angoff (Cizek 2007; Consejo de Europa 2009), ampliado con datos de IRT (según Figueras 2011). El proceso se hace mediante rondas de análisis y opiniones intentando acercar posiciones entre los expertos a los que se pide opinión. En nuestro caso se utilizan tres rondas con datos de IRT (al modo en que se hace en Figueras, 2011). En las pruebas de EIE y EIO la calificación se realiza por el sistema de doble corrección. En la EIO uno de los evaluadores tiene el papel de entrevistador. Los evaluadores llevan a cabo sesiones periódicas de estandarización de criterios

4.7. Certificados

Para conseguir la acreditación en un nivel, hay que superar el 50% de los ítems establecidos para cada nivel (B1 o B2), así como un total de un 75% en su conjunto, compensando un mayor dominio en unas pruebas frente a las otras. Para evaluar las pruebas de EIE y de EIO, siguiendo las recomendaciones del MCER disponemos de una escala de calificación con los siguientes criterios: para EIE cumplimiento de la tarea, coherencia, corrección y alcance y dominio de vocabulario; y para EIO los mismos 4 criterios además de fluidez y capacidad de interacción. La nota se elabora mediante dos evaluadores. Las certificaciones se presentan al candidato con la calificación final de APTO en B1 o APTO en B2, explicitando el nivel que ha alcanzado en cada una de las destrezas. Si el candidato ha obtenido un nivel B1 con alguna destreza de nivel B2, se le guarda la calificación de B2 de esa destreza durante dos convocatorias por si quiere presentarse a una nueva convocatoria para conseguir un certificado del nivel superior en el resto de las destrezas.

4.8. Sedes de administración

El *eLADE* se administra en línea en centros de cualquier lugar del mundo que tengan un acuerdo para tal fin con la Universidad de Granada, así como en la sede del CLM de dicha universidad. En el caso de que no haya ninguna sede próxima al lugar de residencia de un candidato, este puede ponerse en contacto con cualquier institución social o académica y que esta solicite ser sede acreditada cumplimentando el protocolo existente al efecto en nuestra página web ya citada. Los requisitos para ello son disponer de los recursos técnicos que exigen la realización del examen –especificados en la guía de administración– y firmar el acuerdo de confidencialidad por el que –entre otras cosas– toda entidad examinadora se compromete a no revelar ninguna información sobre el contenido de cada examen. Cada centro examinador debe asimismo designar a un responsable de administración que siga los protocolos. Al matricularse se asigna a cada centro examinador o sede una fecha para que sus candidatos hagan la entrevista oral –mediante *Skype*– en un día y hora concretos. Para inscribirse como sede o ver los requisitos: www.clm-granada.com/html/c_espanol/esp/examenes/examen_acreditacion/examen_acreditacion_solicitud_sede.htm

4.9. Detalles técnico-informáticos

El *eLADE* está previsto para ser administrado a nivel mundial, por lo que se construyó en una plataforma que permite reproducir los contenidos del examen en ordenadores que necesiten pocos recursos y que es fácilmente accesible desde cualquier dispositivo con acceso a Internet. Se ha desarrollado en una plataforma propia, a partir de herramientas comunes de la Web 2.0. buscando siempre la mayor compatibilidad posible con cualquier dispositivo. Para más datos técnicos sobre seguridad, servidores, autenticación, codificación de vídeos, salvaguarda de datos, etc. ver Biedma (en prensa) o Martínez Baztán (2015).

5. Discusión

El examen *eLADE* de la Universidad de Granada, nacido para cubrir las necesidades de acreditación específicas para los programas de movilidad, se ha ampliado para evaluar a todo tipo de personas interesadas en acreditar un nivel de lengua tanto por razones académicas como sociales o personales. El hecho de ser un examen en línea y de fácil acceso, para el que no es necesario acudir a una sede determinada que implique un desplazamiento del candidato, con un formato sencillo y atractivo, y validado técnicamente con los estándares internacionales de evaluación de lenguas, lo convierten en una herramienta útil, fiable y necesaria para la comunidad educativa así como para el público en general. Ello nos invita a pensar en la futura acreditación de otros niveles de dominio, A1, A2, C1 y C2, y de este modo contribuir a la expansión de la lengua española con un instrumento como el *eLADE*. En tal caso, seguimos apostando por evaluar los nuevos niveles mediante un examen binivel que sería de A2/B1 y de B2/C1, en principio, utilizando los mismos bancos de ítems para los niveles comunes de cada examen. Pero antes de implementarlo es necesario investigar los resultados e implementar ciertas mejoras, siempre que se disponga de las herramientas adecuadas para ello. En este sen-

tido y a pesar del éxito y la satisfacción por el uso del *eLADE*, hay algunos aspectos sobre los que hay que realizar análisis en espera de posibles mejoras o reformas.

Hay otros elementos que se podrían plantear de cara a nuevas convocatorias, como hacer un uso más fuerte de las posibilidades del formato por ordenador, siempre debidamente acompañadas de análisis que muestren su validez y eficacia; explorar diferentes formatos de ítems, p. ej. desplazando elementos en el escritorio o pantalla del ordenador mediante el ratón (fotos, palabras, etc.) en tareas de emparejamiento o de completar huecos, que podrían dar juego a la dinámica, hacer un manejo más sencillo y aumentar la motivación y satisfacción de uso; también se puede ir hacia una gestión más autónoma o completamente autónoma (habría que estudiar los resultados) del uso del tiempo y del orden en que cada candidato elija realizar cada tarea –CA, CL y EIE– o volver atrás y adelante en ellas; finalmente, es importante comparar resultados de los modelos del examen en línea y en papel, siguiendo el camino trazado por otras investigaciones (Winke 2008) –algo que en el CLM podemos comparar fácilmente con diferentes grupos de estudiantes– así como analizar el importante factor de en qué medida la habilidad en el uso de ordenadores influye en los resultados del examen o en las notas (Madsen 1991), por ejemplo, analizar si se llegan a dar diferencias entre el uso por nativos tecnológicos frente al resto de usuarios, algo que puede tener relevancia en el futuro. El planteamiento y análisis de todos estos aspectos entra dentro de los habituales protocolos de validación de exámenes que empleamos en el CLM de la UGR si adoptamos, como hemos hecho y pretendemos mediante este artículo, un enfoque de evaluación de lenguas garantista (Martínez Baztán, 2011) y transparente que nos asegura que evaluamos lo que pretendemos y lo que nuestros usuarios y agentes sociales demandan. La Universidad de Granada está satisfecha con los resultados e implicada en la constante mejora de las herramientas.

Esperamos contribuir con este artículo a la necesaria difusión del examen y a obtener la credibilidad social y académica así como el crédito de los usuarios y agentes sociales que lo deseen utilizar para que nos sigan ofreciendo su confianza como evaluadores.

Referencias Bibliográficas

ALDERSON, J. Ch.; CLAPHAM, C.; WALL, D. (1998): *Exámenes de idiomas. Elaboración y evaluación*, Madrid: Cambridge University Press. [Título original: (1995): Language Test Construction and Evaluation, Cambridge: Cambridge University Press].

ASSOCIATION OF LANGUAGE TESTERS IN EUROPE (ALTE) (2001): *Principles of Good Practice for ALTE Examinations*, [pdf], [en línea: 21/12/2007], www.alte.org/attachments/files/good_practice.pdf

BACHMAN, L. F. (1990): *Fundamental Considerations in Language Testing*, Oxford, Oxford University Press.

BACHMAN, L. (2004): *Statistical Analyses for Language Assessment*, Cambridge, Cambridge University Press.

BACHMAN, L; PALMER, A. (1996): *Language Testing in Practice*, Oxford, Oxford University Press.

BIEDMA, A.; CHAMORRO, D.; MARTÍNEZ, A.; SÁNCHEZ, A.; SÁNCHEZ, S.; AMADOR, C.; PUERTAS, J. y RODRÍGUEZ, J. (en prensa): "eLADE: The University of Granada Bi-level B1/B2 Accreditation Exam. Construction and Validation Process", *CASALC* (Czech and Slovak Association of Language Centers) Review.

BIEDMA, A.; CHAMORRO, D.; MARTÍNEZ, A.; SÁNCHEZ, A.; SÁNCHEZ, S.; AMADOR, C.; PUERTAS, J. y RODRÍGUEZ, J. (en prensa): "eLADE: Examen para la Acreditación de Dominio del Español de niveles B1 o B2 de la Universidad de Granada" en *Actas del I Congreso FLECOS*, Granada.

BROWN, J. D. (1997): 'Computers in language testing: Present research and some future directions', *Language Learning and Technology* 1(1), 44–59. http://llt.msu.edu/vol1num1/brown

CHALHOB-DEVILLE M. (2010): 'Technology in standardized language assessments'. En R. Kaplan (ed.), *The Oxford handbook of applied linguistics*. Oxford, England: Oxford University Press, pp. 511–26.

CIZEK, G. y BUNCH, M. (2007): *Standard Setting*, Sage Publications, Thousand Oaks, California.

CONSEJO DE EUROPA (2002): *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación*, Madrid: Ministerio de Educación, Cultura y Deporte; Instituto Cervantes. [Título original (2001): *Common European Framework of Reference for Languages Learning, teaching, assessment*, Strasbourg: Council of Europe].

CONSEJO DE EUROPA (2009): *Relating Language Examinations to the CEF, Manual Preliminary Pilot Version. Introduction and Feedback to the pilot phase*, [pdf], Strasbourg: Council of Europe.

CONSEJO DE EUROPA (2011): *Manual for Language test Development and Examining*, [pdf], Strasbourg: Council of Europe.

DOUGLAS, D. (2010): *Understanding language testing*, London, England: Hodder Education.

DUNKEL, P. A. (1999): "Considerations in developing or using second/foreign language proficiency computer-adaptive tests", *Language Learning and Technology* 2(2), 77–93.

EUROPEAN ASSOCIATION FOR LANGUAGE TESTING AND ASESSMENT (EALTA) (2004): *Directrices de EALTA para una buena práctica en el desarrollo de pruebas y la evaluación de lenguas*, [en línea 5/12/2007] http://www.ealta.eu.org/documents/archive/guidelines/Spanish.pdf

FIGUERAS, N.; KAFTANDJIEVA, F.; TAKALA, S. (2011): "Relating a Reading Comprehension Test to the CEFR Levels: A Case of Standard Setting in Practice with Focus on Judges and Items", *The Canadian Modern Language Review/La Revue canadienne des langues vivantes*, 69, 4, (November / novembre), 359–385.

GREEN, R. (2013): *Statistical Analyses for Language Testers*, Palgrave McMillan.

INTERNATIONAL LANGUAGE TESTERS ASSOCIATION (ILTA) (2000), *Code of Etics for ILTA*, [en línea: 2/6/2012], www.iltaonline.com/images/pdfs/ILTA_Code.pdf

MADSEN, H. S. (1991): 'Computer-adaptive testing of listening and reading comprehension: The Brigham Young approach'. En P Dunkel (ed.), *Computer-assisted language learning and testing: Research issues and practice*. New York: Newbury House, pp. 237-57.

MARTÍNEZ BAZTÁN, A. (2008): *La evaluación oral: una equivalencia entre las guidelines de ACTFL y algunas escalas del MCER* (PDF) (tesis doctoral). Universidad de Granada.

MARTÍNEZ BAZTÁN, A. (2011): *La evaluación de lenguas: garantías y limitaciones*, Octaedro, Barcelona.

MARTÍNEZ BAZTÁN, A.; BIEDMA, A.; SÁNCHEZ, A.; CHAMORRO, M. D.; SÁNCHEZ, S. (2015): "Descripción y análisis del examen online de acreditación de dominio de español de los niveles B1 y B2: el eLADE". *Certiuni Journal, 1*, 52-66. www.certiunijournal.com

MUBARAK PATHAN, M. (2012): *Computer Assisted Language Testing [CALT]: Advantages, Implications and Limitations.* www.researchvistas.com, vol. 1 (104).

NORTH, B. y JONES, N. (2009): *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). Further Material on Maintaining Standards across Languages, Contexts and Administrations by exploiting Teacher Judgment and IRT Scaling*. [pdf], Strasbourg: Council of Europe, [en línea: 20/2/2014] http://www.coe.int/t/dg4/education/elp/elpreg/Source/Publications/Standards_language_EN.pdf

WINKE, P. y FEI FEI, (2008): "Computer-Assisted Language Assessment", en N. Van Deusen-Scholl and N. H. Hornberger (eds.), *Encyclopedia of Language and Education*, Volume 4: Second and Foreign Language Education Springer, Science+Business Media LLC, pp. 353-364.