La evaluación de segundas lenguas (L2). Balance y perspectivas

Teresa Bordón
Universidad Autónoma de Madrid
teresa.bordon@uam.es

Resumen: La evaluación de L2 constituye un campo propio de la lingüística aplicada y su evolución va ligada a la de la enseñanza de la lengua. En este artículo se traza un breve recorrido histórico por el campo de esta disciplina y se analizan algunos de los asuntos de mayor relevancia tanto por su aplicación e impacto social, como por el interés que suscitan para la investigación.

Palabras clave: evaluación de L2; tipos de exámenes; requisitos de un buen examen; responsabilidad de los exámenes a gran escala; la evaluación de la lengua oral; los exámenes de aprovechamiento; necesidad de transparencia y equidad en la evaluación de L2.

Abstract: The assessment of L2 has achieved its own space in the field of Applied Linguistics, and its development has run parallel to the teaching of L2. This article presents a brief survey of the historic evolution of this matter and it concentrates in some of the most relevant issues due to its social impact or for being the focus of active research.

Key words: assessment of L2; types of tests; characteristics of a good test; accountability of large scale tests; assessment of oral proficiency; achievement tests; need of transparency and fairness in the assessment of L2.

0. Introducción: presentación y revisión de conceptos básicos.

La evaluación es parte indiscutible de la enseñanza-aprendizaje de una L2 tanto por su valor formativo como por su necesidad práctica. Además, hoy día, la existencia de exámenes a gran escala para certificar niveles de dominio no sólo ha contribuido a reforzar la relación entre enseñanza-aprendizaje de L2 y evaluación, así como a fomentar la investigación en esta línea, sino que también ha supuesto el desarrollo de una poderosa industria. En torno a los exámenes de certificación se mueven grandes sumas de dinero por derechos de examen, contratación de personal experto para el diseño y administración de las pruebas, sin olvidar su influencia en la creación de cursos para la formación de evaluadores certificados, o la aparición de abundantes publicaciones destinadas a la preparación de los candidatos que optan a diplomas.

Antes de empezar el recorrido por la senda de la evaluación, es conveniente aclarar algunos conceptos que, aunque resultan bien conocidos, no siempre quedan perfectamente delimitados.

Uno: No es exactamente lo mismo evaluar que examinar. En una evaluación se recaban y analizan datos significativos acerca de algo y a partir de ellos se toman decisiones. Para llevarla a cabo se puede utilizar una variedad de procedimientos, entre ellos, exámenes. En un examen, en principio, también se hace esto, pero su objetivo fundamental es medir la actuación de alguien. Así, un examen es un tipo de evaluación. En este artículo se va a tratar fundamentalmente de exámenes, pero también se harán referencias a la evaluación ya que no todo puede ser examinable pero sí evaluable.

Dos: Ciñéndonos al campo de la evaluación de L2 y concentrándonos en la técnica del examen, se distinguen fundamentalmente dos tipos:

i. Exámenes destinados a obtener información acerca de los logros de los aprendices en el aula de L2, suelen denominarse génericamente exámenes de aprovechamiento.
ii. Exámenes orientados a determinar un cierto nivel lingüístico, el cual no tiene que haberse adquirido necesariamente a través de una instrucción formal: son los exámenes de nivel de dominio. Estos pueden ser pruebas de acceso, o de clasificación o de certificación de un nivel.

Tres: Todo examen, para que se trate de una herramienta bien construida, debe satisfacer las características de validez: medir lo que supone que debe medir; fiabilidad: ser consistente en sus resultados; y viabilidad[1]: hacer posible su ad-

1 Viabilidad es el término ampliamente aceptado en español para traducir lo que en inglés se conoce como *practicability*. La autora lo emplea dada su implantación, aunque prefiere el de *aceptabilidad*.

ministración. Estos tres requisitos se consideran absolutamente indispensables, y están totalmente consensuados como tales. No obstante, Bachman y Palmer (1996), a estos tres pilares garantes de la calidad del examen, añaden las características de autenticidad, entendida como el grado de correspondencia entre la tarea de examen y los rasgos de una tarea de uso de la lengua meta; interacción, que se refiere a la cantidad de características individuales del candidato y el tipo de implicación de las mismas que adopta para resolver las tareas de examen; y, finalmente, el impacto, que consiste en las repercusiones derivadas del examen en las personas y la sociedad.

Estos autores matizan, asimismo, la relación de estas características respecto de la propia herramienta evaluadora. Así, la fiabilidad y la validez son propias del instrumento examen, refiriéndose directamente a él. La autenticidad y la interacción, por su parte, están relacionadas con el objeto del examen, es decir con la lengua; y son relativas ya que pueden darse en mayor o menor medida. En lo que concierne al impacto, este se entiende como los efectos de los resultados de un examen, los cuales pueden tener una repercusión más o menos importante, ya sea afectando al propio candidato que puede depender de ellos, por ejemplo, para merecer una beca de estudios, conseguir un puesto de trabajo, o acceder a la ciudadanía de un país. El impacto también puede alcanzar a la sociedad, ya que hay exámenes que involucran a muchas personas e instituciones y mueven grandes sumas de dinero.

Los conceptos incluidos en estos tres puntos irán apareciendo en los diferentes apartados de este artículo, que se estructura presentando en primer lugar un breve recorrido histórico para trazar el origen y desarrollo de los exámenes de L2. A continuación, se tratarán algunos de los temas que preocupan en la evaluación de segundas lenguas y, en último lugar, se presentarán algunas de las posibles áreas de interés en este campo en el futuro cercano.

1. Mirando hacia atrás

1.1 Antecedentes

Sullivan (2011:259) recoge que ya los chinos implantaron algún tipo de prueba de lengua en el complejo sistema de exámenes para acceder al funcionariado imperial, un requisito que se mantuvo vigente durante 1.500 años hasta su abolición en 1905. No obstante, para comenzar este balance acerca del desarrollo de la evaluación de L2, no nos adentraremos en los complicados procedimientos de la antigua burocracia china.

Tradicionalmente, para establecer etapas en la evolución de la evaluación se ha venido usando una clasificación similar a la adoptada para clasificar los métodos y enfoques para la enseñanza-aprendizaje de segundas lenguas. Y esto no es casual ya que es a todas luces evidente que la evaluación de la lengua debe estar relacionada, tanto con la manera en que se entiende su naturaleza como con los procedimientos o técnicas adoptados para su enseñanza-aprendizaje.

De este modo, se establece una primera línea divisoria entre etapa precientífica y científica, y dentro de esta última se matiza algo más. Entre las clasificaciones más utilizadas se encuentra la de Spolsky (1976) que distingue una época pre-científica, y, ya dentro de la científica, una psicométrica-estructuralista (años cincuenta y sesenta del siglo xx) y otra psicolingüística-sociolingüística (correspondiente a la década de los setenta). Morrow (1979) haciendo gala de sentido del humor las llama «el jardín del edén» a la precientífica, »el valle de lágrimas» a la psicométrica y «la tierra prometida» a la última. A los años ochenta y noventa se los conoce como etapa comunicativa y todavía no se ha acuñado una denominación estable para el período de tiempo correspondiente a los años transcurridos del siglo xxi.

1.2 Etapa precientífica

En la etapa precientífica, los exámenes para evaluar la lengua consistían fundamentalmente en traducciones, escribir redacciones y realizar lecturas, siendo generalmente los mismos profesores quienes diseñaban los criterios para la evaluación de las pruebas.

No obstante, Sullivan (2011: 259) recoge que ya en 1915 Kelly desarrolla el formato de prueba con ítems de selección múltiple, lo cual permite aplicar esta metodología para la estandarización y usarla para exámenes a grandes grupos. Estos primeros exámenes de L2 a gran escala, por lo menos los que se recogen en la bibliografía especializada, tienen como objeto la lengua inglesa.

De esta época, por ejemplo, data *The Army Alpha*, un examen diseñado por un grupo de especialistas liderado por Robert Yerkes para evaluar a los reclutas norteamericanos para la I Guerra Mundial. Introducido en 1917, el examen medía la habilidad verbal y la numérica, así como la capacidad para seguir instrucciones y el conocimiento de la información. A partir de los resultados, se decidía acerca de las competencias de los soldados que se incorporaban al servicio, el tipo de trabajo que podría desempeñar y su aptitud para el mando. También existía una versión oral del examen pero los soldados analfabetos.

Igualmente, a principios del siglo xx, en 1913, el University of Cambridge Local Examination Syndicate (UCLES) —una institución activa desde 1858

con el objetivo de administrar exámenes y de mejorar los estándares en educación— publica su primer examen para certificar un nivel de inglés como lengua extranjera: el Certificate of Proficiency in English (CPE) destinado a los no nativos de inglés que querían estudiar en universidades del Reino Unido.

La existencia de estos exámenes a gran escala revela el interés desde las primeras décadas del siglo xx por evaluar la L2. A pesar de su mérito, no se puede considerar que se trate de ejemplos de pruebas realizadas con una base científica ya que carecen de fundamento teórico lingüístico o metodológico.

1.3 Etapa científica

En general, se considera que la etapa científica en lo que respecta a los exámenes de L2 comienza en los años cincuenta del siglo xx, el período en el que asimismo se imponen los métodos estructuralistas, especialmente el audiolingual en EE.UU. para la enseñanza aprendizaje de L2. En ese momento, el objetivo fundamental de la evaluación consiste en diseñar y aplicar procedimientos de medición —exámenes— que sean objetivos. Y puesto que el modelo de lengua es el estructuralista y la competencia lingüística se entiende como el conocimiento de los elementos que constituyen la lengua, los exámenes destinados a medir tal competencia se enfocarán en extraer el conocimiento de esos elementos discretos.

En 1961, Robert Lado[2] publica *Language testing*: *the construction and use of foreign language tests: a teacher's book*, considerado actualmente un texto decisivo para el establecimiento del campo de la evaluación de la lengua, como recoge Davidson (2004).

1.3.1 Los años sesenta: el modelo psicométrico

En los años sesenta del siglo xx, la investigación en el campo de la evaluación de la lengua estaba ampliamente dominada por la hipótesis de que la habilidad lingüística constituye un rasgo unitario, según la cual el conocimiento lingüístico se puede averiguar sumando la información obtenida a partir de los resultados en los distintos componentes del mismo. Sin embargo, algunos autores como Oller (1979) se apartan de esta línea [3] proponiendo otro tipo de exámenes que en vez de estar constituidos por pruebas enfocadas a elementos discretos, lo estén por pruebas integradoras como el *cloze test* (texto con huecos) o dictados, que pueden

2 Robert Lado, de origen español, está considerado el padre del análisis contrastivo. Fue un gran lingüista y una figura fundamental en el campo de la evaluación de L2 y de EL2 en particular. Fundó la academia de la lengua española de EE.UU.

3 Bachman (2003) recoge la observación hecha por Oller (1979) según la cual «el campo de la investigación de los exámenes de lengua (*testing*) estaba ampliamente dominado por la hipótesis de que la habilidad lingüística (*language proficiency*) consistía en un único rasgo unitario, así como por una metodología de investigación cuantitativa y estadística».

requerir la utilización de más de una destreza y que incluyen la presencia de un contexto.

No obstante, en lo que respecta a la evaluación de lengua en esta década, va a predominar el uso del modelo psicométrico, con la proliferación de exámenes con pruebas de ítems de selección múltiple destinadas a extraer los diferentes componentes de la lengua. Los exámenes de L2 realizados para obtener respuestas que constituyen elementos discretos se corresponden, por lo tanto, con el modelo estructuralista y conductista de lo que es aprender o adquirir una L2.

Y esto es así, porque un objetivo primordial en ese momento era la objetividad de la medición y la fiabilidad de los resultados. Pero, como se puede adivinar fácilmente, la validez de este tipo de pruebas es más que cuestionable: con este tipo de herramienta de evaluación es posible averiguar lo que el candidato sabe de la lengua —o mejor dicho: de los elementos constitutivos de la estructura de la lengua—, pero no lo que realmente puede hacer con ella en cuanto a comprender, expresarse o interactuar, en situaciones de uso de la lengua.

1.3.2 Los años setenta: crisis del modelo estructuralista

Más adelante, hacia la segunda mitad de los años setenta del siglo xx, algo empieza a moverse en el campo del aprendizaje de segundas lenguas respecto de la pedagogía que debe adoptarse para su enseñanza en el aula. El «fracaso» del modelo estructuralista para conseguir la adquisición de L2 de manera que el aprendiz pueda usarla de manera activa y eficaz, así como el desarrollo de los estudios de psicolingüística y sociolingüística hace que se produzca un giro hacia enfoques de tipo pragmático que entienden la lengua fundamentalmente como un medio de comunicación. De este modo, la evaluación de la lengua y los exámenes que se adopten para determinar niveles de competencia deberán responder a este nuevo modelo comunicativo de lengua.

Recordemos que en Europa, en 1971 ve la luz el influyente texto de Wilkins, *Notional Syllabuses* y que en 1975 El Consejo de Europa publica el *Threshold Level* y cuatro años más tarde (1979) su versión española, *El nivel Umbral*. La publicación de este nivel con todo su repertorio de funciones, nociones, exponentes gramaticales, etc. va a influir en todos los niveles de enseñanza-aprendizaje de L2: desde la producción de nuevos materiales y manuales para satisfacer la demanda de cursos diseñados con esta orientación, hasta, por supuesto, en la manera de evaluar la lengua.

Si bien en ese momento no existe todavía en español un examen destinado a medir el nivel de dominio del aprendiz, sí había aparecido ya alguno en inglés. Este cambio hacia una orientación de corte comunicativo lo señala Weir (2002)

poniendo como ejemplo la revisión del *Certificate of Proficiency in English* (CPE), realizada en 1975. Esta nueva versión del examen para el certificado se realiza de acuerdo con las nuevas orientaciones de entender la lengua como un instrumento de comunicación —alejándose del modelo de competencia unitaria— y dividiendo el examen en partes que corresponden a las destrezas de lengua: comprensión y expresión de manera oral y escrita, aunque se sigue incluyendo una prueba de uso del inglés.

De alguna manera, la década de los setenta se puede considerar como de transición, porque, si bien empiezan a surgir cambios, todavía están muy activos los modelos estructuralistas, tanto en la enseñanza-aprendizaje, como en la evaluación.

Será en 1979, con la celebración del primer *Language Testing Research Colloquium*, cuando se le reconocerá al campo de la evaluación de L2 un espacio propio dentro de la lingüística aplicada. También, en este congreso se da el portazo definitivo al modelo de la habilidad lingüística como rasgo unitario y se emprende el camino hacia los nuevos modelos de competencia comunicativa.

1.3.3 Los años ochenta: período psicolingüístico-sociolingüístico

La llegada de los años ochenta ve el triunfo de los estudios de psicolingüística y sociolingüística, muchos ya iniciados en la etapa anterior, que contribuirán a que la lengua ya no se conciba como algo monolítico. En este período se investiga y trabaja en la definición de un modelo de competencia comunicativa (Widdowson 1978, 1979, 1983), Savignon (1972, 1983), Canale & Swain (1981) que integre tanto competencia del sistema como de su uso. Esta concepción de la lengua no tiene como objetivo la mera especulación teórica sobre la naturaleza de la lengua, sino que al proceder de la lingüística aplicada busca convertirse en el modelo de lengua para su enseñanza y aprendizaje, y consecuentemente en el referente para su evaluación.

En este sentido, los trabajos recogidos en el volumen *New Directions in Language Testing* (1985)[4] reflejan los asuntos que preocupan en el campo del *testing* de L2 y los caminos que quiere tomar la evaluación de la lengua a partir del International Symposium on Language Testing celebrado en Hong Kong en 1983. Destacan como áreas de interés las que se centran en evaluación de la lengua y el currículo, los exámenes a gran escala, la evaluación de la habilidad oral y el complejo y amplio asunto de la validez. Asuntos que siguen vigentes hoy día y sobre los que se continúa debatiendo e investigando.

4 Una recopilación hecha por Y.P. Lee de las ponencias presentadas en el International Symposium on Language Testing, Hong Kong en 1983.

Y en esta década, encontramos un acontecimiento importante relacionado con la evaluación del español L2 a gran escala. Se trata de la creación de los diplomas DELE (Diplomas de Español Lengua Extranjera) que vieron la luz en 1988. Una iniciativa para dotar al español de unos diplomas que garantizan la posesión de un determinado nivel de dominio lingüístico a quien supere las pruebas que los integran. En su inicio, los DELE constaban sólo de tres niveles: inicial, intermedio y superior, aunque posteriormente han experimentado ajustes y reformas, existiendo actualmente para todos los niveles[5] descritos por el MCER. La aparición de los DELE supuso también dotar al español de un instrumento que contribuye al prestigio de la lengua.

También, en estos años (1983), se publican en EE.UU. las *Proficiency Guidelines* de ACTFL (*American Council for the Teaching of Foreing Languages*). Esta institución en colaboración con el ETS (*Educational Testing Service*) y la *Interagency Language Roundtable* adaptó la ya existente escala utilizada por el Departamento de Estado de EE.UU. y otras agencias federales para poder ser utilizada en contextos escolares (secundaria) y otros programas académicos. La escala original de cinco niveles se transforma en una de cuatro: Novato (*novice*), Intermedio (*intermediate*), Avanzado (*advanced*) y Superior (*superior*), que en realidad se convierten en nueve, puesto que el novato y el intermedio se subdividen en tres: bajo, medio y alto; y el avanzado en dos (en una posterior revisión también se especificará en bajo, intermedio y alto) y superior que se especifica en ese único nivel. Las guías de ACTFL que se adoptarán para la OPI (*Oral Proficiency Interview*) experimentarán revisiones en 1986, 1999 y 2012[6]. La OPI, que existe para certificar la competencia lingüística en varias lenguas y entre ellas el español, consistía originalmente en una entrevista cara a cara entre un evaluador, certificado por ACTFL tras un exhaustivo entrenamiento, y un candidato. La prueba no propone un nivel determinado, sino que a lo largo de ella se va estableciendo el nivel del candidato según lo que sea capaz de hacer: cuanto más alto sea el nivel más larga será la entrevista ya que el evaluador necesitará hacer más preguntas, sondeos y comprobaciones para ir fijando el nivel alcanzado. Actualmente existe la posibilidad de realizarla por teléfono y también existe una versión por internet para el inglés.

5 Véase: <http://diplomas.cervantes.es/>.

6 <http://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf>.

1.3.4 La década de los noventa y el inicio del segundo milenio. El triunfo del enfoque comunicativo.

En la década de los noventa se va a culminar la adopción de exámenes y pruebas de corte comunicativo para la evaluación de las segundas lenguas. Es fundamental, como punto de partida para asegurar el constructo, contar con un modelo que especifique en qué consiste la habilidad lingüística en una lengua. En este sentido, resultará muy influyente la propuesta de Bachman (1990), Bacham & Palmer (1996, 2010, que reorganiza y amplía el modelo de competencia comunicativa de Canale & Swain (1980), aunque deja fuera la competencia estratégica. No porque ésta no sea necesaria para la comunicación, sino porque se entiende más bien como una serie de componentes o estrategias metacognitivas que incluyen establecer las metas: decidir qué se va a hacer. Según esta propuesta, el conocimiento lingüístico se compone de[7]: Conocimiento Organizativo que incluye el conocimiento gramatical (fonología, morfología, sintaxis, vocabulario, ortografía), y el conocimiento textual (elementos de cohesión, organización retórica y de la conversación); y Conocimiento Pragmático que se refiere al conocimiento funcional (funciones ideales, funciones manipulativas, funciones heurísticas, funciones imaginativas) y el conocimiento sociolingüístico (dialectos/variedades, registros, expresiones idiomáticas, referencias culturales y figuras del habla).

El modelo de Bachman (1990) a pesar de su indudable influencia ha sido, sin embargo cuestionado. Por ejemplo, McNamara (2003.468) argumenta que el modelo resulta demasiado psicolingüístico y no incluye referencias al contexto social de la lengua.

La aparición de los varios modelos que proponen entender la lengua más allá de los modelos estructuralistas, no impiden que en 1991 Carroll[8] exprese que, a pesar de la investigación y de las propuestas orientadas hacia una evaluación de la lengua desde el paradigma comunicativo, sigue existiendo una gran falta de permeabilidad a la hora de adoptar los modelos comunicativos, tanto entre los profesores de segundas lenguas, como por parte de los administradores de exámenes e, incluso, en los medios académicos que ven con recelo las innovaciones. Pero a pesar de los recelos y la resistencia, ya no habrá vuelta atrás.

Otro importante acontecimiento que repercutirá tanto en la enseñanza como en la evaluación de L2 es la publicación en 1966 en EE.UU. de los *National Standards for Foreign Language Education for the 21st Century*[9], un documento de

7 Bachman y Palmer 1996, adaptado en Bordón 2006: 36.

8 Carroll, B. (1991): «Resistance to change», in Ch. Alderson & B. North (eds.), *Language Testing in the 1990s*, pp. 22–27.

9 El documento completo se puede consultar en: <http://www.actfl.org/node/192>.

gran impacto orientado a guiar la enseñanza, el aprendizaje y la evaluación de las segundas lenguas.

El documento es el fruto de un consenso sin precedentes entre educadores, instituciones gubernamentales, empresarios y la propia comunidad para definir los contenidos) y el papel de la enseñanza de lenguas extranjeras en el sistema de educación norteamericano. Es decir: lo que los estudiantes deben saber y lo que pueden hacer en términos de una lengua extranjera. Lo usan los maestros, la administración escolar, los diseñadores de currículo tanto a nivel estatal como local con el fin de mejorar la educación en lenguas extranjeras.

Su manifiesto filosófico se especifica en el siguiente párrafo:

> La lengua y la comunicación están en el corazón de la experiencia humana. Los Estados Unidos deben educar a estudiantes que estén lingüística y culturalmente equipados para comunicarse con éxito tanto en una sociedad americana plural como en el extranjero. Esta necesidad prevé un futuro en el cual TODOS los estudiantes tendrán que desarrollar y mantener un nivel de competencia en inglés y al menos en otra lengua moderna o clásica. Los niños que acceden a la escuela de un contexto no anglo hablante deben tener también la oportunidad de desarrollar habilidad lingüística en su lengua materna. (ACTFL Statement of Philosophy en <www.educationworld.com/standards/national/lang_arts/>)

Unos años más tarde —en 2001— y con otra perspectiva ve la luz en Europa otro documento fundamental: El *Common European Framework of Reference for Language Teaching, Learning. Assessment*, que un año más tarde (2002) aparecerá traducido en español: *Marco común europeo de referencia para la enseñanza, aprendizaje y evaluación de la lengua* (MCER)[10], que a la pregunta ¿Qué es? Esta es la respuesta que se da:

> El *Marco de referencia europeo* proporciona una base común para la elaboración de programas de lenguas, orientaciones curriculares, exámenes, manuales, etcétera, en toda Europa. Describe de forma integradora lo que tienen que aprender a hacer los estudiantes de lenguas con el fin de utilizar una lengua para comunicarse, así como los conocimientos y destrezas que tienen que desarrollar para poder actuar de manera eficaz. La descripción también comprende el contexto cultural donde se sitúa la lengua. El *Marco de referencia* define, asimismo, niveles de dominio de la lengua que permiten comprobar el progreso de los alumnos en cada fase del aprendizaje y a lo largo de su vida. (Capítulo 1: 1.1. c.v.c.cervantes.es/ensenanza/biblioteca_ele/marco)

10 Versión electrónica en <http://cvc.cervantes.es/ensenanza/biblioteca_ele/marco/>.

Estos dos documentos se pueden considerar como dos manifiestos de política lingüística que, sin duda, han influido en cada lado del Atlántico en cómo se debe abordar la enseñanza-aprendizaje de segundas lenguas y, consecuentemente, su evaluación.

2. La situación actual

En lo que respecta a los asuntos que suscitan más interés en el campo de la evaluación en la actualidad, destacan como temas vigentes:

- Los exámenes a gran escala, destinados a determinar niveles de competencia y proporcionar una certificación.
- El esfuerzo por garantizar la validez y la fiabilidad de las pruebas del examen y de éste en general.
- La evaluación de la expresión oral.
- El interés por la calidad de los exámenes en el aula.
- La necesidad de asegurar la transparencia y la equidad de los exámenes.

2.1 Los exámenes a gran escala

Como se ha presentado en el recorrido histórico, desde principios del siglo xx se siente la necesidad de contar con exámenes para medir la actuación de grupos numerosos de candidatos. Ya en los años cincuenta del siglo xx, se genera en EE.UU. una creciente industria en torno a los exámenes de lengua destinados a determinar niveles de lengua y certificar la habilidad de los candidatos, lo cual supone que se empiece a tomar en serio la necesidad de garantizar que el instrumento de evaluación sea realmente una herramienta válida y fiable.

Los exámenes de nivel de dominio, por lo tanto, son muy importantes por varias razones que se pueden concretar en su:

a) Repercusión. La existencia de este tipo de examen genera investigación en el campo de la evaluación, especialmente en asuntos como la coherencia con modelos de lengua, estudios sobre el alcance y tipos de validez, trabajos sobre la fiabilidad, desarrollo de técnicas y procedimientos de evaluación, diseño de descriptores de niveles, consideración de estrategias de examen.

b) Impacto social. Actualmente, la industria en torno a los exámenes de nivel de dominio destinados a evaluar a miles de candidatos es cada vez más poderosa, generando y moviendo grandes sumas de dinero. Por una parte, estas pruebas constituyen una fuente de ingresos para la institución que los gestiona ya que los candidatos pagan derechos de examen. Por otra,

sirven para promocionar empleos ya que instituciones públicas y centros privados necesitan expertos que los elaboren, así como personal que los administre y evalúe. También su existencia impulsa la creación de cursos orientados tanto a la formación de evaluadores como a la preparación de los candidatos que optan a los diplomas, e igualmente impulsan la publicación de libros y otros materiales por parte de editoriales especializadas.

c) Responsabilidad. Los exámenes de certificación se administran en general a grandes grupos de población y a partir de sus resultados se toman decisiones que pueden afectar de manera muy importante a sus usuarios. Un diploma que garantice un determinado nivel puede decidir el acceso de una persona a determinados estudios, la concesión de una beca, o la obtención de la ciudadanía de un país, por lo tanto tienen una gran responsabilidad por el poder que poseen.

Por lo tanto, a causa de su gran responsabilidad, derivada de su repercusión e impacto social, los exámenes de certificación deben garantizar absolutamente su validez, entendida ésta como la solidez de su constructo y en su apariencia, así como en su interacción con el contexto; también su fiabilidad, referida a la consistencia de sus resultados y de las evaluaciones; y su viabilidad, es decir: la posibilidad de aplicación. Igualmente, por todo lo anterior, deben ser impecables y meticulosos en todas sus fases: diseño, redacción, administración, experimentación y evaluación de los resultados.

De este modo, es necesario, para asegurar la calidad del instrumento medidor, que el examen esté perfectamente hecho: para ello habrá que disponer de equipos de redactores para las pruebas en los que intervengan lingüistas, profesores, evaluadores, estadísticos, e informáticos. También habrá que prever contar con profesionales encargados de evaluar las pruebas de expresión e interacción.

Las consecuencias sociales de los exámenes de lengua no pueden desligarse de la responsabilidad de los examinadores. Por eso, es fundamental su papel en cada uno de los estadios de la confección, administración y evaluación de un examen: de ahí la necesidad de su profesionalización.

Por lo tanto, para garantizar al usuario y a la sociedad en general que el examen es una herramienta «justa» y que todas las fases se realizan con transparencia, será necesario que se proporcione información asequible, clara y completa, asequible a todos los posibles usuarios. Información que permita saber exactamente en qué consiste el examen: número de pruebas y sus características, su orden de presentación y su duración, cuáles son los objetivos y contenidos de cada nivel descrito, criterios de evaluación, etc. Y por supuesto será una prioridad respetar y aplicar códigos de buena conducta, como los que proponen la

Association of Language Testers in Europe (ALTE) y la European Association for Language Testing and Assessment (EALTA) que están publicado en varias lenguas, entre ellas el español[11].

Los exámenes de certificación, como es sabido, tienen como objetivo garantizar que quien supere el nivel descrito por el propio examen, posee ese nivel. Algunos de ellos sólo garantizan el nivel alcanzado durante un tiempo determinado, teniendo fecha de caducidad la vigencia del nivel certificado.

En el contexto europeo, los exámenes de certificación se adscriben, en principio, a los niveles descritos por el MCER, como es el caso de los diplomas DELE[12] para el español. En EE.UU., por su parte, son bien conocidos como exámenes de nivel de dominio que certifican una determinada competencia lingüística el Test of English Foreign Language (TOEFL)[13] y la Oral Proficiency Interview (OPI)[14].

2.2 La validez

Probablemente se trate del requisito más importante, ya que garantiza que el examen mide lo que se propone medir. En las clasificaciones al respecto, se especifican distintos tipos de validez que, en general, se han reducido a tres: validez del constructo, del contenido y criterial (Cronbach: 1971)

No obstante, el concepto de ella que sigue siendo dominante hoy día es el desarrollado por Messick (1975, 1980), que considera que el constructo es fundamental para garantizar la validez y que cualquier otro tipo de evidencia debe servir para comprender mejor el constructo. Así, a partir de Messick (1989), la validez no es sólo una característica del examen, sino un concepto que conlleva la aceptabilidad del uso de un examen en un contexto concreto. Supone, asimismo, un proceso continuo de recabar evidencia acerca de lo que mide un examen, lo cual implica la posibilidad de hacer inferencias a partir de los resultados del mismo y de las actuaciones en él.

De esta manera, la validación de un examen de lengua continúa a lo largo de toda su vida. Es decir, no termina tras su diseño, aplicación y experimentación. Y supone tener en cuenta su impacto a corto y largo plazo, así como los cambios que se deriven de su puesta en práctica.

---

11 Se puede encontrar en: <http://www.alte.org/attachments/files/code_practice_es.pdf http://www.ealta.eu.org/documents/archive/guidelines/Spanish.pdf>.

12 <http://diplomas.cervantes.es>.

13 <http://www.ets.org/es/toefl>.

14 <http://www.languagetesting.com/oral-proficiency-interview-opi>.

Por lo tanto, dada su repercusión en la calidad del instrumento evaluador, la validez es una de las áreas más estudiadas en el área de la evaluación y es uno de los asuntos que genera más investigación.

2.3 La fiabilidad

La fiabilidad se refiere fundamentalmente a la consistencia de los resultados del examen. Es decir una misma prueba administrada a grupos semejantes proporcionará resultados similares, o pruebas semejantes administradas al mismo grupo también darán resultados parecidos. De no ser así, deberemos dudar de la fiabilidad de la prueba.

También es importante asegurar la fiabilidad de una evaluación, cuando se trata de valorar o calificar pruebas que no proporcionan respuestas cerradas susceptibles de ser sometidas a evaluaciones objetivas realizadas con máquinas o plantillas. En el caso de las pruebas que proporcionan respuestas abiertas como es en la producción de textos originales orales y escritos, que, por ahora, requieren de un evaluador humano, la presencia de la subjetividad en la valoración de la actuación del candidato es un hecho que puede afectar la fiabilidad de la calificación. La misma muestra de lengua puede no ser valorada igual por distintas personas. En estos casos, para garantizar una cierta objetividad y consecuentemente la fiabilidad de la evaluación será necesario contar con descriptores de nivel de dominio claros y rigurosos. Y no sólo esto, puesto que los descriptores por muy bien redactados que estén no van a poder evitar la utilización de palabras o frases como «casi siempre» o «algunos fallos» que se pueden prestar a diversas interpretaciones. Por lo tanto, habrá que formar a evaluadores para entrenarlos en el uso de los descriptores y asegurar la estabilidad de su aplicación. De ahí la necesidad de profesionalizar al evaluador cuando se trata de exámenes de certificación.

En los siguientes descriptores[15] diseñados para evaluar muestras originales de lengua escrita en lo que respecta a la competencia discursiva se pueden encontrar ejemplos de apreciaciones que pueden resultar ambiguas.

> Comunicación
>
> 4. Se ha logrado plenamente la finalidad comunicativa de la tarea: todos los puntos dados como orientación han sido tratados. Buena organización; ideas presentadas de forma coherente y lógica.
>
> 3. Se ha logrado la finalidad comunicativa de la tarea en términos generales, aunque puede faltar alguno de los puntos dados como orientación. Organiza-

15 Criterios para la evaluación de la prueba de expresión escrita del examen TELE, diseñado por Bordón, T. 1983, recogido en Bordón, T. (2006: 247-248).

> ción adecuada; puede haber pequeñas inconsistencias, pero no obstaculizan la transmisión de las ideas más importantes.
>
> 2. Se ha logrado sólo en parte la finalidad comunicativa de la tarea; muchos de los puntos dados como orientación no han sido tratados. Se entienden algunas ideas principales, pero la organización desigual puede hacer difícil la transmisión del mensaje en alguna ocasión.
>
> 1. Se ha logrado sólo mínimamente la finalidad comunicativa de la tarea; se ha tratado alguno de los puntos dados como orientación.

Fijémonos, por ejemplo, solo en el descriptor para el punto 3. Aparecen frases como «en términos generales», «puede faltar alguno de los puntos», «pequeñas inconsistencias» que lo más probable es que no todos los que vayan a aplicar el descriptor entiendan del mismo modo. Es más que posible que los descriptores tengan defectos de redacción y no sean perfectos, sin embargo la garantía de la fiabilidad del descriptor no reside tanto en la calidad del propio descriptor sino en la correcta interpretación del alcance de sus definiciones: esto solo se puede conseguir entrenando a quienes vayan a usarlos, de manera que todos den el mismo valor a «en términos generales», «algunos de los puntos» y «pequeñas inconsistencias».

Los dos requisitos revisados —fiabilidad y validez— son absolutamente indispensables para garantizar la calidad de un examen y no se puede dejar de tomarlos en consideración. Existe el peligro de que una prueba pueda ser válida en cuanto a su constructo, su apariencia y otros factores y sin embargo no ser fiable porque sus resultados no son consistentes, bien porque se aplica mal, bien porque se evalúa mal. Igualmente, un examen puede dar resultados muy fiables si consta de pruebas objetivas que se califican mecánicamente y sin embargo no ser válido, en cuanto que las pruebas que lo constituyan no supongan tareas de examen que respondan a usos efectivos de la lengua en situaciones de utilización real de la lengua.

3 La evaluación de la lengua oral

La evaluación de la lengua oral constituye uno de los grandes retos para evaluadores y examinadores, especialmente en los exámenes de nivel de dominio, dado su alto impacto individual y social. Por lo tanto, se deberá cuidar todo el proceso de elaboración, administración y evaluación de las pruebas que se incluyan en el examen.

El primer paso consistirá en delimitar el alcance de la destreza y de este modo garantizar el constructo. Por lo tanto, habrá que decidir el alcance de lengua oral y si esta se va a entender solo como expresión o como expresión e interacción.

Será fundamental diseñar procedimientos y tareas para obtener muestras de lengua evaluables, ya que cualquier actuación lingüística no constituye una muestra evaluable. Para serlo, el candidato debe proporcionar una cantidad de lengua que permita extraer la información que constituye el objetivo de la prueba y que proporciona realmente evidencia de lo que la persona podría hacer en una situación de uso real de la lengua hablada. De este modo, se buscará también que el candidato hable por lo menos diez minutos: esto para asegurar los niveles inferiores (A1, A2), pero para los niveles altos (C1, C2) conviene emplear más tiempo ya que es necesario comprobar una mayor cantidad de habilidades del candidato.

Para evaluar la actuación oral de un candidato se podrán utilizar pruebas que se enfoquen en la expresión: desde describir personas o situaciones con apoyo gráfico, hasta argumentar una opinión a partir de unas instrucciones; o tareas de examen que incluyan interacción, como utilizar fichas con vacíos de información para hablar entre pares o mantener una conversación con un examinador. Y no hay que olvidar que un examinador bien entrenado para la evaluación de pruebas orales es capaz de sacar una mejor muestra de lengua (en cuanto a calidad y cantidad) que alguien no acostumbrado a este papel. Por eso, la interacción entre pares puede verse afectada por factores no lingüísticos, como la timidez de un candidato frente a la agresividad del otro, desajustes por diferencias de edad o cultura, de manera que no se obtengan muestras de lengua realmente válidas.

Igualmente, hay que hacer hincapié en la necesidad de proporcionar al candidato instrucciones claras y sencillas de las pruebas, así como el tiempo del que se dispone para realizarlas.

Puesto que la actuación oral de los candidatos solo se puede evaluar de manera subjetiva, es decir haciéndolo personas, también habrá que definir criterios de evaluación en forma de descriptores, ya que es imposible evaluar muestras originales de lengua con procedimientos de medición objetivos. Para garantizar la fiabilidad de la evaluación, se recomienda la utilización de una doble corrección, y así asegurar la estabilidad de la valoración asignada a la actuación del candidato.

En los exámenes de certificación, se deberían grabar las pruebas orales ya que de esta manera se conserva evidencia de la actuación del candidato y de la intervención del evaluador, facilitando de este modo la posibilidad de rendir cuentas de la calificación en el caso de que fuera necesario. La grabación es también una herramienta para garantizar la validez de las pruebas ya que permite analizar los efectos de la técnica utilizada para extraer las muestras de lengua en relación con

el constructo. Contar con un documento que recoge la actuación oral del candidato permitirá, asimismo, contrastar los resultados de la evaluación: el mismo evaluador puede escuchar la muestra en otra ocasión y comparar su calificación con la asignada en un primer momento, o bien dos evaluadores diferentes pueden escuchar la misma muestra y comparar los resultados, lo cual garantizará la fiabilidad de la calificación. La grabación constituye, asimismo, la mejor manera de obtener muestras de lengua auténticas de personas en situaciones de examen, las cuales se pueden utilizar para trabajar en la formación de evaluadores cualificados.

Realizar pruebas orales de manera rigurosa, como es dedicar tiempo suficiente para extraer una muestra evaluable, contar con evaluadores profesionalizados, o realizar doble corrección, exige, no obstante, la necesidad de disponer de mucho tiempo, así como acometer una gran inversión en recursos humanos, lo cual supone un alto coste económico, que no todos los centros o instituciones que realizan exámenes con pruebas orales para grandes grupos de candidatos pueden o quieren asumir.

4. La evaluación en el aula: los exámenes de aprovechamiento

El impacto social de un examen de aprovechamiento se considera menor que el de un nivel de dominio, pero no por ello se debe olvidar la meticulosidad a la hora de diseñar, administrar y evaluar pruebas para los aprendices de EL2.

No obstante, aunque los profesores seamos cuidadosos a la hora de elaborar, administrar y valorar o calificar pruebas con las que evaluar a nuestros estudiantes, puede ser que en algún momento del proceso ocurra algún tipo de fallo. Sin embargo, las repercusiones de un error en un examen de aprovechamiento no serán tan graves ni tendrán el mismo impacto que si se tratara de un examen de certificación. Se supone que a lo largo de un período de instrucción se administran varias pruebas a los estudiantes y además se tiene evidencia de muestras de lengua del alumno por su actuación en el aula, tareas y ejercicios que lleva a cabo, etc. Es decir, además de exámenes, lo más probable es que el profesor lleve a cabo una evaluación continua y disponga de otras fuentes de información, que valorará a la hora de calificar a su estudiante.

Si bien los exámenes en situaciones de aprendizaje formal pueden ser necesarios por una serie de razones como su imposición por razones de política educativa, sin embargo hay que considerar otras maneras de llevar a cabo la evaluación de los aprendices de L2

Como recoge Hamp-Lyons (2007), el contexto y las necesidades de la evaluación en el aula no son los mismos que para los exámenes a gran escala y los destinados a certificar niveles de competencia. Este tipo de pruebas requieren discriminar, categorizar y calificar a grandes grupos, mientras que en el aula el objetivo de la evaluación sería valorar lo individual.

Esta autora defiende la existencia de dos maneras (las denomina culturas) de entender la evaluación: una que tiene que ver con el aprendizaje (*learning culture*) y otra que se refiere a los exámenes (*exam culture*). En la primera, la evaluación está influida por consideraciones sobre la enseñanza y el aprendizaje, mientras que en la segunda la evaluación se entiende como una preparación para exámenes externos.

Argumenta, también, que si se quieren introducir cambios orientados hacia una evaluación que adopte procedimientos de corte más humanista será necesario hacer hincapié en la preparación de los profesores respecto de la evaluación, ya que la «cultura del examen» sigue siendo la ideología dominante en el discurso educativo, económico y político. Incide asimismo en la necesidad de que profesores y expertos en evaluación sean más abiertos en la manera de considerar los puntos fuertes de ambas «culturas» de manera que la evaluación de los aprendices se pueda beneficiar de ello.

Hamp-Lyons (2013) [16]considera que uno de los retos más importantes para llevar a cabo una evaluación basada en el aula, es decir una evaluación de un tipo más humanista y que no adopte necesariamente exámenes, es definir una serie de principios que ella resume como:

- Principios para el aula, que incluyen: definir objetivos claros, diseñar tareas de aprendizaje y de evaluación apropiadas, comunicar los criterios de evaluación a profesores y alumnos, proporcionar retroalimentación de alta calidad y proveer oportunidades para la autoevaluación y la evaluación en parejas.
- Principios técnicos, con ellos Hamp-Lyons se refiere a que la evaluación por parte del profesor debe suponer: interpretar de manera crítica la evidencia de la actuación de los alumnos, obtener diversas fuentes de evidencia, aplicar el conocimiento de principios de medida apropiados y valorar los juicios acerca de los significados de los resultados
- Principios éticos, respetarlos implica que: la evaluación en el aula debe ser justa y ser consciente de sus consecuencias tanto previstas como im-

16 Adaptado del texto del power point de la conferencia plenaria «The challenge of classroom-based assessment», pronunciada por Liz Hamp-Lyons en el 10 Congreso de EALTA, celebrado en Estambul en 2013. <http://www.ealta.eu.org/conference/2013/keynotes/L_zHamp%20Lyons%20The%20challenge%20of%20classroom-based%20assessment.pdf>.

previstas, influir positivamente en la motivación y el aprendizaje de los alumnos, ser eficaz y factible, así como fortalecedora para los profesores.

Tener en cuenta estos principios constituye una buena guía para el docente responsable de diseñar procedimientos de evaluación para sus alumnos, respetando su individualidad y los procesos implicados en la enseñanza y aprendizaje de una lengua.

5. Mirando al futuro

En este apartado se esbozan algunas de las que creemos constituyen áreas de preocupación y de futura investigación en el campo de la evaluación de L2, si bien siguen siendo objeto de interés las presentadas en los apartados anteriores.

5.1 Preocupación por la transparencia y equidad de los exámenes

Una cuestión fundamental en el campo de la evaluación, y especialmente en lo que respecta a los exámenes a gran escala que constituyen herramientas con un gran poder e impacto, es la necesidad de asegurar que sean transparentes y justos. La utilización de los resultados de los exámenes a gran escala destinados a determinar y certificar niveles de lengua no se limita a obtener información sobre la competencia en L2 de los sujetos que se han sometido a la prueba. Hoy día se han convertido —y llevan camino de hacerlo cada vez más— en herramientas políticas que utilizan instituciones y gobiernos para decidir cuestiones fundamentales sobre el futuro de las personas, como convertir la certificación de un determinado nivel lingüístico en un requisito fundamental ya no solo para el acceso a la ciudadanía de un país, sino incluso para la entrada en él.

Algunos especialistas, como McNamara (2005, 2006, 2008) y Shohamy (1997, 2008) se han interesado especialmente en indagar sobre esta cuestión, reflexionando en sus libros y artículos sobre las dimensiones éticas y políticas del uso de los resultados de los exámenes de L2, y cómo su efecto puede proyectarse incluso hasta rozar el respeto (o mejor dicho: la falta de respeto) de los derechos humanos.

5.2 Incorporación de la tecnología

La utilización de las nuevas tecnologías es una realidad en la enseñanza de segundas lenguas y por lo tanto también se refleja en la evaluación. Ya es posible encontrar en la red pruebas que permiten la autoevaluación del aprendiz de L2,

como es DIALANG[17], que facilita a su usuario diagnosticar su nivel en la lengua en que desee evaluarse. Este sistema de autoevaluación —no se considera un examen— se desarrolló con la colaboración de varias instituciones europeas de educación superior y utiliza los niveles de competencia definidos en el MCER. A través de él se pueden evaluar las destrezas de leer, escribir y escuchar, así como la competencia gramatical y léxica. Actualmente, la Universidad de Lancaster (Reino Unido) se ocupa de su mantenimiento.

En EE.UU. el examen oral por medio de la entrevista OPI, definida de acuerdo con los niveles descritos por ACTFL también existe en una versión en Internet[18].

Actualmente se está desarrollando la investigación enfocada al diseño de programas informáticos que permitan evaluar y calificar no sólo pruebas que proporcionan respuestas cerradas, sino también la producción de textos originales escritos[19].

5.3 La evaluación de la interculturalidad

El MCER presta atención a este asunto y en su capítulo 5, dedica el apartado 5.1.2.2.[20] a especificar las destrezas y las habilidades interculturales que los usuarios del Marco (docentes, evaluadores) deben tener presentes. Igualmente, en la introducción del Plan Curricular del Instituto Cervantes[21] se incluye la dimensión del alumno como *hablante intercultural*, como alguien «capaz de identificar los aspectos relevantes de la nueva cultura a la que accede a través del español y establecer puentes entre la cultura de origen y la de los países hispanohablantes». Y matiza que esta denominación «ha de entenderse en sentido amplio (…) al papel del alumno en todo lo relacionado con la dimensión cultural».

De esta manera, la interculturalidad se ha incorporado ya al currículo de L2 de programas e instituciones. Si esto es así, la interculturalidad constituirá entonces también un objeto de evaluación. Y evaluar esta dimensión constituye un nuevo reto que ya ha producido algunos proyectos interesantes como el ICCinTe[22] que toma en consideración la propuesta y las pautas del MCER.

---

17 <http://www.lancaster.ac.uk/researchenterprise/dialang/about>.

18 Más información en: <www.languagetesting.com/oral-proficiency-interview-by-computer-opic>.

19 En la Tesis Doctoral de Paz Ferrero García de Jalón (Universidad Autónoma de Madrid, abril de 2011) se presenta una serie herramientas automáticas para procesar funciones y analizar el léxico, la sintaxis y la semántica de textos escritos que se experimentan y cuyos resultados se contrastan con los de evaluadores.

20 <http://cvc.cervantes.es/ensenanza/biblioteca_ele/marco/cap_05.htm#p512>.

21 <http://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular/niveles/01_objetivos_introduccion.htm>.

22 <http://archive.ecml.at/mtp2/Iccinte/results/en/assessing-competence.htm>.

6. Conclusión

Es evidente que los exámenes son necesarios: constituyen un instrumento de evaluación que permite tomar decisiones que de otro modo requerirían mayor inversión de tiempo y medios. En este sentido los exámenes son herramientas útiles, siempre y cuando la herramienta sea válida, fiable y aplicable para el fin para el que ha sido diseñada. La evaluación constituye un área fundamental dentro de la lingüística aplicada cuya evolución va ligada a la de la propia enseñanza de las segundas lenguas y constituye, asimismo, un amplio campo investigación que proporciona interesantes y estimulantes retos.

7. Bibliografía

Bachman, L.F. y Palmer, S. (1996). *Language Testing in* Practice. Oxford: Oxford University Press.

Bordón, T. y Liskin-Gasparro, J. (2014): «The Assessment and Evaluation of Spanish». En M. Lacorte (ed.) *The Routledge Handbook of Hispanic Applied Linguistics*. New York and London. Routledge: 258–274.

Bordón, T. y Liskin-Gasparro, J. (2007): «Evaluación» en M. Lacorte (coord.) *Lingüística aplicada del español* . Madrid. Arco Libros: 211–251.

Bordón, T. (2006): *La evaluación de la lengua en el marco de E/L2: Bases y procedimientos*. Madrid: Arco Libros.

Bordón, T. (2004): «Panorama histórico de algunas de las cuestiones fundamentales en la evaluación de segundas lenguas». *CARABELA*, Madrid, SGEL, nº55: 5–30.

Carroll, B. (1991). «Resistance to change». En Ch. Alderson y B. North (eds.). *Language Testing in the 1990s*. Modern English Publications and the British Council: 22–27.

Hamp-Lyons, L. (2007). «The Impact of Testing Practices on Teaching» en J. Cummins y Ch. Davison (eds.) *International Handbook of English Language Teaching*. New York. Springer: 487–504.

Lado, R. (1961). *Language Testing. The Construction and Use of Foreign Language Tests*. London, Longman.

Lee, Y.P. *et alii* (eds.) (1985). *New Directions in Language Testing*, Oxford, Pergamon.

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

McNamara, T.; Shohamy, E. (2008). «Language tests and human rights». *International Journal of Applied Linguistics*. John Wiley & Sons Ltd. Vol. 18, No. 1: 89–95

McNamara, T. (2005). «Introduction». En E. Hinkel (ed.) *Handbook of Research in second language Teaching and Learning*. Mahwah, N.J. Lawrence Erlbaum Associates: 775–778.

McNamara, T. (2003). «Looking back, looking forward: rethinking Bachman», *Language Testing*, 2003, 20 (4): 466–473.

McNamara, T. (2000), *Language Testing*. Oxford. Oxford University Press.

Messick, S. (1989): «Validity» en R.L. Linn (ed.) *Educational Measurement* (pp. 13–104). New York, Palgrave Macmillan.

Oller, J.W. (1979). *Language Tests at School*. London. Longman.

O'Sullivan, B. (2011). «Language Testing». En J. Simpson (ed.) *Routledge Handbook of Applied Linguistics*. New York, Routledge: 259–273.

Shohamy, E. (1997), «Testing methods, testing consequences: are they ethical?». *Language Testing* 14, pp. 340–349.

Weir, C. y M. Milanovic (eds.) (2002). *Innovation and Continuity: Revising the Cambridge Proficiency Examination,* UCLES/CUP.