LA FIABILIDAD DE LA PRUEBA DE COMPRENSIÓN AUDITIVA EN LOS DELE: DE LA *HISTORIA OFICIAL* A LA REALIDAD[1]

Abel Cárdenas Martínez

Universidad de Nanzan (Japón)

1. Introducción

Los Diplomas de Español como Lengua Extranjera (DELE), certificaciones oficiales otorgadas por el Instituto Cervantes en nombre del Ministerio de Educación y Ciencia de España, se están convirtiendo en uno de los sistemas de evaluación del dominio del idioma español de más importancia en el mundo. La magnitud de su impacto se puede observar en el aumento constante no sólo en el número de candidatos que toman los exámenes cada convocatoria, sino también en el número de centros que los ofrecen y las instituciones que los aceptan y los promueven. La influencia del sistema, además, se ha extendido a otros sistemas de evaluación de países hispanohablantes (p. ej., los CELA de la Universidad Nacional Autónoma de México), que han tomado los exámenes DELE como base para elaborar sus propios instrumentos de evaluación de la proficiencia del español como lengua extranjera. A pesar de la importancia de los DELE, sin embargo, poco se sabe de las cualidades psicométricas de las diferentes pruebas de este sistema de evaluación. ¿Son fiables y válidos los resultados que se obtienen a través de los exámenes DELE? Ni el Instituto Cervantes, el organismo principal responsable del sistema, ni la Universidad de Salamanca, la institución encargada de la elaboración y calificación de los exámenes, han mostrado hasta la fecha evidencia empírica algu-

1. Este estudio se realizó con ayuda del 2005 Pache Research I-A-2 Fund de University Nanzan.

na de que los exámenes o, mejor dicho, los resultados de los exámenes sean fiables y válidos para los propósitos para los que fueron creados.

Después de más de una década de constantes, pero fallidos, intentos de llamar la atención de las instituciones mencionadas para que actúen de una manera más profesional, ya sea publicando estudios empíricos de validación o proporcionando información a especialistas interesados en llevar a cabo este tipo de trabajos, el autor de este artículo decidió tomar la iniciativa para suplir esta deficiencia. El trabajo que aquí se presenta forma parte de un proyecto a largo plazo que intentará encontrar evidencias empíricas para poder contestar la pregunta que se ha mencionado anteriomente. El objetivo principal de esta comunicación es la exploración del primer aspecto de la pregunta, es decir, *la fiabilidad*, una de las cualidades que todo instrumento de evaluación debe poseer y que se ha definido como "la medida en que los resultados de un examen son estables, consistentes y libres de error" (Saville, 2003: 69). En la primera parte realizaremos un análisis de los documentos oficiales para determinar la manera en que los creadores y administradores de los exámenes DELE han abordado el tema de la fiabilidad de los resultados de las pruebas hasta la fecha. Este breve análisis, que formará lo que se considerará como la *historia oficial*, servirá de base para la segunda parte, y central, de este trabajo, en la que se presentará una investigación que tuvo como objetivo primordial determinar el grado de fiabilidad de los resultados de las pruebas de comprensión auditiva de dos de los tres niveles del sistema de evaluación DELE. Con el análisis de los datos obtenidos se mostrará la discrepancia que existe entre lo que se dice en la *historia oficial* y lo que ocurre en la realidad.

2.

2.1. La fiabilidad de los DELE: la *historia oficial*

La literatura sobre el sistema de evaluación de los DELE es escasa. Los contados trabajos que se han publicado (Pisonero y García, 1991; Eguiluz et ál., 1996; Eguiluz y Eguiluz, 2004a y 2004b; Fernández, 2004; Parrondo, 2004a y 2004b; Prieto, 2004), incluyendo el de este autor (Cárdenas, 2001) son sólo trabajos descriptivos o promocionales de los DELE, pero ninguno se podría considerar como investigación empírica de los exámenes y menos aún sobre la fiabilidad de los resultados que se obtienen con los mismos.

Las contadas publicaciones oficiales de las instituciones involucradas en el

sistema de evaluación que nos ocupa tampoco abordan con detalle el tema de la fiabilidad. En la *Guía para la obtención de los DELE.*, se nos presenta la primera *historia oficial* cuando se menciona que:

> Los exámenes son diseñados y posteriormente corregidos por un equipo de más de cincuenta profesores de la Universidad de Salamanca. La experiencia de esta universidad, especializada en los últimos 30 años en la enseñanza del español, *garantiza la calidad de los exámenes, así como la fiabilidad de los resultados*[2]. (MEC, 1992: 5)

Esto nos hace pensar que la experiencia que había adquirido la Universidad de Salamanca en la enseñanza del español como lengua extranjera parecía ser suficiente para garantizar la fiabilidad de las pruebas en los inicios del sistema de evaluación.

En una publicación más reciente, *Descripción General del Sistema de Evaluación DELE*, se afirma:

> El sistema de examen de los Diplomas de Español ofrece un servicio de evaluación independiente, *altamente fiable*, en el que se combina la utilización de las más modernas tecnologías en la corrección automática de las pruebas objetivas con la experiencia de más de 500 especialistas en evaluación de lengua española en el mundo en pruebas no objetivas. La homogeneidad de las condiciones de celebración de las convocatorias de examen y las medidas de seguridad que se adoptan permiten al Instituto Cervantes *garantizar la fiabilidad de las pruebas*[3] convocatoria a convocatoria, centro a centro. (IC, 2005: 5)

De acuerdo con la *historia oficial* que nos presenta estas citas estamos, pues, ante un sistema de evaluación perfecto o casi perfecto, lo que conduce al Instituto Cervantes a extender una garantía de la fiabilidad de los resultados. Sin embargo, algo que podemos destacar es que después de más de una década y media de existencia del sistema de evaluación, en ninguna de las contadas publicaciones oficiales aparecen evidencias que avalen esta garantía de fiabilidad y, por supuesto, de validez de los resultados de las pruebas de los DELE. Se menciona, por ejemplo, que el sistema es altamente fiable por la utilización de la tecnología en la corrección automática de las pruebas objetivas y la participación de especialistas en la evaluación de pruebas subjetivas, pero no hay referencias a reportes de estudios de fiabilidad de los resultados de las diferentes pruebas. ¿Cuál es el grado de fiabilidad de los resultados de cada

2. Énfasis del autor en cursiva.

3. Íd.

una de las tres pruebas que se califican por corrección automática, es decir, las pruebas de gramática y vocabulario, comprensión auditiva y comprensión lectora? ¿Cuál es el grado de fiabilidad de los correctores de las pruebas de expresión oral y expresión escrita? Las respuestas a estas preguntas de vital importancia brillan por su ausencia en la *historia oficial* que nos presenta el Instituto Cervantes.

El insistir en la necesidad de publicar evidencias que demuestren la fiabilidad y validez de los resultados de una prueba no es nada nuevo. Los especialistas en el campo de la evaluación siempre la han considerado como una responsabilidad y obligación de las instituciones, organizaciones e individuos que ofrecen y administran un sistema de evaluación. Los códigos de ética o estándares[4] elaborados por asociaciones profesionales tales como AERA (American Educational Research Association, 1999), ILTA (International Language Testing Association, EALTA (European Association for Language Testing and Assessment) y ALTE (Association of Language Testers in Europe), a la que, por cierto, pertenecen tanto el Instituto Cervantes como la Universidad de Salamanca, son un recordatorio constante de que no es suficiente afirmar que los resultados de tal o cual prueba son *altamente fiables o válidos*, sino que estas afirmaciones se deben avalar con evidencias. Saville, de hecho, indica que entre las responsabilidades de los miembros de ALTE encontramos las siguientes:

> *Recabar y analizar datos* de una muestra adecuada de candidatos *para calcular el nivel de dificultad, discriminación, fiabilidad y error de medición del examen.*
>
> *Proporcionar información a los usuarios* sobre el contexto, propósito, uso, contenido y *la fiabilidad* de los resultados del examen[5]. (Saville, 2005: 3)

2.2. La fiabilidad de las pruebas de comprensión auditiva en la realidad

2.2.1. Objetivo del estudio

El objetivo general que se fijó para el estudio que aquí se presenta era muy

4. Los documentos relevantes se pueden obtener en ILTA <http://www.iltaonline.com/>, EALTA <http://www.iltaonline.com/> y ALTE <http://www.iltaonline.com/>

5. Énfasis del autor en cursiva.

concreto: encontrar evidencias que avalaran y/o cuestionaran la fiabilidad de los exámenes DELE, fiabilidad que, como hemos visto, ha sido y es garantizada pero nunca demostrada. Nuestro plan original contemplaba explorar la fiabilidad de todas las pruebas del sistema de evaluación; sin embargo, la falta de cooperación y apoyo de las instituciones responsables nos obligó a concentrarnos en las pruebas de corrección objetiva, especialmente dentro del contexto japonés. Apegándonos a la temática central de este congreso, es decir, las destrezas orales, nos limitaremos aquí a contestar a la siguiente pregunta específica: ¿Cuál es el grado de fiabilidad de los resultados de las pruebas de interpretación de textos orales y comprensión auditiva del nivel inicial y el nivel intermedio de los DELE al administrarse a diferentes grupos de estudiantes universitarios japoneses?[6]

2.2.2. Participantes

El estudio contó con la participación voluntaria de 5 grupos de estudiantes de una universidad japonesa (2 para el nivel inicial y 3 para el nivel intermedio). La mayoría de los participantes era del sexo femenino (80%) y su edad fluctuaba entre los 19 y los 23 años. En el momento de hacer la investigación los 104 estudiantes del grupo que denominaremos *inicial* estaban a punto de terminar el segundo año de la licenciatura que, normalmente, en Japón se conoce como estudios hispánicos, y que incluye aproximadamente un promedio de 462 horas de asignaturas de español como lengua extranjera en los primeros dos años de la carrera. El grupo *intermedio*, con 121 participantes, estaba formado por 3 grupos diferentes de estudiantes de tercer y cuarto año de la carrera universitaria, quienes habían tomado entre un mínimo de 498 y un máximo de 1.370 horas de clases de español.

2.2.3. Procedimiento

A cada uno de los grupos se le administró una versión oficial ya retirada de los exámenes para obtener el DELE del nivel correspondiente. Como es de todos conocido, cada nivel consta de 5 pruebas en total, tres de las cuales, incluyendo la de comprensión auditiva, son pruebas que en la literatura sobre la evaluación (Alderson, 1995; Bachman, 2004; Brown, 1996) se conocen

6. Cf. Cárdenas (2007) para el reporte completo.

como tipo objetivo o corrección objetiva y en las que los candidatos seleccionan entre dos o más opciones las respuestas que consideren adecuada.

Las pruebas del nivel inicial se administraron en dos ocasiones diferentes. Las pruebas del nivel intermedio, sin embargo, se aplicaron en tres años consecutivos. La supervisión de la administración de las pruebas en todo momento estuvo a cargo del autor y otros colegas de la institución donde se realizó la investigación. Como sucede en una administración oficial, las hojas de respuesta fueron calificadas con el uso de medios tecnológicos. Una vez que se obtuvieron los resultados, el autor se encargó de la preparación de los datos para realizar un estudio más detallado de las propiedades psicométricas de las pruebas con ayuda de los programas informáticos TiaPlus o *Test and Item Analysis* (Cito, 2006) y TAP o *Test Analysis Program* (Brooks, 2003-2005), que se basan en lo que se conoce como la *Teoría Clásica de los Exámenes*.

2.2.4. Resultados y discusión

En la Tabla 1 se puede ver un resumen del análisis de los resultados que se obtuvieron de las administraciones de la prueba de interpretación de textos orales y de la prueba de comprensión auditiva: número de ítems, puntuación media, etc. Obsérvese la última fila de la tabla para ver lo que más nos interesa, es decir, el grado de fiabilidad de los resultados. En esta fila tenemos el coeficiente alfa de Cronbach que nos indica la consistencia interna de las pruebas y que se considera como muestra de fiabilidad. Es pertinente recordar aquí que el coeficiente de fiabilidad varía de 0 a 1 y que muchos expertos en evaluación (p. ej.: McNamara, 2000; Weir, 2005) consideran que éste debe ser superior a 0,90 cuando se trata de exámenes de alto impacto de uso internacional[7]. Si tomamos este criterio como base, desafortunadamente nos encontramos con datos alarmantes. Para sorpresa del autor, la prueba de interpretación de textos orales en el nivel inicial obtuvo un coeficiente alfa de 0,49, una cifra extremadamente baja para un sistema de evaluación de uso internacional. Si observamos los resultados del mismo tipo de prueba en el nivel intermedio, la situación es igualmente preocupante, ya que la prueba de comprensión auditiva obtuvo también el mismo coeficiente de fiabilidad, es decir, 0,49. Si recordamos lo que en la literatura se dice sobre el significado de fiabilidad (p. ej.: Brown, 2005: 175), se podría decir que los resultados que se

7. Algunos expertos consideran que lo mínimo aceptable es 0,80.

obtienen a través de la aplicación de estas pruebas a los participantes en este estudio se ven afectados o influidos negativamente hasta en un 51% por factores ajenos a los verdaderos conocimientos o habilidades de los candidatos. Por lo tanto, tomando en cuenta los datos de esta tabla, en cuanto concierne a la fiabilidad, los resultados de este estudio muestran una gran discrepancia entre la *historia oficial* que nos cuenta el Instituto Cervantes y lo que ocurre en la realidad:

Tabla 1: Resumen de los resultados

NIVEL	DELE INICIAL	DELE INTERMEDIO
Prueba	Int. de Textos Orales	Comprensión Auditiva
Núm. de Estudiantes	104	121
Núm. de Ítems	23	12
Puntuación Media	16,45	5,23
Desviación Típica	2,74	2,24
Coef. alfa de Cronbach (fiabilidad)	0,49	0,49

Además de analizar el coeficiente de fiabilidad, nuestro estudio no estaría completo si no intentáramos buscar algunas de las razones por las cuales se han obtenido niveles de fiabilidad tan bajos. Puesto que las limitaciones de espacio no nos permiten hacer un estudio exhaustivo, nos limitaremos al análisis de la calidad de los reactivos o ítem, factor que según expertos del campo de la evaluación de las lenguas extranjeras (Alderson, 1995; Bachman, 2004; Brown, 2005; Hughes, 2003) influye de una manera determinante en el nivel de fiabilidad. Por lo tanto, para examinar la calidad de los reactivos se realizó un análisis de la prueba de interpretación de textos orales del nivel inicial con ayuda del programa TAP o *Test Analysis Protocol* (Brooks, 2003-2005).

Tabla 2: Análisis de los ítems de la prueba de interpretación de textos orales

N.º de Ítem	Coef. de Dificultad	Índice de Discriminación	Correlación Biserial Puntual	Ítems problemáticos según diferentes criterios			
				Bachman (2004)	ALTE (2005)	Brown (2005)	U. Cambridge (2003)
01	0,91	0,13	0,33	X	X	X	X
02	0,98	0,06	0,25	X	X	X	X
03	0,81	0,36	0,40	X	X	X	
04	0,84	0,24	0,29	X	X	X	
05	0,63	0,46	0,33				
06	0,60	0,56	0,49				
07	0,52	0,21	0,24	X	X	X	X
08	0,93	-0,02	-0,03	X	X	X	X
09	0,59	0,40	0,37				
10	0,71	0,26	0,27	X	X	X	
11	0,54	0,42	0,39				
12	0,81	0,28	0,37	X	X	X	
13	0,81	0,01	-0,03	X	X	X	X
14	0,81	0,36	0,44	X	X	X	
15	0,78	0,26	0,27	X	X	X	
16	0,46	0,34	0,39				
17	0,27	0,22	0,24	X	X	X	X
18	0,79	0,16	0,20	X	X	X	X
19	0,78	0,09	0,09	X	X	X	X
20	0,47	0,22	0,20	X	X	X	X
21	0,73	0,21	0,31	X	X	X	
22	0,78	0,40	0,45	X		X	
23	0,91	0,18	0,28	X	X	X	X

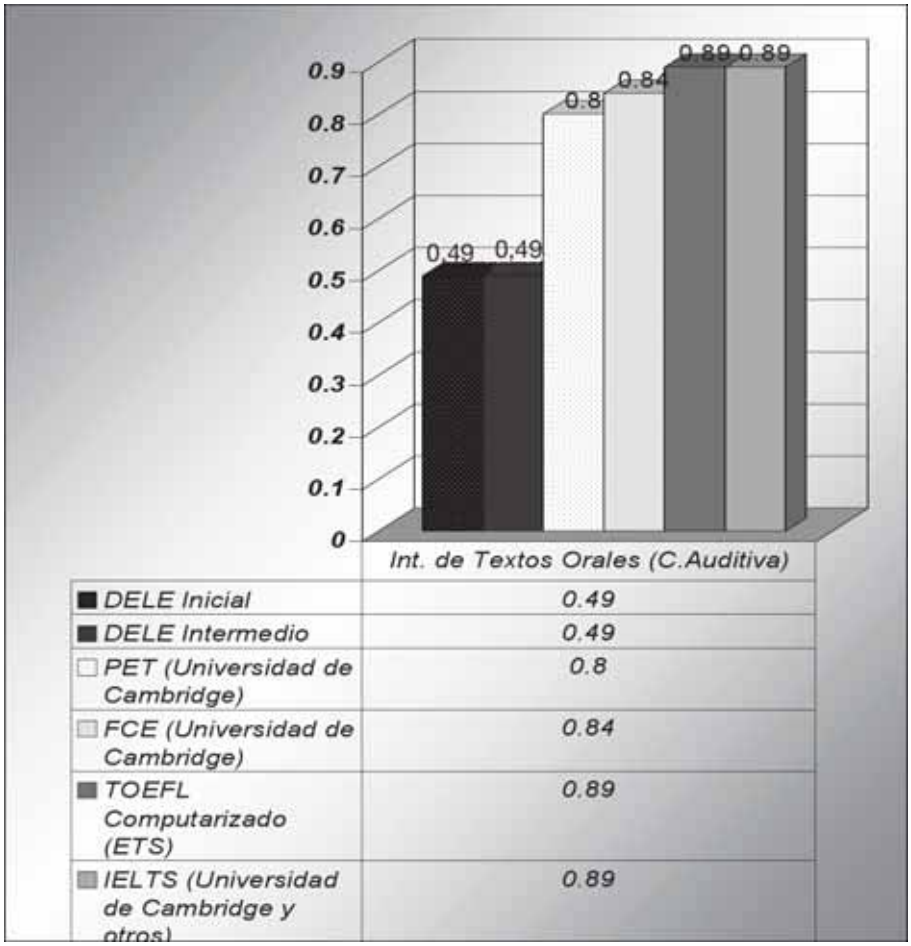
En la primera columna de la Tabla 2 tenemos el número de cada uno de los reactivos o ítems. A continuación, tenemos el coeficiente de dificultad o facilidad, que indica el porcentaje de estudiantes que dieron la respuesta correcta. En la tercera y cuarta columnas tenemos el índice de discriminación y la correlación biserial puntual, que en términos generales nos indica el grado o poder de discriminación de cada uno de los ítems, es decir qué tanto o en qué medida un reactivo es capaz de distinguir entre los examinados que saben y los que no saben. Cuanto más se acerque esta cifra a 1, mejor es el poder de discriminación del ítem.

Para juzgar la calidad de cada uno de los reactivos, nos auxiliaremos de los criterios de los expertos y las organizaciones involucradas en el campo de la evaluación que aparecen en la tabla. En cuanto al coeficiente de dificultad los expertos no parecen coincidir, pero éste puede variar entre un mínimo de 0,20 a un máximo de 0,85. Sin embargo, los expertos sí coinciden en el índice de discriminación que tiene que ser mayor a 0,30. Si procedemos a aplicar los criterios de Bachman (2004) y los de Brown (2005), dos especialistas en evaluación de renombre internacional, nos encontramos con un resultado alarmante, ya que sólo 5 de un total de 23 reactivos cumplen los requisitos especificados para un buen ítem (los cinco que *no* están marcados con X). Si utilizamos los criterios de ALTE (2005), la asociación de instituciones europeas que administran sistemas de evaluación de lenguas extranjeras, a la que pertenecen tanto el Instituto Cervantes como la Universidad de Salamanca, nuevamente nos llevamos otra gran sorpresa ya que sólo 6 de los reactivos cumplen con estos alineamientos: ítems n.º 5, 6, 9, 11, 16 y 22. Esto nos indica que a pesar de ser miembros de una asociación profesional de instituciones encargadas de administrar exámenes, la Universidad de Salamanca y el Instituto Cervantes no se están apegando a las guías o pautas que la organización propone para la elaboración de ítems y pruebas de calidad. Nuestro análisis estaría incompleto si no consultáramos los criterios que la Universidad de Cambridge (Saville, 2003) fija para sus exámenes. Después de todo, los DELE se han basado y es casi seguro que se seguirán basando en el sistema de evaluación de la institución inglesa. Nótese que con la aplicación de estos criterios la prueba de interpretación de textos orales sale mejor librada, ya que 13 de los reactivos sí cumplen con los requisitos. Obsérvese, sin embargo, que aun en este caso, más del 40% del total de los reactivos se considerarían como ítems problemáticos que tendrían que ser modificados o eliminados de la prueba. En resumen, síganse los criterios que se sigan, ya sean los de los expertos en evaluación, ya sean los alineamientos de las instituciones con las

que la Universidad de Salamanca y el Instituto Cervantes están vinculados, los datos muestran que la prueba de interpretación de textos orales tiene un gran porcentaje de ítem problemáticos o de mala calidad y ésta puede ser una de las razones principales por las que la prueba ha obtenido un coeficiente de fiabilidad tan bajo.

Para darnos una idea de lo grave que parece ser la situación, pasemos ahora al Gráfico 1 para comparar los coeficientes de fiabilidad de diferentes sistemas de evaluación en el mundo:

Gráfico 1: Comparación de coeficientes de fiabilidad de diferentes sistemas de evaluación



Con ella nos podremos dar cuenta del gran contraste que existe entre los coeficientes de fiabilidad obtenidos en los DELE y los proporcionados por

otros sistemas de evaluación. Obsérvese la gran diferencia que hay entre el coeficiente de 0,49 para las pruebas de interpretación de textos orales del DELE inicial y de comprensión auditiva del DELE intermedio y los coeficientes de componentes similares en el *Preliminary English Test* o PET (0,80 según reporta Geranpayeh en comunicación personal del 23 de enero de 2006), el 0,84 del *First Certificate in English* o *FCE* de la Universidad de Cambridge y el 0,89 del TOEFL (*Test of English as a Foreign Language*) y el IELTS (*International English Language Testing System*). Si bien es cierto que algunos podrían argumentar que no es justo comparar los DELE con sistemas de evaluación como el TOEFL y el IELTS por tratarse de exámenes con objetivos diferentes, sí podemos compararlos con los exámenes de los sistemas de evaluación a los que están tratando de emular, es decir, con el PET y el FCE. Los resultados que observamos en la gráfica demuestran que, aun si nos limitamos a estos dos casos, los coeficientes de fiabilidad de la pruebas de interpretación de textos orales en el DELE inicial y la prueba de comprensión auditiva en el DELE intermedio están muy por debajo de las cifras que reporta la Universidad de Cambridge para exámenes similares. Esto parece ser una indicación de que, desafortunadamente, el Instituto Cervantes y la Universidad de Salamanca, a pesar de haber adaptado los alineamientos y formatos de las pruebas del sistema de evaluación de la institución británica, no parecen haber adoptado los mismos criterios estándares o de calidad para asegurarse de que la fiabilidad de los resultados de sus instrumentos de evaluación esté a la altura de otros sistemas de evaluación en el mundo.

3. Conclusión

Los datos que aquí se han presentado indican que para los estudiantes japoneses que participaron en este estudio los resultados de las pruebas objetivas de interpretación de textos orales y comprensión auditiva para obtener tanto el nivel inicial como el nivel intermedio de los DELE tienen un nivel de fiabilidad extremadamente bajo. Es indudable que la posibilidad de incurrir en un gran porcentaje de error de medición en este tipo de pruebas amenaza la fiabilidad, y por ende, la validez del resultado global obtenido a través del sistema de evaluación DELE, ya que no debemos olvidar que un instrumento de medición que no mide con precisión lo que dice que mide no puede arrojar resultados válidos. Los datos proporcionados son evidencia de que para los participantes en el estudio la garantía de fiabilidad de los resultados del sistema de evaluación DELE a la que hace referencia el Instituto Cervantes en sus

publicaciones no parece ser efectiva. La *historia oficial* que nos cuenta el Instituto Cervantes, por lo tanto, dista mucho de lo que ocurre en realidad. Es verdaderamente una lástima que el Instituto Cervantes que tanto ha hecho por la difusión de la lengua y la cultura de los países hispanohablantes dañe su imagen con instrumentos de evaluación que están muy por debajo de los estándares de otros sistemas de evaluación internacional. Por lo tanto, a los usuarios del sistema, tanto individuos como instituciones, se les invita a interpretar con mucha cautela los resultados obtenidos en los exámenes DELE en la toma de decisiones, sobre todo en aquellas consideradas de alto impacto para los involucrados. Al Instituto Cervantes, responsable principal del sistema de evaluación, y a la Universidad de Salamanca, institución encargada de elaborar, calificar y evaluar las pruebas del DELE se les exhorta, por enésima vez, a asumir la responsabilidad profesional que tienen de mostrar con evidencias (y no sólo con palabras) que lo que aquí se ha presentado no sucede en otros contextos. Los miles de candidatos, quienes por cierto pagan una muy buena cantidad de dinero para ser evaluados en cada convocatoria para los exámenes DELE en todos los centros del mundo, merecen ser evaluados con instrumentos que arrojen resultados fiables, válidos y justos.

Bibliografía

ALDERSON, J. et ál.: *Language Test Construction and Evaluation*, Cambridge: Cambridge University Press, 1995.

ASSOCIATION OF LANGUAGE TESTERS IN EUROPE: *Materials for the Guidance of Test Item Writers,* 2005.

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION: *Standards for Education and Psychological Testing*, Washington, DC: AERA, 1999.

BACHMAN, L. F.: *Statistical Analyses for Language Assessment*, Nueva York: Cambridge University Press, 2004.

BROOKS, G. P.: *TAP, Test Analysis Program*, 2003-2005.

BROWN, J. D.: *Testing in Language Programs: A Comprehensive Guide to English Language Assessment*, Nueva York: McGraw-Hill ESL/ELT, 2005.

CÁRDENAS, A.: «La evaluación oral en los exámenes del Diploma de Español como Lengua Extranjera: una guía práctica para profesores y estudiantes», *Academia, Language and Literature,* vol. 69, Nagoya: Universidad de Nanzan, 2001, 25-67.

CÁRDENAS, A.: «La fiabilidad de los Diplomas de Español como Lengua Extranjera y los candidatos japoneses: ¿motivo de preocupación?», *Cuadernos de Canela*, vol. XVIII, (en prensa).

CITO, M. R. D.: *TiaPlus. Test and Item Analysis,* Amhem: Cito, 2006.

EDUCATIONAL TESTING SERVICES: *TOEFL CBT Score User Guide. 2000-2001 Edition,* Princeton: ETS, 2001.

EGUILUZ, J. et ál.: «Criterios para la evaluación de la producción escrita», *Didáctica del español como lengua* extranjera, Madrid: Fundación Actilibre, 1996, 75-94.

EGUILUZ, J. y Á. EGUILUZ: «La evaluación de la expresión escrita», *Vademécum para la formación de profesores. Enseñar español como segunda lengua (L2) / lengua extranjera (LE)* (Dir. SÁNCHEZ, J. e I. SANTOS), Madrid: SGEL, 2004a, 1.005-1.024.

EGUILUZ J. y Á. EGUILUZ: «La evaluación de la comprensión lectora», *Vademécum para la formación de profesores. Enseñar español como segunda lengua (L2) / lengua extranjera (LE)* (Dir. SÁNCHEZ, J. e I. SANTOS), Madrid: SGEL, 2004b, 1.025-1.041.

FERNÁNDEZ, J.: *Nuevos modelos de diplomas de español como lengua extranjera* [en línea] <http://cvc.cervantes.es/obref/congresos/rosario/ponencias/internacional/fernandez_j.htm>

GERANPAYEH, A.: Comunicación personal del 23 de enero de 2006.

HUGHES, A.: *Testing for Language Teachers*, Cambridge: Cambridge University Press, 2003.

JONES, N.: «Reliability as One Aspect of Test Quality», *Research Notes*, vol. 4, University of Cambridge, 2001, 2-5.

INSTITUTO CERVANTES: *Descripción general del sistema de evaluación DELE* [en línea] <http://diplomas.cervantes.es/docs/ficheros/200602080001_7_11.pdf>

MCNAMARA, T.: *Language Testing*, Oxford: Oxford University Press, 2000.

MINISTERIO DE EDUCACIÓN Y CIENCIA: *Guía para la obtención de los diplomas de español como lengua extranjera* (DELE), Madrid: Ministerio de Educación y Ciencia, 1992.

PARRONDO, J. R.: «Aspectos éticos de la evaluación: impacto de la actividad evaluadora y códigos deontológicos de los examinadores», *Carabela,* 55, 2004a, 31-44.

PARRONDO, J. R.: «Modelos, tipos y escalas de evaluación», *Vademécum para la formación de profesores. Enseñar español como segunda lengua (L2) / lengua extranjera (LE).* (Dir. SÁNCHEZ, J. e I. SANTOS), Madrid: SGEL, 2004b, 967-981.

PISONERO, I. y Á. GARCÍA: «Diplomas de español como lengua extranjera», *Cable: Revista de español como lengua extranjera*, 7, 1991, 3-4.

PRIETO, J. M. et ál.: «La elaboración de una prueba de nivel: la reforma de los DELE», *Carabela*, 55, 2004, 85-140.

SAVILLE, N.: «The Process of Test Development and Revision within UCLES-EFL», *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913-2002* (Ed. WEIR, C. J.), Cambridge: Cambridge University Press, 2003, 57-116.

SAVILLE, N.: «Setting and Monitoring Professional Standards: A QMS Approach», *Research Notes*, 22, Cambridge: University of Cambridge, 2005, 2-5.

WEIR, C. J.: *Language Testing and Validation: An Evidence-Based Approach*, Basingstoke: Palgrave Macmillan, 2005.