LA FIABILIDAD DE LOS DELE Y LOS CANDIDATOS JAPONESES: ¿UN MOTIVO DE PREOCUPACION?[1]

Abel CÁRDENAS MARTÍNEZ
abel@nanzan-u.ac.jp

Introducción

Para los que estamos involucrados en la enseñanza y evaluación del español como lengua extranjera (ELE), es indudable que ninguna otra institución en el mundo ha hecho lo que el Instituto Cervantes en pro de la difusión de la lengua y la cultura de los países hispanohablantes. El establecimiento de sedes en diferentes partes del mundo, la puesta en marcha del excelente Centro Virtual Cervantes, la organización de los congresos de la lengua, la reciente apertura del Centro de Formación de Profesores y la administración y participación en el sistema de evaluación de los Diplomas de Español como Lengua Extranjera, son ejemplos concretos de las importantes contribuciones que esta institución española ha aportado a nuestro campo de estudio. El gobierno español y el Instituto Cervantes merecen, por lo tanto, un reconocimiento de parte de todos los especialistas de ELE por la gran labor que vienen realizando desde hace ya varios años.

Así como debemos reconocer y aplaudir los aciertos, los profesionales de ELE también tenemos la responsabilidad de señalar aquellas áreas que merecen atención. Aunque este autor reconoce que en el sistema de evaluación de los Diplomas de Español como Lengua Extranjera o "los DELE" sí ha habido mejorías (la reciente renovación de la página Web de los diplomas, por ejemplo), es precisamente en este aspecto en donde queda todavía mucho camino por recorrer. A pesar del incremento en el número de países en los que se administran los exámenes y el constante aumento en el número de examinados, así como también la existencia de códigos de ética o práctica que lo requieren, hoy, a más de una década y media del inicio de este sistema de evaluación, siguen sin publicarse estudios que avalen la fiabilidad ni la validez de los resultados obtenidos 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 alguna de que los exámenes o mejor dicho los resultados de los exámenes son 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 éstas 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 decidió tomar la iniciativa para suplir esta deficiencia. El trabajo que aquí se presenta forma parte un proyecto a largo plazo que intentará encontrar evidencias empíricas

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

para poder contestar a una pregunta de vital importancia: ¿son fiables y válidos los resultados que se obtienen a través del sistema de evaluación DELE? El objetivo principal de este artículo es, pues, compartir con los lectores los primeros frutos de una investigación en la que nos propusimos explorar la primera parte de la pregunta: *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, p. 69).

Repaso de la Literatura

El llevar a cabo un repaso de la literatura sobre el tema que nos concierne es relativamente fácil, ya que prácticamente ésta es inexistente. Una búsqueda en la bibliografía de didáctica del español como lengua extranjera de la página Web del Instituto Cervantes terminó con un listado de 8 trabajos en los que se hace referencia a los DELE, pero ninguno relacionado con el concepto de fiabilidad en la evaluación. La búsqueda en otras publicaciones resultó igualmente desalentadora. El tema de los DELE o la fiabilidad de sus exámenes no parece ser de interés de los investigadores de ELE. Los contados trabajos que se encontraron (Pisonero y García Santa Cecilia, 1991; Eguiluz y Vega Santos, 1996; Eguiluz y Eguiluz, 2004a; Eguiluz y Eguiluz, 2004b; Fernández, 2004; Parrondo, 2004a; Parrondo, 2004b; Parrondo, 2004c; Prieto, 2004), incluyendo el de este autor (Cárdenas, 2001), son 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 D.E.L.E., publicada por el Ministerio de Educación y Ciencia, 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, p. 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 sus inicios. Posteriormente, en una publicación más reciente del Instituto Cervantes, la *Descripción General del Sistema de Evaluación DELE* se afirma que:

> *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*

[2] Énfasis del autor.

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 convocatoria a convocatoria, centro a centro [3] (IC 2005, p. 5)

El Instituto Cervantes como podemos ver en las cita de arriba, actualmente garantiza la fiabilidad de los resultados del sistema de evaluación respaldándose en la manera de calificar o corregir los exámenes, la uniformidad de condiciones y las medidas de seguridad con las que se administran los mismos. 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 evaluación en 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 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 elaborados por asociaciones profesionales tales como AERA (American Educational Research Association, 1999), ILTA (International Language Testing Association), entre otras son un recordatorio constante de que no es suficiente afirmar que los resultados de tal o cuál prueba son *altamente fiables o válidos*, sino que estas afirmaciones se deben avalar con evidencias. De hecho, el Código de Práctica y el programa de calidad en la evaluación de ALTE (Asociación Europea de Organismos Certificadores de la Competencia Lingüística), a la que pertenecen tanto el Instituto Cervantes como la Universidad de Salamanca, claramente especifican que sus miembros deberán asegurarse de cumplir con las siguientes responsabilidades en el aspecto de fiabilidad:

Documentar y explicar la manera en que se califica y se estima la fiabilidad, así como también la manera en que se recaban y se analizan los datos correspondientes relacionados al desempeño de los examinadores y calificadores de las pruebas de expresión oral y escrita.

Recabar y analizar datos de una muestra adecuada de candidatos para calcular el

[3] Énfasis del autor.

> *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*[4]*.* (Saville 2005, p. 3)

En resumen, con este breve repaso de la literatura nos podemos dar cuenta de que el estudio de la fiabilidad de los exámenes DELE no ha recibido la debida atención por parte de los investigadores del campo de ELE. Pero lo que resulta todavía más alarmante es que en un período de más de 15 años este aspecto no haya sido estudiado, documentado ni difundido por los administradores, creadores y calificadores de los exámenes: el Instituto Cervantes y la Universidad de Salamanca.

Objetivo del Estudio

El estudio que aquí presentamos es el primer intento de llenar este vacío en el campo de la evaluación del español como lengua extranjera. El objetivo general que nos fijamos era muy concreto: encontrar evidencias que avalaran 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 y especialmente dentro del contexto japonés. La pregunta específica que guió este trabajo es la siguiente: ¿Qué nivel de fiabilidad tienen los resultados de las pruebas de interpretación de textos orales (comprensión auditiva), interpretación de textos escritos (comprensión lectora) y conciencia comunicativa (gramática y vocabulario) de los niveles inicial e intermedio de los DELE al administrarse a un grupo de estudiantes universitarios japoneses?

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 en lo que normalmente en Japón se conoce como estudios hispánicos, que incluye aproximadamente un promedio de 462 horas de asignaturas de español como lengua extranjera en los primeros dos años de la carrera[5]. 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 1370 horas de clases de español.

---

[4] Énfasis del autor.

[5] El programa académico de los dos primeros años de la carrera incluye asignaturas de lengua española (gramática, conversación, lectura y composición), además de materias optativas y obligatorias de contenido no lingüístico (cultura) que se ofrecen en japonés.

Procedimiento

A cada uno de les grupos se les administró una versión oficial ya retirada de los exámenes para obtener el Diploma de Español como Lengua Extranjera del nivel correspondiente. Como es de todos conocido cada nivel consta de 5 pruebas en total: *interpretación de textos escritos (compresión de lectura), interpretación de textos orales (comprensión auditiva), conciencia comunicativa (gramática y vocabulario), producción de textos escritos, y expresión e interacción orales*. Las tres primeras son pruebas que en la literatura sobre evaluación (Alderson, 1995; Bachman, 1990; Bachman y Palmer, 1996; Brown, 2005) se conocen como tipo objetivo o corrección objetiva en las que los candidatos seleccionan entre dos o más opciones la respuesta que ellos consideren adecuada. Las pruebas de *producción de textos escritos y expresión e interacción orales,* por el contrario, implican la aplicación de criterios subjetivos por calificadores de la Universidad de Salamanca o miembros del tribunal local (véase Cárdenas 2001; Cárdenas 2005 para descripciones completas de las diferentes partes). Las pruebas del nivel inicial y del nivel intermedio se aplicaron respectivamente a dos y tres grupos diferentes de estudiantes. La supervisión de las cinco administraciones 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 respuestas fueron calificadas usando 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*.

Resultados y discusión

En la tabla No. 1 podemos ver un resumen del análisis de los resultados que se obtuvieron de la administración de las diferentes pruebas: el número de estudiantes, número de ítems o reactivos de cada prueba, puntuación media, etc. En términos generales, las pruebas del nivel inicial no representan realmente un gran grado de dificultad para los estudiantes del segundo año de la carrera en la universidad donde se realizó el estudio. Como se puede observar la prueba de *conciencia comunicativa* es la que resultó más fácil para los participantes, ya que obtuvieron una puntuación media del 23.21 que representa el 77% del total. La puntuación media más baja, se obtuvo en la prueba de *interpretación de textos escritos* o lectura, con un promedio de 13.63 (el 68% del total). La actuación de los integrantes del grupo intermedio, sin embargo, no fue tan buena como la de los del nivel inicial. El promedio más alto se obtuvo en la prueba de *compresión lectora* (62% del total), seguido de promedios todavía aún más bajos en las pruebas de *gramática y vocabulario* (56%) y *comprensión auditiva* (43%) lo que indicaría que para la mayoría de los integrantes del grupo intermedio las pruebas sí representan un mayor grado de dificultad.

Si nos concentramos en la última fila de la tabla, podemos ver el aspecto que más

nos interesa en este trabajo, es decir, el nivel 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, McNamara (2000) y Weir (2005), por ejemplo, consideran que éste debe ser superior a 0.90 cuando se trata de exámenes de alto impacto de uso internacional[6]. Si empezamos por lo positivo, podemos observar que en nuestro estudio la prueba de *conciencia comunicativa* en el nivel inicial obtuvo un coeficiente alfa de 0.82. Si bien esta cifra no es la que muchos expertos en evaluación considerarían como óptima para una prueba o examen de gran importancia, sí está dentro de un límite de lo aceptable. Si nos referimos a una prueba similar en el nivel intermedio, la situación empieza a cambiar, ya que el coeficiente alfa de 0.79 de la prueba de *gramática y vocabulario* no alcanza a estar dentro del mínimo aceptable.

Tabla No. 1 Resumen de los Resultados

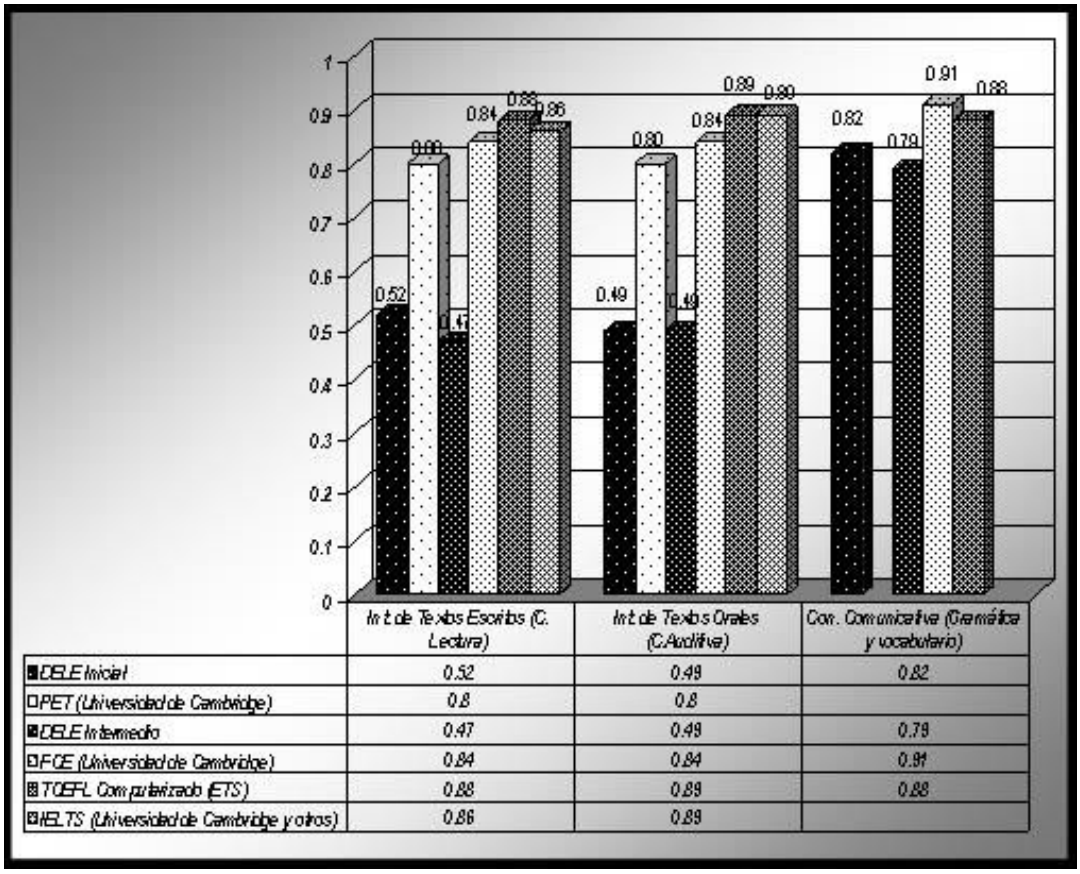
	DELE INCIAL			DELE INTERMEDIO		
	Int. de Textos Escritos	Int. de Textos Orales	Conciencia Comunicativa	Comprensión Lectora	Comprensión Auditiva	Gramática y Vocabulario
Número de Estudiantes	104			121		
Número de Ítems	20	23	30	12	12	60
Puntuación Media	13.63	16.45	23.21	7.46	5.23	33.85
Desviación Típica	2. 63	2.74	4.72	2. 15	2.24	7.55
Coeficiente alfa de Cronbach	0.52	0.49	0.82	0.47	0.49	0.79

Los resultados de las otras pruebas, desafortunadamente, muestran datos todavía más alarmantes. Para sorpresa del autor, los coeficientes de fiabilidad que se obtuvieron en la administración de la prueba de *interpretación de textos orales* (0.49) y la prueba de *interpretación de textos escritos* (0.52) en el nivel inicial son extremadamente bajos para una prueba de un sistema de evaluación de uso internacional. Si observamos los resultados de las mismas pruebas en el nivel intermedio, la situación es igualmente preocupante, ya que en este caso los coeficientes alfa que se obtuvieron (0.47 en la prueba de *compresión lectora* y 0.49 en la prueba de *comprensión auditiva*) también resultan ser bajísimos. Según los resultados de este

[6] Algunos expertos consideran que lo mínimo aceptable es 0.80.

estudio, podemos darnos cuenta de que ninguna de las tres pruebas de corrección objetiva en el nivel intermedio, ni dos de ellas de ellas en el nivel inicial alcanzan a obtener un coeficiente de fiabilidad aceptable. Si recordamos lo que en la literatura se dice sobre el significado de fiabilidad, se podría decir que los resultados que se obtienen a través de la aplicación de estas pruebas a los participantes en este estudio, se ven afectados o influenciados negativamente hasta en un 53 % por factores ajenos a los conocimientos o habilidades verdaderas de los candidatos. En lo que a fiabilidad concierne, por lo tanto, 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.

Gráfica No. 1. Comparación de Fiabilidad en Diferentes Sistemas de Evaluación en el Mundo



	Int. de Textos Escritos (C. Lectura)	Int. de Textos Orales (C.Auditiva)	Com. Comunicativa (Gramática y vocabulario)
DELE Inicial	0.52	0.49	0.82
PET (Universidad de Cambridge)	0.8	0.8	
DELE Intermedio	0.47	0.49	0.79
FCE (Universidad de Cambridge)	0.84	0.84	0.91
TOEFL Computarizado (ETS)	0.88	0.89	0.88
IELTS (Universidad de Cambridge y otros)	0.86	0.89	

La gráfica No. 1, en la que podemos ver una comparación de los coeficientes de fiabilidad de diferentes sistemas de evaluación en el mundo, nos muestra lo grave que parece ser la situación.[7] Si comenzamos nuevamente, por lo aceptable, nos podemos

[7] En la gráfica se han agrupado las pruebas de los otros sistemas de evaluación, con la prueba correspondiente o similar del sistema de evaluación DELE aunque en algunos casos

dar cuenta de que el coeficiente de fiabilidad de la prueba de *conciencia comunicativa* en el DELE inicial está cerca de los coeficientes que se reportan en otros sistemas de evaluación. Jones (2001), afirma que el coeficiente de fiabilidad de la prueba del Use of English del First Certificate (FCE) de la Universidad de Cambridge está en el rango del 0.91. El Educational Testing Service, en su boletín del 2000-2001, reporta un coeficiente de 0.88 para la prueba de Structure and Writing del mundialmente conocido TOEFL (Test of English as a Foreing Language). Si bien, el coeficiente de 0.82 de la prueba de *conciencia comunicativa* del DELE inicial no logra igualar o alcanzar las cifras reportadas por otros sistemas de evaluación, éste sí se podría considerar como aceptable. El coeficiente de fiabilidad del mismo componente en los exámenes del DELE intermedio (0.79), aunque cercano a la misma cifra, ya dejaría dudas en la mente de muchos especialistas en evaluación.

El gran contraste entre el sistema de evaluación DELE y los demás sistemas lo encontramos en los otros dos tipos de pruebas. Obsérvese la gran diferencia que hay entre el coeficiente de 0.49 para las pruebas de *interpretación de textos orales* y *comprensión auditiva* del DELE inicial y el 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 FCE de la Universidad de Cambridge y el 0.89 del TOEFL y el IELTS (International English Language Testing System). Nótese también la diferencia muy marcada entre los coeficientes de fiabilidad de las pruebas de *interpretación de textos escritos* y *comprensión lectora* del DELE (0.52 en el nivel inicial y 0.47 en el nivel intermedio) con las cifras de 0.80[8] o más en los otros sistemas de evaluación. Si bien es cierto que se podría 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 del sistema de evaluación en el que se han basado, es decir el PET y el FCE. Los resultados que observamos en la gráfica demuestran que aún si nos limitamos a estos dos casos, los coeficientes de fiabilidad de las pruebas de *interpretación de textos orales* o *comprensión auditiva e interpretación de textos escritos o comprensión lectora* de los DELE están muy por debajo de las cifras que reporta la Universidad de Cambridge para exámenes similares. Esto parece ser una indicación de que el Instituto Cervantes y la Universidad de Salamanca, a pesar de haber adaptado los lineamientos y formatos de las pruebas del sistema de evaluación de la institución inglesa, estas instituciones españolas, desgraciadamente, no parecen haber adoptado los mismos criterios de calidad o estándares para asegurarse de la fiabilidad de sus instrumentos de evaluación.

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

---

no se trata exactamente del mismo componente. Por ejemplo, en el caso del PET solo hay dos pruebas escritas: la primera es una prueba de comprensión auditiva y la segunda es una combinación de comprensión de lectura y expresión escrita.

[8] La fiabilidad de la prueba lectura y expresión escrita del PET está en el rango de 0.80 a 0.87.

obtenido estos coeficientes de fiabilidad tan bajos. La literatura sobre el campo de la evaluación de las lenguas extranjeras, (Alderson 1995; Hughes 2003; Bachman 2004; Brown 2005) señala varios factores que pueden interferir, entre los que se encuentran la longitud de las pruebas y la calidad de los ítems o reactivos. En cuanto al primer aspecto, la longitud de las pruebas, todo parece indicar que en nuestro estudio ésta podría ser una de las causas. Si observamos nuevamente las cifras de la tabla No. 1, nos daremos cuenta de que en general las pruebas más largas tienden a tener un coeficiente alfa de Cronbach más alto. Las pruebas de *gramática y vocabulario* o *conciencia comunicativa* en los dos niveles, con coeficientes de fiabilidad de 0.79 y 0.82 tienen entre 30 y 60 ítems o reactivos. Por el contrario, las pruebas de *interpretación de textos orales* e *interpretación de textos escritos* con 23 y 20 ítems en el nivel inicial y las de *comprensión lectora* y *comprensión auditiva* con 12 reactivos cada una en el nivel intermedio, son las que obtuvieron coeficientes de fiabilidad más bajos (entre el 0.47 y el 0.52). Existe pues también en nuestro estudio una aparente relación entre el número total de ítems y el coeficiente de fiabilidad.

Para examinar la calidad de los ítems o reactivos, se realizó un análisis exhaustivo de la prueba de *interpretación de textos orales*, que fue la que obtuvo el coeficiente de fiabilidad más bajo en el nivel inicial. La tabla No. 2 nos muestra parte de los datos obtenidos del análisis realizado con ayuda del programa TAP o *Test Analysis Protocol* (Broks 2003-2005). En la primera columna tenemos el número de cada uno de los reactivos. A continuación, tenemos el coeficiente de dificultad o de facilidad, que indica el porcentaje de estudiantes que proporcionó la respuesta correcta. En la segunda y la tercera columna tenemos el índice de discriminación y la correlación biserial puntual, que en términos generales nos indican 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 estudiantes o examinados que saben y los que no saben. Entre más se acerque a 1 esta cifra, mejor es el poder de discriminación.

Para juzgar la calidad de cada uno de los ítems o reactivos de la prueba, nos podemos auxiliar de los criterios de los expertos en el campo de la evaluación. Por ejemplo, para Bachman (2004), considerado como el mejor experto en el mundo de la evaluación de las lenguas extranjeras, un buen ítem debe tener un coeficiente de dificultad del 0.25 al 0.75 y un índice de discriminación de 0.30 o mayor. Brown (2005), otro de los expertos coincide con Bachman en el índice de discriminación, pero sugiere un coeficiente de dificultad del 0.30 al 0.70. Si revisamos los datos que tenemos en la tabla y aplicamos los criterios de estos dos expertos, nos llevaremos otra gran sorpresa, ya que solo 5 de un total de 23 reactivos cumplen con los requisitos especificados para un buen ítem (los cinco que no están marcados con X). Podríamos utilizar también los criterios de ALTE, 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. En sus guías para la elaboración de exámenes de lenguas extranjeras (ALTE, 2005) recomienda un coeficiente de dificultad del 0.20 al 0.80, y un índice de discriminación y una correlación biserial puntual de 0.30 o mayor. Nuevamente, si revisamos las cifras, nos encontramos con un resultado alarmante:

solo 6 de los reactivos cumplen con estos lineamientos: ítems no. 5, 6, 9, 11, 16 y 22.

Tabla No. 2. Análisis de Ítems de la Prueba de Interpretación de Textos Orales

No. de Ítem	Coef. de Dificultad	Índice de Discriminación	Correlación Biserial Puntual	Ítems problemáticos según diferentes criterios			
				Bachman (2004)	Brown (2005)	ALTE (2005)	U. de Cambridge (Saville,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

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 les propone para la elaboración de ítems de calidad. Nuestro análisis estaría incompleto, si no consultáramos los criterios que la Universidad de Cambridge fija para sus exámenes, después de todo, como se ha mencionado, los DELE se han basado y es casi seguro que se seguirán basando en el sistema de evaluación de la institución inglesa. Saville (2003), en un excelente trabajo sobre los exámenes de la Universidad de

Cambridge, indica que para ser considerados como adecuados, los reactivos deben tener un coeficiente de dificultad que va del 0.30 al 0.85 y una correlación biserial puntual de 0.25 o mayor. Nótese que con la aplicación de estos criterios, la prueba de *interpretación de textos orales* sale mejor librada, ya que 13 de los ítems o reactivos sí cumplen con los requisitos. Obsérvese, sin embargo, que aún 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. El gran porcentaje de ítems problemáticos o de mala calidad puede también ser, por lo tanto, una de las razones por las que la prueba de interpretación de textos orales ha obtenido un coeficiente de fiabilidad tan bajo.

Regresando a la pregunta del título de este trabajo, ¿es la fiabilidad de los resultados de los DELE un motivo de preocupación? Para contestar esta pregunta, recordemos brevemente la manera en que se determina si un candidato pasa o aprueba los exámenes DELE con la tabla de abajo.

Tabla No. 3 Sistema de Calificación en los DELE

GRUPO 1		GRUPO 2		GRUPO 3	
Int. de Textos Escritos o Comp. Lectora (Objetiva)	20 puntos	Conciencia Comunicativa o Gramática y Vocabulario (Objetiva)	20 puntos	Int. de Textos Orales o Comp. Auditiva (Objetiva)	15 puntos
E. Escrita (Subjetiva)	15 puntos			E. Oral (Subjetiva)	30 puntos
Total	35 puntos				45 puntos
Puntuación mínima para pasar (70 %)	**24.5 puntos**		**14 puntos**		**31.5 puntos**

De acuerdo con los resultados de nuestro estudio, la prueba de *conciencia comunicativa* en el nivel inicial y la prueba de *gramática y vocabulario* del nivel intermedio (grupo 2) obtuvieron un coeficiente de fiabilidad de 0.82 y 0.79 lo que indica que los resultados de estas pruebas para los estudiantes japoneses de este estudio están dentro (o casi dentro) de lo aceptable. Sin embargo, no podemos decir lo mismo de los grupos 1 y 3 (las columnas sombreadas en gris). Sabemos que las pruebas de *interpretación de textos escritos* y *comprensión lectora* y las pruebas de *interpretación de textos orales* y *comprensión auditiva* obtuvieron un coeficiente de fiabilidad comprendido entre 0.47 y 0.52, lo que indica que los resultados de estas pruebas tienen un gran porcentaje de error de medición (hasta un 53%) causado por factores ajenos a la verdadera habilidad de los estudiantes. Si esto sucede con las pruebas objetivas, que en teoría deberían tener un coeficiente de fiabilidad más alto, ¿qué se puede esperar del nivel de fiabilidad de las pruebas de expresión escrita y oral en donde entran en juego una serie de factores subjetivos (diferentes tareas, diferentes calificadores o miembros

del tribunal, etc.) que aumentan la posibilidad de error en la medición? Es obvio que al combinar una prueba objetiva, que ya sabemos tiene un bajo nivel de fiabilidad, con una subjetiva que es más susceptible a variaciones, como sucede en los grupos 1 y 3, que los resultados en estos grupos se verán afectados negativamente. Ahora bien, si sabemos que para obtener el resultado de apto o aprobado, los candidatos deben obtener como mínimo el 70% de puntos de cada uno de los grupos, y sabemos que los resultados de de dos de éstos grupos se verán afectados negativamente por el bajo nivel de fiabilidad, ¿no es esto un motivo de preocupación? Es obvio que sí, sobre todo si sabemos que para los candidatos el obtener una fracción de un punto menos de la puntuación mínima requerida en cualquiera de los tres grupos, puede significar un *no aprobado o no apto* en los exámenes.

Conclusión

Para concluir, 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 escritos* o *comprensión lectora* y la de *interpretación de textos orales* o *comprensión auditiva* de los exámenes para obtener tanto el nivel inicial como el nivel intermedio de los Diplomas de Español como Lengua extranjera tienen un nivel de fiabilidad extremadamente bajo. La posibilidad de incurrir en un gran porcentaje de error de medición en estas pruebas, indudablemente, repercute en la fiabilidad global, y por ende, en la validez del resultado total, ya que no debemos olvidar que un instrumento de medición que no mide con precisión lo que afirma medir, 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. 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 en el mundo. A los usuarios del sistema de evaluación DELE, tanto individuos como instituciones, por lo tanto, se les invita a interpretar con mucha cautela los resultados obtenidos en los exámenes 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, nuevamente se les exhorta a asumir la responsabilidad profesional que tienen de mostrar con evidencias 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 de los exámenes DELE en todos los centros del mundo, merecen ser evaluados con instrumentos que arrojen resultados fiables, válidos y justos.

Bibliografía

Alderson, J. C., Clapham, C. y Wall, D. (1995). *Language test construction and*

evaluation. Cambridge, Cambridge University Press.

Association of Language Testers in Europe (2005). *Materials for the guidance of test item writers*.

American Educational Research Association (1999). *Standards for education and psychological testing*. Washington, DC, AERA.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, New York, Cambridge University Press.

Brooks, G. P. (2003-2005). TAP, *Test Analysis Program*.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. New York, NY, McGraw-Hill ESL/ELT.

Cárdenas, A. (2001). La evaluación oral en los exámenes del Diploma de Español como Lengua Extranjera: una guía práctica para profesores y estudiantes. ACADEMIA Language and Literature 69: 25-67.

Cárdenas, A. (2005). La evaluación de la proficiencia en español como lengua extranjera: introducción y contextos. ACADEMIA Language and Literature 77: 191-208.

CITO, M. R. D. (2006). *TiaPlus*. Amhem, NL.

Educational Testing Services (2001). *TOEFL CBT score user guide. 2000-2001 Edition.* ETS.

Eguiluz, J. y C. M. Vega Santos (1996). Criterios para la evaluación de la producción escrita. Actas de Expolingua 1994-1995 (Cuadernos del tiempo libre). 3: 75-94.

Eguiluz, P. J. y P. A. Eguiluz (2004). La evaluación de la comprensión lectora. *Vademécum para la formación de profesores. Enseñar español como segunda lengua (L2)/lengua extranjera (LE).* L. J. Sánchez. Madrid, SGEL: 1025-1041.

Eguiluz, P. J. y P. A. Eguiluz (2004). La evaluación de la expresión escrita. *Vademécum para la formación de profesores. Enseñar español como segunda lengua(L2)/lengua extranjera (LE)*. L. J. Sánchez Madrid, SGEL: 1005-1024.

Fernández, G. J. (2004). Nuevos modelos de diplomas de español como lengua extranjera. III Congreso Internacional de la Lengua. 2005.

Geranpayeh, A. Comunicación personal del 23 de enero del 2006.

Hughes, A. (2003). *Testing for language teachers*. Cambridge, UK, Cambridge University Press.

Jones, N. (2001). Reliability as one aspect of test quality. ResearchNotes (4): 2-5.

Instituto Cervantes (2005). Descripción general del sistema de evaluación DELE, en http://diplomas.cervantes.es/docs/ficheros/200602080001_7_11.pdf

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

Ministerio de Educación y Ciencia (1992). Guía para la obtención de los diplomas de español como lengua extranjera (D.E.L.E.)

Parrondo, R. J. R. (2004a). Aspectos éticos de la evaluación: impacto de la actividad evaluadora y códigos deontológicos de los examinadores. Carabela 55: 31-44.

Parrondo, R. J. R. (2004b). El Instituto Cervantes y los diplomas de español como lengua extranjera.

Parrondo, R. J. R. (2004c). Modelos, tipos y escalas de evaluación. *Vademécum para la*

formación de profesores. Enseñar español como segunda lengua (L2)/lengua extranjera (LE). L. J. Sánchez. Madrid, SGEL: 967-981.

Pisonero, I. y A. García Santa-Cecilia (1991). Diplomas de español como lengua extranjera. CABLE: revista de español como lengua extranjera(7): 3-4.

Prieto, H. J. M., Díaz, S. C., Domínguez, L. Ch. y Martí, M. E. (2004). La elaboración de una prueba de nivel: la reforma de los DELE. Carabela 55: 85-140.

Saville, N. (2003). The_process of test development and revision within UCLES-EFL. *Continuity and innovation: revising the Cambridge Proficiency in English examination 1913-2002*. C. J. M. Weir, M. Cambridge, UK, Cambridge University Press. Studies in Language Testing 15: 57-116.

Saville, N. (2005). Setting and monitoring professional standards: a QMS approach. ResearchNotes 22: 2-5.

Weir, C. J. (2005). *Language testing and validation : an evidence-based approach.* Basingstoke, Palgrave Macmillan.