FIABILIDAD Y VALIDEZ EN TESTS DE LENGUA. A PROPÓSITO DE UN TEST DE ELECCIÓN MÚLTIPLE DE ESPAÑOL COMO LENGUA EXTRANJERA[1]

PETER JAN SLAGTER
Universidad de Utrecht

1. Introducción

En este artículo nos proponemos exponer algunos aspectos problemáticos de la construcción de tests de lengua -concretamente su fiabilidad y validez- teniendo en cuenta distintos aspectos de enfoques teóricos en torno a la definición de los rasgos que se pretenden medir. Para aclarar las claves de nuestra argumentación tenemos que hacer, en muy breves palabras, algunas consideraciones generales. Después ilustraremos nuestra exposición con ejemplos basados en experiencias sobre la elaboración de tests de lengua de elección múltiple (Rodríguez Diéguez, 1980: cap. 11). Este tipo de test es uno de cinco subtests de una serie de tests de semejantes características que se ha desarrollado y utilizado en la Universidad de Utrecht. Puede que llame la atención que se haya combinado *elección múltiple* con *enseñanza comunicativa*. Pese a las apariencias hemos comprobado que, dependiendo de cómo se formulen las preguntas, estos tests, y concretamente estos subtests, aportan datos que permiten pronunciarse sobre aspectos importantes de la competencia comunicativa de los hablantes en cuestión. En la presente contribución comenzaremos por centrarnos en aspectos de fiabilidad y validez (cfr. Rodríguez Diéguez, 1980: cap. 9).

2. Fiabilidad

En la interpretación de datos procedentes de tests en general, y de tests de lengua en particular, interesa conocer el valor que representan los distintos coeficientes que pueden calcularse para cada test y para cada uno

[1] Sin el 'software' y el incansable apoyo de Huub van den Bergh este artículo no tendría la forma que tiene. Los errores que contiene y la limitada sofisticación que caracteriza el análisis y las conclusiones se deben a quien firma estas páginas.

de sus ítemes. Cuando se trata de tests de ítemes independientes, como se supone que lo son los de elección múltiple, se suelen aplicar las fórmulas psicométricas comúnmente aceptadas para el tipo de test en cuestión. Una vez aceptada la hipótesis básica de un test de elección múltiple, la independencia interna de los ítemes, se suele aplicar alguna de las fórmulas *split-half* (división en dos) que mide la correlación entre una y otra parte del test. Las mitades pueden elegirse de diferentes formas, el coeficiente debería ser más o menos constante para no dar lugar a la sospecha de que, de hecho, se miden diversos rasgos en un solo test, por lo cual evidentemente ya no se permitiría utilizar ninguna de estas fórmulas.

3. Validez

A la hora de exponer los aspectos básicos del concepto de validez, la medida con que el test se aplicará al rasgo, suele insistirse en la validez aparente (la impresión que causa en los sujetos), la de contenido o constructo (el grado de convergencia que tiene con la materia o el rasgo que pretende medir, según la opinión de los expertos), la predictiva (pronósticos confirmados sobre el futuro desarrollo de la competencia de los sujetos) y la validez concurrente o validez de criterio (la correlación con otros tests que miden el mismo rasgo). Sobre la relación entre aspectos, y de ahí sobre cuáles tienen menos importancia y cuáles priman, difieren los expertos. Shohamy & Inbar (1991) resumen esta discusión señalando que Messick (1981, 1988) considera independientes la validez de contenido, la de criterio y la de constructo, y que Cronbach (1971) y Bachman (1990) emplean la de contenido y de criterio sólo como medio para obtener validez de constructo (Shohamy & Inbar, 1991: 23).

4. Competencia comunicativa

Observando las huellas de Hymes (1972) en sus propuestas de ampliación de la competencia lingüística chomskiana y los trabajos del equipo del Consejo de Europa a partir de esos mismos años, puede decirse que ya antes de publicar Canale & Swain en 1980 por primera vez su análisis de los aspectos más importantes de la competencia comunicativa, se había iniciado el debate sobre cuántas competencias existen realmente y sobre cómo se podrían llegar a distinguir. A partir de esta publicación se acepta generalmente que existen la competencia gramatical, la sociolingüística y la estratégica. Canale (1983), en una revisión parcial, añade a esta serie la competencia discursiva. Bachman & Palmer (1985) distinguieron dos tipos de competencia, la de organización y la pragmática, a las que se subordinan cuatro tipos: el gramatical, el discursivo, el ilocutivo y el sociolingüístico. Bachman (1990) afirma primero que "está

comúnmente aceptado que la *habilidad* lingüística no es una destreza única y unitaria, sino que consta de varios constructos diferentes relacionados que se añaden a un constructo general de *habilidad* lingüística" (Bachman 1990: 68), pero en el mismo libro Bachman se retracta en parte aclarando que la interpretación (un refinado análisis de rasgo múltiple y medida múltiple) que dieron de los datos (Bachman-Palmer, 1985), retrospectivamente, no parece tener una base tan sólida.[2] Así volvemos a una situación relativamente confusa: si no se puede demostrar que las competencias que se distinguen en teoría tienen una base empírica, ¿no tendremos simplemente que aceptar la existencia de una competencia única?

5. Enseñanza comunicativa y enseñanza estructural

Esta discusión, en principio teórica, tiene consecuencias prácticas. Las propuestas del Consejo de Europa y las numerosas publicaciones sobre la competencia comunicativa en general han hallado eco entre autores de libros de texto, editores y profesores de lenguas modernas. Entre los lingüistas especializados en cuestiones de adquisición, desde Oller y Krashen hasta hoy, las consecuencias que pudieran tener las disquisiciones sobre estas distintas competencias son mucho menores. Aquellos que enarbolaban la bandera de lo comunicativo en teoría debieron quedar confundidos con el resultado indeciso de la discusión, teniendo en cuenta que parecía haber menos justificación para arremeter contra los defensores de la tradición -aquellos que sólo pretenden enseñar la estructura de la lengua y no su uso-. Si todo viene a ser una misma cosa, ¿por qué molestarse? Dicho de otro modo: si no se detectan diferencias apreciables de nivel en cuanto a las distintas competencias en un mismo alumno, es decir, si se aprecia la misma proporción de aciertos en unos que en otros, o bien la enseñanza impartida ha tenido un éxito absoluto o es que se trata de factores inseparables.

6. Enseñanza comunicativa del español como lengua extranjera (ELE)

Los enfoques comunicativos -no "el enfoque comunicativo" porque nunca se ha pretendido una sola versión del evangelio- hallaron pronto más eco

[2] "As has been emphasized throughout this book [Bachman 1990], the question of how many component abilities there are, and how, if at all, they are related to each other, is essentially an empirical one. The specific component abilities included in the Bachman-Palmer scale are given here primarily for illustrative purposes, and not as a definitive model of language proficiency. Indeed, Bachman and Palmer's own research indicates that this particular configuration of components is only partially supported by empirical results". (Bachman, 1990: 358, nota 1)

entre autores de manuales que entre lingüistas teóricos. De hecho, la mayor parte de la investigación sobre el español que pudiéramos calificar de funcional era coproducto de la confección de materiales didácticos (entre mis favoritos los libros del Equipo Pragma, y más recientemente *Intercambio*), mientras que otros hicieron suya la calificación de "comunicativo" defendiendo una postura mucho más gramatical y formal -en opinión mía demasiado- para poder llamarse así. Es sabido que los enfoques se formulan, se llevan al aula, pero no logran su última versión operativa hasta plasmarse en tests de lengua. Llegados a ese punto, defensores de una y otra postura terminan enseñando sus cartas. En nuestras clases en Utrecht hemos intentado respetar la esencia de lo comunicativo tal y como se encontraba en los dos títulos del Equipo Pragma, *Para empezar* y *Esto funciona*. En los párrafos que siguen daremos cuenta de algunos aspectos de esta labor y expondremos cómo nos fue.

7. Tests comunicativos y tests DP

Farhady (1983) resume en breves palabras la controversia entre los tests de *discrete point* (DP) y los tests *integrados* (IN). Parecería lógico que en un enfoque comunicativo, funcional o nocio-funcional se optase por tests integrados que tuvieran en cuenta tanto la interacción que se crea en la clase de lengua como el contexto de uso auténtico, la vida real (Baker 1989, Hughes 1989). Farhady, sin embargo, realza los puntos positivos de los tests DP y presenta un ejemplo de cómo redactar ítemes DP para un test nocio-funcional (NF) centrado en dos funciones, expresión y comprobación de actitudes intelectuales y persuasión (van Ek 1976, Slagter 1979), describe además cómo preparó un test de competencia comunicativa para el inglés como segunda lengua. Llama la atención su manera de caracterizar las alternativas que selecciona para cada contexto comunicativo elegido. Las alternativas de cada pregunta comprenden cuatro variantes: (1) socialmente adecuada y lingüísticamente correcta, (2) socialmente adecuada pero lingüísticamente incorrecta, (3) lingüísticamente correcta, pero socialmente inadecuada, (4) incorrecta e inadecuada en ambos sentidos. Tras múltiples consultas a hablantes nativos y pretesting con sucesivas generaciones de alumnos se han conseguido cada vez mejores versiones del test en cuestión. En cuanto a los ítemes que analizaremos en este artículo veremos propuestas más centradas en aspectos lingüísticos que las de Farhady.

8. Tests de *proficiencia* frente a pruebas de nivel

Teniendo en cuenta las consideraciones sobre las distintas competencias, las distintas propuestas nocio-funcionales que no detallamos aquí y las

experiencias de Farhady, diseñamos en el marco de una serie de tests de nivel sucesivos -que incluían tests parciales de diverso carácter, como traducción holandés-español, comprensión lectora, un test de lagunas y un test de conocimientos morfológicos- una serie de tests de elección múltiple correspondientes a distintos niveles y basados en los materiales que usamos, pero procurando crear en la medida de lo posible, pequeños contextos nuevos con material fácilmente reconocible, para probar la destreza del alumno ante una situación o un problema sociolingüístico nuevo. No nos interesaba que los alumnos reprodujeran conocimientos. Nuestra aspiración era construir tests de habilidad, no de sus logros, pero por la forma de trabajar que adoptamos es evidente que nuestros tests tienen más de lo segundo (¿en qué medida dominan los alumnos lo que les enseñamos?) que de habilidad. Obviamente es cuestionable que se midan las destrezas comunicativas por medio de tests escritos. La línea comunicativa se refleja porque se trata de intercambios entre dos personas, por los temas y por la elección de elementos propios del lenguaje hablado. Presentándoles este tipo de preguntas a los candidatos, en realidad les pedimos un juicio sobre la adecuación de tal o cual variante a un contexto creado. No negamos que este procedimiento refleje más sus conocimientos que su dominio oral en situaciones reales. Para medir este último aspecto disponemos de otros tests. Los ítemes que reseñaremos a continuación comprenden preguntas que reflejan una situación de uso en la que el candidato responde eligiendo la segunda parte que considera adecuada de un par adyacente o rellena una laguna en el intercambio. Redactar este tipo de preguntas es una tarea interesante, creadora, pero muy laboriosa. Redactarlas, pedir la opinión de los profesores responsables de estos grupos de estudiantes, consume cantidades ingentes de tiempo por el pretesting, la interpretación de los datos provisionales y las consiguientes revisiones.

9. Construcción y análisis psicométrico del test

El test que comentaremos se aplicó por primera vez a principios de enero de 1990. Se trata de un test de repesca para el primer módulo (cursillo equivalente a 160 horas de estudio, de las que 30 son lectivas) sobre *Para empezar A*. Se administró sin pretesting alguno. Esto se debe a que, por una serie de razones que no detallaremos aquí, fue a partir de esas fechas cuando emprendimos la tarea de dedicar más atención al análisis de nuestros propios tests. En estas circunstancias sólo es posible ajustar a posteriori las normas de aprobados y suspensos. En sucesivas administraciones hemos podido comprobar que las correlaciones entre este subtest y los demás de cada examen por separado eran satisfactorias. No

hemos podido encontrar pruebas de que en este subtest de elección múltiple se tratase de un constructo diferente del de los demás subtests. En vista de la discusión que sigue, y debido a efectos pedagógicos, aún no hemos decidido suprimir los demás subtests, por lo que pese a la fiabilidad y facilidad de administración y corrección que ofrece, nos hemos quedado sólo con este (en este contexto quizás sea procedente añadir que tampoco hemos podido comprobar con análisis factoriales a posteriori que el subtest de elección múltiple se pudiera subdividir en conjuntos de preguntas diversas, unas posiblemente más relacionadas con factores como la competencia gramatical, u otros relacionables con otras competencias).

Tomaron parte en el examen 75 estudiantes, su puntuación media fue de 29.187, la desviación estándar 5.844, y la varianza 34.154. El coeficiente de fiabilidad fue 0.725 (alpha o KR 20),[3] y el error típico 3.062.

10. Análisis de ítemes

El análisis de cada ítem evidenció a posteriori que algunos ítemes tenían un valor negativo en la correlación ítem - resto de los ítemes. Quiere ello decir que puede que hubiera candidatos que a pesar de responder correctamente a las demás preguntas, fallaron en estos ítemes. Esto suele indicar que se trata de un ítem defectuoso. Veamos algunos de ellos con más detalle (los números son los originales del test completo):

24. \+ Oye, ¿puedes tú escribirle una carta a Julio Romero de Torres?
o ¿Quién, yo? ———————————————

A. No recuerdo.
B. No tiene teléfono.
C. ¿Por qué me lo dices a mí?
D. Venga, hombre, sí que puedes hacerlo tú.

ITEM	ALTERNATIVAS	CANDS	PORCENTAJES
24	A	4	0.053
24	B	1	0.013
24	C	42	0.560
24	D	27	0.360

[3] Para los análisis hemos utilizado siempre la fórmula KR-20.

Habíamos calificado la respuesta D como la única correcta, porque parecía combinarse con la negativa implícita en el intercambio. Todos los candidatos menos 27 discreparon con razón, porque aunque nuestra argumentación puede que sea ingeniosa, el contexto deja mucho que desear. Las respuestas A y B son tan claramente inadecuadas que se descartan con facilidad, reduciendo la elección múltiple de cuatro a dos opciones. Es sorprendente que un solo candidato optara por B: ¿qué relación lógica hay entre no tener teléfono y tener que enviar una carta? La alternativa C tampoco es apropiada porque sea la única correcta, sino más bien por ser la más plausible o la menos improbable. La correlación ítem/test de esta pregunta es -.158.

28. \+ ¿Qué te pasa, Fernandito?
o Nada, mamá, sólo que Maribel ha dicho que no quiere jugar conmigo ya porque _____

Pero ya encontraré otra amiga para jugar con el Lego.

A. no le gusto
B. no me gusta
C. no nos gusta
D. no te gusto

ITEM	ALTERNATIVAS	CANDS	PORCENTAJES
28	A	34	0.453
28	B	35	0.467
28	C	5	0.067
28	D	1	0.013

Puede ser que el uso de "gustar" para expresar gustos y aficiones y la posición contraria sea un problema digno de mayor atención y evaluación, si bien esta no es la manera más adecuada de dedicarse a ello. Nos encontramos con una pregunta de dos alternativas, A y B. Fernandito puede haber expresado su poca simpatía por Maribel y viceversa, pero el candidato eso no lo puede saber. La correlación ítem/test es -.151.

29. \+ ¿Por qué llevas unos zapatos tan feos?
o _____

A. ¿Cómo? ¿Los traes tú?
B. Porque es lo más barato.
C. Porque están lloviendo.
D. Porque están un regalo.

Esta pregunta se centra en varios problemas a la vez y ninguno de ellos parece de tipo comunicativo. La alternativa A insiste en la diferencia entre "llevar" y "traer", sin ser por ello una respuesta lógica. La respuesta B es la correcta, la C contiene un plural (están) imposible y D hace hincapié en la diferencia entre "ser" y "estar". "Estar" no se puede combinar con un sustantivo como predicado, error en el que incurren con facilidad los principiantes por distintas razones. Este ítem ilustra un tipo de pregunta que habría que evitar a toda costa, pero que en la práctica se da con gran facilidad: se confunden aspectos pragmático-discursivos con elementos puramente morfológicos. Si partimos con Canale & Swain (1980) y Swain (1983) de la idea de que la competencia comunicativa se compone de diferentes subcompetencias, deberíamos evitar mezclar elementos en una misma pregunta; si pensamos que estas diferencias son ociosas habría que tener en cuenta los efectos *backwash*, es decir, la retroalimentación que se produce en el alumno estudiado. ¿Qué le estimula a fiarse de sus conocimientos sobre el mundo que le rodea? Preferirá depositar su confianza en conocimientos morfosintácticos. Los resultados son los siguientes:

ITEM	ALTERNATIVAS	CANDS	PORCENTAJES
29	A	9	0.120
29	B	15	0.200
29	C	7	0.093
29	D	44	0.587

La correlación ítem/test de esta pregunta es -.287.

32. \+ ¿Qué haces tú todavía aquí en la cantina?
o Es que tengo otra clase a las tres y ————————

A. no ha terminado la clase todavía.
B. tengo que ir a casa.
C. ya no he terminado mi traducción todavía.
D. ya no voy a casa.

Para empezar, el candidato tiene que saber que las clases en nuestra facultad son de dos horas siempre, de 9 a 11, de 11 a 1, de 1 a 3 y de 3 a 5. A las cinco acaba la jornada y todo el mundo se va a casa a cenar a las seis o seis y media. Dato cultural. Además, tiene que tenerse en cuenta que este intercambio se produce probablemente entre la una y las tres.

También puede ser antes de la una. La respuesta A, según creo, es incorrecta siempre, B sugiere que el segundo interlocutor irá a casa en coche con otra persona a la que no tiene más remedio que esperar (explicación harto rebuscada), C contiene la frecuente confusión de "ya no" y "todavía no", y D se había calificado como respuesta correcta. La correlación item/test es de -.161.

ITEM	ALTERNATIVAS	CANDS	PORCENTAJES
32	A	7	0.093
32	B	1	0.013
32	C	31	0.413
32	D	36	0.480

11. Coeficientes de dificultad y correlación ítem/test

Como primer paso, a la hora de valorar la calidad de nuestro test podemos analizar los coeficientes de dificultad. Una respuesta que eligieron menos de la mitad de los candidatos o más del 75% suele considerarse difícil y demasiado fácil respectivamente. Hemos elegido estas preguntas, no por su relativa dificultad o facilidad, sino por su coeficiente negativo de correlación ítem/test. Un valor negativo, hemos dicho antes, es señal de que ese ítem funciona en sentido contrario a como queremos que discrimine: confunde precisamente a los candidatos que contestan correctamente a las demás preguntas. Y, como es lógico, no es justo achacar las deficiencias de construcción del test a los candidatos.

Eliminar estas cuatro preguntas del cómputo final haciendo como si no existieran y como si los candidatos no las hubieran contestado jamás, tiene un resultado a primera vista sorprendente. Los resultados psicométricos calculados con las mismas medidas que la versión íntegra del test ofrecen la siguiente imagen:

	ORIGINAL	CORREGIDA
Candidatos	75.000	75.000
Puntuación media	28.187	27.693
Desviación stándard	5.844	6.105

Varianza	34.154	37.270
Coeficiente de fiabilidad	0.725	0.988
Error típico de medición	3.062	0.673

Consignar tres dígitos detrás de la coma, sobre todo cuando se trata de candidatos, evidentemente es excesivo, pero el programa de análisis que utilizábamos por esas fechas se ha molestado en ofrecerlos. Lo que llama la atención es el espectacular crecimiento del coeficiente de fiabilidad. Este tipo de mejora, lamentamos tener que confesar, no se ha repetido jamás en ninguna otra de estas operaciones y sólo se explicaría por una casi ideal distribución de los niveles de destreza de los candidatos.

12. Validación con hablantes nativos

Con estos datos podríamos dar por finalizado nuestro análisis: con un 0.98, 0.99 casi, los candidatos tienen una máxima garantía de que sus notas son buenas y de que el que acaba primero en la lista efectivamente es el mejor de todos en esta prueba, y así sucesivamente.

El hablante nativo no existe. Un grupo de hablantes nativos tampoco se pondrá fácilmente de acuerdo, por ejemplo, en cuestiones de estilo o de adecuación sociolingüística, pero es lógico suponer que un panel de hablantes nativos contestará a todas las preguntas de este test con absoluta unanimidad. Haciéndolo así se proporcionan pruebas independientes de la calidad de los ítemes. Veamos los resultados obtenidos con 16 profesores de español como lengua extranjera durante un cursillo de reciclaje celebrado en Málaga en la primavera de 1990. En este artículo me centro en sus respuestas a las preguntas del test. También comentaron aspectos que afectan a la validez del mismo. Ahora bien, cuando todos ellos optan por una misma respuesta y ésta se corresponde con la clave, no hay ningún problema, como lo demuestran los datos en la mayoría de las preguntas. Si no contestan unánimemente ello quiere decir que o bien las preguntas son ambiguas o simplemente defectuosas.

Por eso serán un motivo de invalidez las dudas con las que se verían confrontados los hablantes nativos cuando tengan que responder ante este mismo test. El apartarse de la respuesta correcta se podría explicar por efecto del azar, del error típico, etc. Nadie es perfecto. Pero si los alumnos holandeses aciertan mejor que los hablantes nativos, ¿qué pasa? A los hablantes nativos simplemente no pueden resultarles más difíciles las preguntas que a los hablantes no nativos holandeses. La segunda pregunta es la siguiente: ¿qué ítemes tienen una varianza más elevada en las respuestas de los hablantes nativos del español, reflejando así mayores

dudas a la hora de contestarlas? He aquí los datos del grupo de hablantes nativos al que antes aludíamos:

Sujetos nativos:	16
Número de preguntas:	50
Puntuación media:	43.81
Varianza:	3.58
Homogeneidad/fiabilidad:	0.01
Error típico:	1.88

Datos completos del grupo de hablantes nativos:

ITEM	PROMEDIO	VARIANZA	RIT
1	0.62	0.25	0.08
2	1.00	0.00	0.00
3	1.00	0.00	0.00
4	0.75	0.20	-0.48
5	0.94	0.06	-0.02
6	0.94	0.06	-0.29
7	0.81	0.16	0.09
8	1.00	0.00	0.00
9	0.44	0.26	-0.58
10	0.94	0.06	0.43
11	1.00	0.00	0.00
12	1.00	0.00	0.00
13	1.00	0.00	0.00
14	0.94	0.06	-0.16
15	1.00	0.00	0.00
16	0.12	0.12	-0.51
17	0.88	0.12	0.32
18	0.75	0.20	-0.28
19	1.00	0.00	0.00
20	1.00	0.00	0.00
21	1.00	0.00	0.00
22	0.88	0.12	0.44
23	0.94	0.06	0.43
24	0.06	0.06	0.04
25	0.94	0.06	-0.02
26	0.56	0.26	0.13
27	1.00	0.00	0.00
28	1.00	0.00	0.00
29	0.81	0.16	-0.17
30	0.94	0.06	0.59
31	1.00	0.00	0.00
32	1.00	0.00	0.00
33	1.00	0.00	0.00
34	0.88	0.12	0.20
35	0.94	0.06	0.43
36	0.75	0.20	-0.06
37	1.00	0.00	0.00
38	0.75	0.20	-0.06
39	1.00	0.00	0.00
40	0.81	0.16	0.28
41	0.94	0.06	0.59
42	0.81	0.16	-0.33
43	1.00	0.00	0.00

44	1.00	0.00	0.00
45	1.00	0.00	0.00
46	1.00	0.00	0.00
47	1.00	0.00	0.00
48	1.00	0.00	0.00
49	0.75	0.20	0.58
50	0.94	0.06	-0.16

En esta tabla damos datos redondeados. Estos mismos datos se representan en los gráficos que se reproducen en los apéndices.

Cuando todos los candidatos responden correctamente a todas las preguntas, el test, como es lógico, deja de discriminar entre buenos y malos candidatos. Recogemos los datos de los informantes nativos en el apéndice 4. Con 16 de ellos un solo fallo significa el 6% de disminución en el porcentaje de aciertos. A continuación y para mayor claridad, reproducimos todos los datos originales, redondeados de la misma forma, del grupo de alumnos holandeses y los datos de los hablantes nativos. El subrayado marca los casos anómalos que nos proponemos comentar.

Candidatos:	75
Preguntas:	50
Puntuación media:	29.19
Varianza:	9.87
Homogeneidad/Fiabilidad:	0.73
Error típico:	1.65

ITEM	PROMEDIO Hol./España	VARIANZA Hol./España	RIT alumnos holandeses
1	0.27/0.62	0.20-0.25	0.04
2	0.08/1.00	0.07-0.00	0.21
3	0.37/1.00	0.24-0.00	0.19
4	0.53/0.75	0.25-0.20	0.31
5	0.88/0.94	0.11-0.06	0.17
6	0.57/0.94	0.25-0.06	0.22
7	0.37/0.81	0.24-0.16	0.12
8	0.72/1.00	0.20-0.00	0.32
9	0.44/0.44	0.25-0.26	0.14
10	0.53/0.94	0.25-0.06	0.19
11	0.93/1.00	0.06-0.00	0.26
12	0.83/1.00	0.15-0.00	0.22
13	0.71/1.00	0.21-0.00	0.29
14	0.52/0.94	0.25-0.06	0.25
15	0.16/1.00	0.14-0.00	0.13
16	0.41/0.12	0.25-0.12	0.30
17	0.21/0.88	0.17-0.12	0.06
18	0.75/0.75	0.19-0.20	0.09
19	0.92/1.00	0.07-0.00	0.04
20	0.77/1.00	0.18-0.00	0.05
21	0.92/1.00	0.07-0.00	0.32
22	0.55/0.88	0.25-0.12	0.30
23	0.57/0.94	0.25-0.06	0.23
24	0.36/0.06	0.23-0.06	-0.16

25	0.40/0.94	0.24-0.06	0.06
26	0.89/0.56	0.10-0.26	0.32
27	0.92/1.00	0.07-0.00	0.30
28	0.45/1.00	0.25-0.00	-0.15
29	0.20/0.81	0.16-0.16	-0.29
30	0.80/0.94	0.16-0.06	0.30
31	0.45/1.00	0.25-0.00	0.21
32	0.48/1.00	0.25-0.00	0.16
33	0.61/1.00	0.24-0.00	0.09
34	0.47/0.88	0.25-0.12	0.24
35	0.81/0.94	0.15-0.06	0.07
36	0.52/0.75	0.25-0.20	0.14
37	0.55/1.00	0.25-0.00	0.12
38	0.61/0.75	0.24-0.20	0.19
39	0.75/1.00	0.19-0.00	0.57
40	0.39/0.81	0.24-0.16	0.28
41	0.79/0.94	0.17-0.06	0.26
42	0.51/0.81	0.25-0.16	0.36
43	0.60/1.00	0.24-0.00	0.40
44	0.68/1.00	0.22-0.00	0.39
45	0.88/1.00	0.11-0.00	0.47
46	0.83/1.00	0.15-0.00	0.64
47	0.35/1.00	0.23-0.00	0.25
48	0.83/1.00	0.15-0.00	0.45
49	0.53/0.75	0.25-0.20	0.15
50	0.51/0.94	0.25-0.06	0.06

Estos mismos datos se recogen en el gráfico 2 que reproducimos en los apéndices.

13. Interpretación

En esta discusión sobre los casos defectuosos es fácil pensar que se trate de un test inaceptable. Véanse, sin embargo, los datos de la primera columna de la tabla precedente, donde los coeficientes que van antes de la barra (/) corresponden a los holandeses, para comprobar que el conjunto de ítemes es variado y que no es ni demasiado fácil ni demasiado complicado. En esta misma columna, la de los promedios de aciertos holandeses/españoles, llaman la atención los ítemes 9, 16, 18, 24 y 26; en la de las varianzas son los ítemes 1, 9, 18 y 26. Tres de estos 9 ítemes se presentan como sospechosos en ambas cuentas. Hay que recordar que sólo dedicamos atención a la pregunta 24 en función de los resultados de los sujetos holandeses.

Llaman ahora la atención otras preguntas que todavía no habíamos tomado en consideración. Se trata de las siguientes:

1. \+ ¿Tú crees que este examen va a ser más fácil que el anterior?
   o ———————————————————

   A. ¿He dicho yo eso?
   B. ¿No crees que tiene que ser el mismo que la otra vez?
   C. ¿No te gustará esa idea?
   D. ¿Por qué preguntas eso ahora cuándo es demasiado tarde?

Los informantes holandeses contestaron de esta manera:

ITEM	ALTERNATIVA	CANDS	PORCENTAJE
1	A	20	26.67
	B	36	48.00
	C	3	4.00
	D	16	21.33

Los hablantes nativos optaron por A (10 veces), B (2 veces) y D (4 veces). Los estudiantes holandeses no se dan cuenta de "*el* mismo" en la alternativa B, al igual que algunos de los hablantes nativos. La alternativa D resulta sugerente a muchos estudiantes holandeses, lo mismo que a cuatro hablantes nativos que no se fijan en la tilde en "cuándo". Entre estos últimos una clara mayoría opta por la respuesta correcta. Sería más convincente que todos ellos eligieran la primera.

9. \+ Claudia, éste es Juanjo, un amigo mío. Juanjo, te presento a Claudia, mi mejor amiga.

o _____

A. ¿Cómo está usted, señorita?
B. ¿Estás bien?
C. Hola, Claudia.
D. Hola, ¿cómo estás?

He aquí las respuestas de los estudiantes holandeses en datos absolutos y en porcentajes:

A	6	8.00
C	33	44.00
D	36	48.00

Los hablantes nativos españoles eligen siete veces la respuesta C y nueve la D. La mayor parte de los holandeses se decide por D (en la que falta la tilde en "¿cómo?") y muchos menos por C (la única correcta, en mi opinión). Dan que pensar las respuestas de los nativos como grupo. Si bien es verdad que la ortografía y la puntuación españolas parecen ser un escollo insalvable para los extranjeros y motivo de desprecio absoluto por parte de los nativos, un análisis más detenido debería haber captado la atención de estos últimos. Lo más coherente es señalar que las respuestas C y D son adecuadas en un contexto comunicativo, que A es aceptable

dependiendo del contexto y, por fin, que B es extraña. Con ello habremos conseguido sorprender a los nativos poco atentos, pero que sea o no correcto incluir ítemes de este tipo puede ser discutible.

16. + ¿Conoces ya a muchos españoles? ¿Qué opinas de nosotros? Y concretamente, ¿qué piensas de mí?

o ———————————————

A. Me han parecido muy simpáticos tus compañeros.
B. ¿No tienes abuela?
C. Nos conocemos ahora exactamente quince minutos.
D. Pues, *muchísimo, ¿comprendes?*

16

A	21	28.00
B	2	2.67
C	31	41.33
D	21	28.00

En la respuesta A parece quererse escurrir el bulto, pero ¿cómo saber si no se trata de un amante latino, insensible a esta pulla? La respuesta B es la mejor desde una perspectiva sociolingüística, el que no tiene abuela habla así. La C contiene un error estructural: "desde hace quince minutos" o "hace exactamente quince minutos que nos conocemos". La D es imposible porque no se puede opinar "mucho" sobre una persona. Los informantes nativos tienen opiniones divididas: 8 a favor de A, 6 optan por B y 2 por C. Demasiada varianza e indecisón por parte de ellos para mantener la pregunta.

26. + Hola, buenas noches, perdone que le moleste, soy el vecino de abajo. ¿Podría subir un poco el volumen de la tele? Es que se ha estropeado el sonido de la tele. La imagen está bien, o sea que si vuelvo a bajar yo y si usted es tan amable...

o Claro, hombre, ———————————————
¿Cuál de las siguientes reacciones ***no*** *es adecuada*?

A. ahora mismo.
B. claro.
C. pase, pase.
D. un momento.

Los sujetos holandeses se deciden como sigue:

26

A	7	9.33
B	1	1.33
C	67	89.33

Esta situación es opuesta a lo normal: es normal llamar la atención del vecino por exceso de ruido, no por su ausencia. Todas las alternativas inducen a confusión, máxime cuando la respuesta correcta tiene que ser la que *no* sea lógica. Los sujetos nativos reflejan esta confusión: A se lleva 5 votos, B 2, y C 9.

En cuanto a las preguntas de valor Rit negativo (correlación ítem-test), la 28, 29 y 32 dan lugar a las siguientes reacciones entre los sujetos españoles:

28: todos los 16 eligen A, la respuesta correcta según la llave
29: A 2, B 13, C 1
32: los 16 optan por D, la respuesta correcta según la llave

En lo que se refiere a las preguntas 28 y 32, con esta validación *a posteriori* a base de los datos producidos por los candidatos holandeses, puede que decidamos excluirlas, y quizás sin argumentos convincentes. El hecho de que respondan de forma unánime todos los nativos indica que las preguntas quizás sean difíciles pero no defectuosas. La 29, que también analizamos antes, presenta más dudas teniendo en cuenta las respuestas de los españoles.
Un caso especial es la pregunta 11:

11. @ Mira, Julita, ése del bigote es Pepe. ———————— y estudia mucho.

A. Es muy bien
B. Es muy estupendo
C. Es muy largo
D. Es muy serio

ITEM	ALTERNATIVAS	CANDS	PORCENTAJES
11	A	2	0.027
11	B	1	0.013
11	C	2	0.0,27
11	D	70	0.933

Todos los sujetos españoles marcaron la alternativa D, porque suponían que la C, aunque admisible, probablemente no se correspondería con el nivel de conocimientos de lengua hablada y coloquial de los estudiantes holandeses: en esta reserva se detecta fácilmente el papel dictatorial del

libro de texto. El uso permite desvíos de las reglas pedagógicas. Otro ejemplo de esto mismo se observa en las diferencias entre las reglas de uso de pretéritos imperfectos e indefinidos, en su formulación para extranjeros y en el grado de anarquía aparente que impera en el uso de hablantes nativos. Pero un uso estricto, restringido y quizás chirriante no es poco mérito en boca de extranjeros.

14. Validación a base de preguntas abiertas

Los alumnos de nuevos cursos no se encontrarán con los ejemplos que utilizamos para explicar el procedimiento seguido en los últimos años para mejorar los ítemes defectuosos; se usaron los primeros datos como pretesting, se sacaron tests paralelos de igual peso a base de los coeficientes de dificultad y de la varianza. En el resto de este artículo queremos prestar una cierta atención a un tipo de evaluación de ítemes mucho más transparente que los análisis estadísticos utilizados hasta ahora. Farhady (1983) explica los pasos que dio hasta llegar a un test de competencia comunicativa de inglés como segunda lengua, revisando y limando constantemente las preguntas que había redactado. Nosotros, en múltiples ocasiones, invitamos a hablantes nativos a completar las preguntas de elección múltiple como si fueran preguntas abiertas. Para aclarar esto reproducimos aquí las preguntas que acabamos de analizar. Deberíamos haber empezado por esta vía... Vea el lector, sin embargo, qué contestaría en las lagunas, y compare sus respuestas después con las alternativas que sugerimos antes. Para que el test no insista demasiado en aspectos gramaticales tendremos que afinar mucho más en sofisticación pragmática.

24.	+	Oye, ¿puedes tú escribirle una carta a Julio Romero de Torres?
	o	¿Quién, yo? ————————————
28.	+	¿Qué te pasa, Fernandito?
	o	Nada, mamá, sólo que Maribel ha dicho que no quiere jugar conmigo ya porque ———————————— Pero ya encontraré otra amiga para jugar con el Lego.
29.	+	¿Por qué llevas unos zapatos tan feos?
	o	————————————
1.	+	¿Tú crees que este examen va a ser más fácil que el anterior?
	o	————————————
9.	+	Claudia, éste es Juanjo, un amigo mío. Juanjo, te presento a Claudia, mi mejor amiga.
	o	————————————
11.	@	Mira, Julita, ése del bigote es Pepe ———————————— y estudia mucho.

16. + ¿Conoces ya a muchos españoles? ¿Qué opinas de nosotros? Y concretamente, ¿qué piensas de mí?
o ———————————————

26. + Hola, buenas noches, perdone que le moleste, soy el vecino de abajo. ¿Podría subir un poco el volumen de la tele? Es que se ha estropeado el sonido de la tele. La imagen está bien, o sea que si vuelvo a bajar yo y si usted es tan amable...
o Claro, hombre, ———————————————

32. + ¿Qué haces tú todavía aquí en la cantina?
o Es que tengo otra clase a las tres y ———————————————

Las veces que se han solicitado este tipo de respuestas a hablantes nativos hemos podido comprobar que sólo en muy pocos casos coincidían con nuestras alternativas, señal de que en la mayoría de las preguntas el alumno, cuando acierta, no selecciona la respuesta correcta, sino la más plausible, o la menos inadecuada.

15. Conclusiones

A lo largo de este artículo nos hemos referido a nuestro *test* como si fuera algo más que una prueba de nivel. Mientras no se resuelva la discusión sobre las distintas competencias o la competencia única es arriesgado afirmar que se trate de una cosa u otra. La cuestión es que discrimina bien este conjunto de ítemes. Es tentador pensar que *mide algo relativamente bien operacionalizado*, pero queda mucho por hacer. En cuanto a la confección del test que analizamos, tenemos que concluir que además de llevar a cabo los controles habituales centrados en cuestiones de fiabilidad (pretesting, análisis de resultados provisionales y correcciones de todo tipo), es imprescindible analizar la validez de los ítemes en términos de varianza. La pauta la marca el grado de divergencia entre los datos obtenidos con informantes del público meta y con nativos. Hay que tener en cuenta que los hablantes nativos no constituyen un bloque homogéneo con criterios inalterables. A falta de criterios de validez, la discusión se reduce a consideraciones de fiabilidad. Hemos visto que también son engañosos los coeficientes de fiabilidad a los que llegamos con sólo unos ligeros retoques. A pesar de todas estas posibles objeciones, poder contar con sujetos ajenos al grupo de profesores directamente adscritos al centro que imparte la enseñanza es una incomparable ventaja a la hora de ultimar tests que pueden tener graves consecuencias para los alumnos.

BIBLIOGRAFÍA

BAKER, D., 1989, *Language testing. A critical survey and practical guide*, London, Edward Arnold.

CANALE, M. & M. Swain, 1980, "Theoretical bases of communicative approaches to second languages teaching and testing", *Applied Linguistics*, 1,1, pp. 1-47.

CANALE, M., 1983, "From communicative competence to communicative language pedagogy". En C. Richards & Richard W. Schmidt (eds.). *Language and communication*, Burnt Mill, Harlow Essex, Longman, pp. 2-27.

CRONBACH, L. J., 1971, "Test validation". En R.L.Thorndike, (ed.). *Educational measurement* (2nd ed.), Washington DC, American Council on Education, pp. 443-507.

EK, J. A. van, 1976, *The Threshold Level*, Estrasburgo, Consejo de Europa.

MARTÍN PERIS, E., L. Miquel López, N. Sans Baulenas y M. Topolevsky Bleger, 1987, *Para empezar. Curso comunicativo de español*, Madrid, Edi 6.

MARTÍN PERIS, E., L. Miquel López, N. Sans Baulenas y M. Topolevsky Bleger, 198?, *Esto funciona. Curso comunicativo de español*, Madrid, Edi 6.

FARHADY, H., 1983, "New directions for ESL proficiency testing". En John W. Oller (ed.), pp. 253-269.

HATCH, E., & H. Farhady, 1982, *Research Design and Statistics for Applied Linguistics*, Rowley, Mass., Newbury House.

HUGHES, A., 1989, *Testing for language teachers*, Cambridge, Cambridge University Press.

HYMES, E., 1972, "On communicative competence". En J. B. Pride & J. Holmes (eds.). *Sociolinguistics*, Harmondsworth, Penguin, pp. 269-293.

MESSICK, S., 1981, "Evidence and ethics in the evaluation of tests", *Educational Researcher*, 10, pp. 9-20.

MESSICK, S., 1988, *Meaning and values in test validation; the science and ethics of assessment*, Paper presented at the AERA Convention, New Orleans.

OLLER JR., J. W. (ed.), 1983, *Issues in Language Testing Research*, Rowley, Mass., Newbury House.

RODRÍGUEZ DIÉGUEZ, J. L., 1980, *Didáctica general. 1. Objetivos y evaluación*, Madrid, Editorial Cincel.

SHOHAMY, E. y O. Inbar, 1991, "Validation of listening comprehension tests: the effect of text and question type", *Language Testing*, 8,1, pp. 23-40.

SLAGTER, P. J., 1979, *Un nivel umbral*, Estrasburgo, Consejo de Europa.

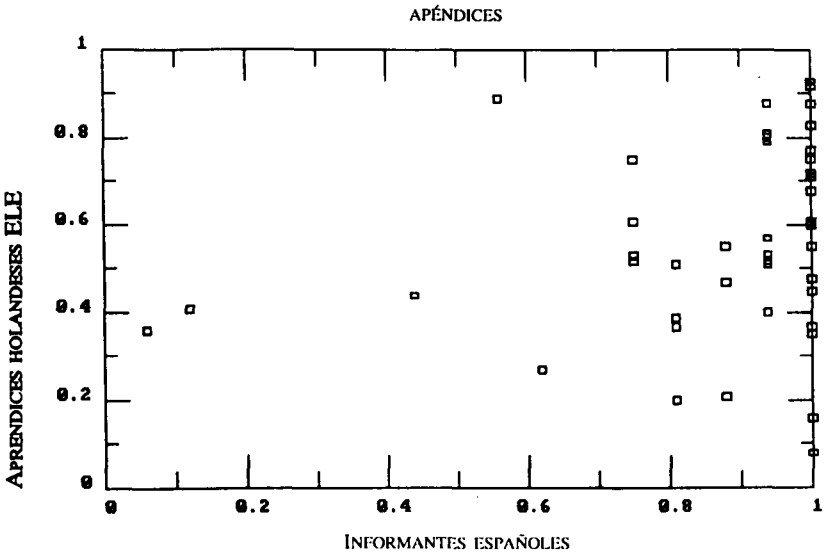

GRÁFICO 1

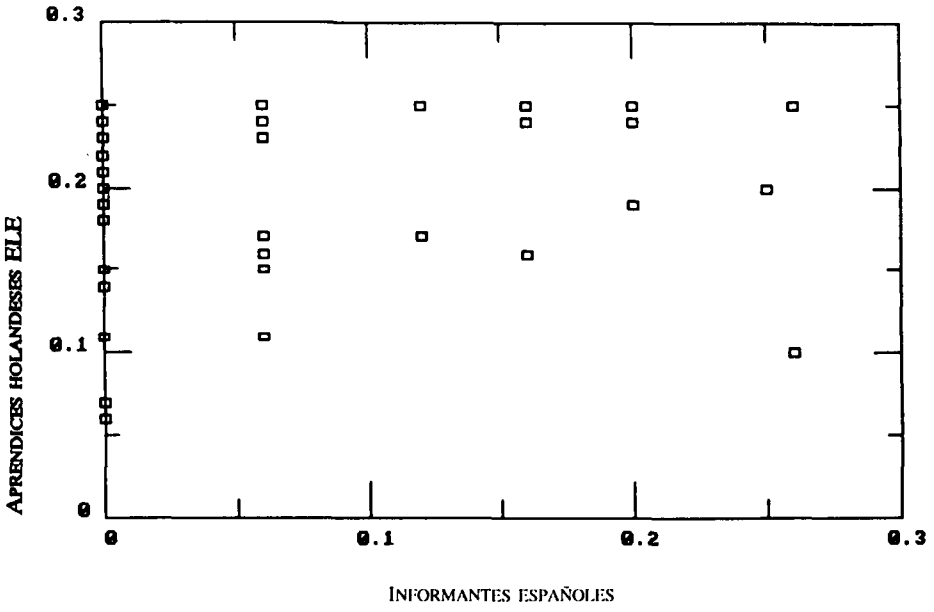

GRÁFICO 2