Modelos de evaluación en tests objetivos

Honesto Herrera Soler
Facultad de Ciencias Económicas y Empresariales de Madrid

RESUMEN: En este trabajo estudio el "test de elección múltiple" y "el c-test"como herramientas idóneas para la configuración de tests a gran escala y, por supuesto, para trabajar en clase. Estos modelos de evaluación tienen también sus limitaciones al basarse en la identificación de la opción correcta en un caso y en el otro en la recuperación de los elementos mutilados. Son referentes básicos indirectamente el "c-test" y directamente el test de elección múltiple en "The Programme for International Student Assessment" (PISA: 23), en el "U.S. Department of Education and the Council of Chief State School (1998)" y "The Common European Framework of Reference (2005)". Ambos pueden ser alternativa o complemento de los distintos ítems objetivos y suministran información muy útil para el profesor más allá de los aciertos. La validez, fiabilidad y economía de estos modelos de evaluación en su aplicación y en su corrección justifican su uso en el aula.

PALABRAS CLAVE: Test de elección múltiple, C-Test, formato, inicio mutilación, C-test guiado/ no guiado, fractiles, errores.

1. INTRODUCCIÓN

En este trabajo la manzana de Ortega y Gasset, que puede ser verde, rosa o amarilla según la perspectiva desde donde la miremos, se hace actualidad. El perspectivismo de Ortega y Gasset lo podemos proyectar al tema de la evaluación y, más en concreto, a cualquier examen con el que pretendamos saber el nivel de nuestros alumnos en un momento determinado. El examen tendrá un color distinto según la perspectiva desde la que se considere. El diseñador lo verá de distinta manera que lo pueda ver el evaluador y el estudiante de distinta manera que el diseñador y el calificador. En cualquiera de los casos al final un resultado, unos números, en los que cada uno de los actores tiene su parte de responsabilidad.

Ante un examen los tres protagonistas tienen sus preocupaciones y sus dudas aunque de distinto orden. Las preocupaciones del diseñador se centran en la validez del contenido y en su valor predictivo mientras que las del evaluador en evitar sesgos personales y conseguir que sus puntuaciones sean fiables y lo más objetivas posibles. Todos sabemos que las dudas y preocupaciones del estudiante vendrán generadas por el nivel de dominio que tenga de la materia y el grado de dificultad que tengan las cuestiones

a las que tiene que responder, en definitiva por la nota que pueda sacar. El alumno es el destinatario del examen y su repercusión en la autoestima y en los efectos que puede tener en su curriculum justifican todo tipo de miedos, recelos y nervios.

Hoy quiero centrarme principalmente en la figura del diseñador y sólo, en cierta medida, en la figura del evaluador y en la del estudiante de bachillerato. El diseñador cuando tiene la tarea de configurar un test tiene en cuenta:

1. el modelo del examen que elige,
2. el material que va a usar,
3. la una estructura que le quiere dar al examen,
4. la cantidad de información que pretende obtener del estudiante,
5. los aspectos concretos que quiere evaluar, etc. etc.

En todo este proceso, los interrogantes se suceden y se multiplican en su mente:

1. ¿ Es un diseño equilibrado tanto desde el punto de vista técnico como desde el contenido? i.e. ¿ocupa un punto medio en un continuo que va de lo muy fácil a lo muy difícil o, dicho en términos estadísticos, nos permite esperar una distribución normal de los resultados?
2. ¿Es el examen una muestra representativa de todo aquello que se pretende evaluar?
3. ¿Hay riesgo de sesgo de que factores culturales o sociales incidan en un determinado grupo de alumnos? Pensemos, por ejemplo, en la redacción. Es muy distinto pedirles que hagan una redacción sobre la natación que sobre la natación sincronizada. Si se opta por la natación sincronizada quienes conozcan el tema tendrán ventaja sobre aquellas personas que no sepan nada de este tipo de natación.
4. ¿Qué modelo de test puede ayudar a simplificar el proceso de evaluación en un momento determinado?

Este último interrogante me lleva a plantear dos modelos de evaluación de carácter objetivo: el test de elección múltiple y el C-test por su economía en el diseño, en su aplicación y en su corrección. Ambos nos permiten:

1. distanciarnos de perspectivas subjetivas y centrarnos en datos, que no tienen el filtro del subjetivismo,
2. generalizar los resultados y establecer contrastes y correlaciones con otros modelos,
3. trabajar con tests fiables y válidos. Dos rasgos básicos en cualquier tipo

test. Si en los años 70 la característica más importante de un test era la fiabilidad hoy en día la validez se considera tan importante como la validez. Autores como Hughes (1989) consideran que los factores de fiabilidad y validez son complementarios y otros como Weir (2005) incluso piensan que la fiabilidad es una forma de validez en el test.

4. incorporarlos a las nuevas tecnologías: internet y WebCT.

Asumo que ambos modelos no son la panacea de la evaluación y que tienen sus limitaciones, pero, aparte de su simplicidad en el diseño, aportan una información muy significativa, que no se suele analizar, y que es de gran relevancia. Se recurre a ellos para medir el proceso de aprendizaje, para calificar y clasificar a los alumnos, pero sólo se presta atención a los aciertos y se olvida el efecto que puede tener el diseño o el valor de las respuestas incorrectas y de las omisiones. Razones lo suficientemente importantes para que proponga una relectura y una reflexión sobre lo que aportan ambos modelos.

2. C-TEST

Comenzaré por este último modelo presentando dos variantes. Los objetivos van dirigidos, por una parte, a observar el efecto del diseño en el nivel de respuesta y, por otra, a analizar si el nivel de correlación con otros tests convencionales lo justifican como alternativa o complemento.

2.1. El C-test como respuesta al "cloze"

En 1981 Klein-Braley y Raatz diseñaron el C-Test, como respuesta al "cloze", o tests de cierre. El cloze test fue una idea de Taylor en 1953, tomando como referente la psicología de la Gestalt. El C-Test es un cloze modificado y reducido por lo que para entender lo que supone de innovación en los tests de cierre procede que se haga un poco de historia del cloze.

Taylor entendió que era una prueba económica para evaluar el nivel de competencia lingüística en general y el nivel de la expresión escrita en particular. Como ya sabemos el diseño de este test se basa en un texto considerablemente largo en el que se suprimen una de cada 7 palabras si el diseño es sistemático, aunque si el diseño es racional la palabra eliminada puede estar en intervalos entre la 5ª y la 12ª . El alumno tiene que recuperar el término eliminado. El cloze test se basa en la teoría de la "redundancia", característica del lenguaje natural, que nos ayuda a descodificar un mensaje mutilado escrito u oral en virtud del conocimiento de las reglas y patrones que tenemos de nuestra propia lengua y cultura. La eliminación de elementos redundantes exige que el lector u oyente tenga que hacer inferencias para la comprensión del texto. Su aplicación ha supuesto una serie de limitaciones que cuestionan su finalidad y que paso a enumerar:

1. en su diseño puede haber varias respuestas posibles para un mismo hueco lo que hace poco fiable su corrección,
2. en una interpretación estricta el diseñador no controla lo que se está evaluando, ya que una vez elegida la primera palabra el resto viene condicionado por ese inicio,
3. se necesita un texto extremadamente largo para obtener un número de huecos que garantice la fiabilidad del test,
4. hablantes nativos y cultos pueden obtener peores resultados que hablantes no nativos, por lo que se pone en duda la validez del test.

2.2. C-Test como alternativa al cloze

Estas y otras limitaciones motivaron que se propusiera en los 80 al C-Test como alternativa. Es un modelo que se basa también en la Psicología de la Gestalt y que supera las limitaciones mencionadas anteriormente. Desde que hizo su aparición en el ámbito de la evaluación ha sido motivo de estudio y debate en los foros de investigación. De los estudios hechos sobre este modelo tan sólo quiero mencionar tres tendencias y algunos autores que han trabajado en ellas:

1. el análisis de las variables concretas que determinan la facilidad o dificultad de los ítems: Esteban 2007 y 2005; Esteban y Herrera 2003,
2. la competencia lingüística global: Eckes y Grotjahn 2006; Rashid 2002; Babaii y Ansary 2001,
3. y especificaciones y puntualizaciones en el diseño: Babaii y Moghaddam (2006) y Jafarpur (1999).

2.3. Diseño del C-Test

Por su economía en el diseño, administración y valoración es un modelo muy apropiado para aplicarlo en clase. Su diseño es muy simple. En cualquier texto con el que se quiera trabajar se deja intacta la primera oración y a partir de ahí se aplica "la regla del dos" suprimiendo la mitad de cada segunda palabra a partir de la segunda palabra de la segunda oración. Veamos el siguiente ejemplo:

Los medios audiovisuales están presentes en nuestra vida cotidiana. En es___ ponencia _____ propondrán diver____ actividades _____ pueden reali_____ con anun_____, videoclips y pel _____ en _____ clase _____ lenguas extran_____.

La recuperación de este texto requiere que el alumno trabaje no sólo el léxico sino también algunos aspectos morfo-sintácticos tales como el género, el número y el aspecto verbal.

Los parámetros, por tanto, que se requiere aplicar cuando se aplica este modelo de evaluación son los siguientes:

1. aplicar la regla del 2,
2. los monosílabos se puede optar por mutilarlos o eliminarlos. Su eliminación significa que hay que recuperar pronombres, preposiciones, conjunciones u otros términos,
3. el proceso de eliminación continúa hasta obtener 20 ó 25 huecos en cada texto, ya que el C-test consta de 100 huecos en total,
4. se requiere, por tanto, 4 o 5 textos cortos para potenciar la diversidad temática,
5. los textos se ordenan de menor a mayor dificultad para mejorar la distribución de las puntuaciones,
6. los tests se corrigen dando un punto a cada palabra que se recupera correctamente, la máxima puntuación es 100,
7. las palabras de una sola letra, los números y los nombres propios no se tienen en cuenta en el diseño.

3. LA PRIMERA VARIANTE DEL MODELO DEL C-TEST: INICIO DE LA MUTILACIÓN EN LA SEGUNDA O EN LA TERCERA PALABRA.

En esta primera variante del C-test se observa, por una parte, el efecto en los resultados según se inicie la mutilación en la segunda o la tercera palabra y, por otra, su correlación con otros modelos convencionales de tests. Si se aplica este modelo en clase y se va más allá de los resultados se obtendrá una información valiosísima sobre los aspectos semánticos y morfo-sintácticos que nuestros alumnos dominan y no dominan, sobre la facilidad o dificultad del tema, sobre las ventajas o desventajas del formato, etc. El conocimiento de todos estos detalles permitirá actuar sobre ellos, contribuirá a la mejora y calidad de nuestra docencia y nos convertirá en auténticos investigadores en el aula. Simplemente, desde nuestra clase, desde nuestra aula, se habrá dado respuesta a no pocos interrogantes sobre el rendimiento de nuestros alumnos.

3.1. Hipótesis

Queremos constatar si hay diferencias significativas según el inicio de la mutilación de palabras y analizar su nivel de correlación con los tests o subtests que se aplican en la Escuela Oficial de Idiomas en la asignatura de inglés.

3.2. Metodología

3.2.1. Alumnos

Se trabaja con 151 alumnos matriculados en 3º curso de la EOI Jesús Maestro de Madrid. Pertenecen a 8 clases elegidas al azar entre todas las existentes. Su distribución es aleatoria. Se forman dos grupos el uno responde el C-Test A y el otro el C-Test B.

3.2.2. Material

Se diseñan dos C-tests (C-test A y C-test B) extraídos de 4 textos con 25 palabras mutiladas en cada uno. La única diferencia estriba en el planteamiento del inicio de la mutilación. El C-Test A comenzaba en la segunda palabra y el C-Test B en la tercera palabra (Anexo 1). Los cuatro textos se ordenaron teóricamente de menor a mayor dificultad. Se les aplica también el test "tutor", extraído de la red, para constatar la homogeneidad de los grupos (Anexo 2).

3.2.3. Proceso

Las dos formas del C-Test y la batería de los tests de la EOI se administraron con una semana de diferencia.

3.3. Resultados

Lo primero que constatamos es la homogeneidad. No hay diferencias significativas entre los dos grupos en el "test tutor. Estos resultados justifican la aleatoriedad en la distribución de los grupos y permiten seguir adelante con el análisis contrastivo. El análisis de los datos de esta variante del C-Test lo reduciré tan sólo a la presentación de dos tablas que responden a las hipótesis que se han planteado sobre el formato y su correlación con otros tests. Con la otra variante del C-Test que se estudiará a continuación completaré algunos aspectos de este modelo. El efecto del inicio de la mutilación se observa en la siguiente tabla I.

Una lectura elemental del estadístico "t" en esta tabla permite observar que hay diferencias significativas entre los subtests con valores que van de (p 0.011 a p .000), que es lo que se esperaba y no las hay cuando se comparan los resultados

Totales del C-Test A y del C-Test B, i.e. no importa si se comienza en la segunda o en la tercera palabra (p 0.366), resultado que se considera un gran hallazgo. El inicio de la mutilación, por tanto, no es relevante, no incide en los resultados de los alumnos. (Ver tabla en la página siguiente)

Tabla I. Diferencias en los subtests y test total

		Prueba T para la igualdad de medias						
		t	gl	Sig. Bilateral	Diferencia de medias	Error típ. de la diferencia	95% Intervalo de confianza para la diferencia	
							Inferior	Superior
Ttph	Se han asumido varianzas iguales	6,743	149	,000	4,18375	,62042	2,95780	5,40970
	No se han asumido varianzas iguales	6,695	128,642	,000	4,18375	,62491	2,94732	5,42017
TtRe	Se han asumido varianzas iguales	-3,050	149	,003	-2,29589	,75276	-3,78335	-,80844
	No se han asumido varianzas iguales	-3,067	140,883	,003	-2,29589	,74857	-3,77579	-,81600
TtEn	Se han asumido varianzas iguales	-2,561	149	,011	-1,76957	,69086	-3,13471	-,40442
	No se han asumido varianzas iguales	-2,571	146,044	,011	-1,76957	,68835	-3,12999	-,40915
TtVi	Se han asumido varianzas iguales	-3,230	149	,002	-2,25939	,69947	-3,64156	-,87722
	No se han asumido varianzas iguales	-3,227	147,886	,002	-2,25939	,70013	-3,64294	-,87584
CTest	Se han asumido varianzas iguales	-,906	149	**,366**	-2,14110	2,36276	-6,80995	2,52775
	No se han asumido varianzas iguales	-,907	148,988	,366	-2,14110	2,36046	-6,80540	2,52320

Tabla II. Correlaciones C-test A

		CTest.A	Reading	Listening	Writing	EOI
CTest.A	Correlación de Pearson	1	,581(**)	,547(**)	,369(**)	,644(**)
	Sig. (bilateral)		,000	,000	,001	,000
	N	77	77	77	77	77
Reading	Correlación de Pearson	,**581**(**)	1	,377(**)	,520(**)	,795(**)
	Sig. (bilateral)	,000		,001	,000	,000
	N	77	77	77	77	77
Listening	Correlación de Pearson	**,547**(**)	,377(**)	1	,408(**)	,789(**)
	Sig. (bilateral)	,000	,001		,000	,000
	N	77	77	77	77	77
Writing	Correlación de Pearson	,**369**(**)	,520(**)	,408(**)	1	,779(**)
	Sig. (bilateral)	,001	,000	,000		,000
	N	77	77	77	77	77
EOI	Correlación de Pearson	,**644**(**)	,795(**)	,789(**)	,779(**)	1
	Sig. (bilateral)	,000	,000	,000	,000	
	N	77	77	77	77	77

** La correlación es significativa al nivel 0,01 (bilateral)

La segunda hipótesis era constatar si las puntuaciones de los alumnos tenían algún tipo de correlación con los tests convencionales de la E.O.I. (Tabla II).Las correlaciones son semejantes en los dos grupos, por razones de espacio se presenta las del grupo A. En la tabla (II) se observa una correlación significativa al nivel 0.01 del C-test con las distintas áreas de evaluación de la E.O.I. Los niveles son más altos en "Reading" (0.581) y en "Listening" (0.547) y más bajos en "Writing" (0.369). Los valores de la correlación con "Reading" podrían justificarse en base a que el C-test podría encuadrase en esta área de evaluación. La correlación con "Listening" podría también deberse al hecho de que lo mismo que el "Reading" agrupan subtests que se etiquetan como tests dirigidos en los que las respuestas del alumno están condicionadas e incluso en ocasiones son cerradas. Esta interpretación justificaría los bajos valores de correlación con el área de "Writing" porque en esta destreza el alumno estructura su propia respuesta, es un subtest abierto y en su valoración influyen factores como la organización, originalidad y nivel comunicativo del alumno así como los criterios personales del profesor evaluador. La correlación 0.644 del test en su conjunto de la E.O.I, superior a los valores de los subtests subraya la validez del modelo.

4. EFECTO DE LA DENSIDAD, DE LA VARIACIÓN Y DEL FORMATO EN EL C-TEST

En esta segunda variante se estudia la influencia que puede tener la densidad, la variación y el formato en los textos y su correlación, en este caso, con los estudiantes de bachillerato. Se hace un seguimiento de los factores que influyen en el grado de dificultad del C-test. Si los ítems son guiados o no, si los temas son conocidos o no, si corresponden al nivel de los alumnos, si los términos son abstractos o concretos o factores como la longitud, familiaridad, redundancia de los ítems.

4.1. Metodología

4.1.1. Participantes

En ésta ocasión no son alumnos de la Escuela Oficial de Idiomas sino alumnos de bachillerato. Este C-test se aplica a 162 alumnos de 2º de Bachillerato de cuatro Institutos de Enseñanza Secundaria de la Comunidad de Madrid.

4.1.2. Material

Se procuró partir de textos que aseguraran a priori algunas cuestiones, como la homogeneidad de nivel, la autenticidad y el interés en cuanto al tema. Por ello, en la selección de textos se optó por cuatro fragmentos procedentes de exámenes de Inglés de las pruebas de acceso a la universidad de otros años.

4.1.3. Proceso

Todos los sujetos realizaron un C-test construido a partir de cuatro textos distintos con 25 omisiones cada uno. La mitad hizo el modelo A y la otra mitad el modelo B y todos hicieron el test de acceso a la universidad "Cavemen?" para valorar la homogeneidad de los grupos. Se tomaron además las calificaciones de Inglés en la 2ª evaluación y las logradas en la Selectividad oficial de junio de ese año para estudiar el nivel de correlación.

4.1.4. Formato

Tabla III. Estructura y modelos del C-test aplicado

C-TEST A	**Textos y diseño**	**C-TEST B**	**Textos y diseño**
C-TEST 1 Ítems 1-25	Road accidents - - - - - -	C-TEST 1 Ítems 1-25	American imperialism - - - - - -
C-TEST 2 Ítems 26-50	Evolution - - - - - -	C-TEST 2 Ítems 26-50	Women doctors - - - - - -
C-TEST 3 Ítems 51-75	American imperialism _____	C-TEST 3 Ítems 51-75	Road accidents _____
C-TEST 4 Ítems 76-100	Women doctors _____	C-TEST 4 Ítems 76-100	Evolution _____

En los dos primeros subtests del modelo A y del modelo B figuraban los espacios correspondientes a las omisiones de cada ítem. Esta ayuda se eliminó en los dos últimos subtests de ambos modelos (Tabla III). Se distribuyeron alternativamente según estaban sentados en el aula.

4.2. Resultados y discusión

En esta ocasión comenzaré presentando las correlaciones de los resultados del C-test con los otros tests para su validación: Cavemen?, PAU de junio de 2001 y calificaciones en Inglés en la 2ª evaluación (tabla IV). Los valores del C-test con todos los demás: 0.723, 0.722 y 0.750 avalan la validez del C-test. La fiabilidad se comprobó con los estadísticos del "split-half" y del "α de Cronbach" con los siguientes valores: 0.890 y .892 que muestran una consistencia interna próxima a los límites óptimos. La primera hipótesis de estudio que se plantea en este modelo es la incidencia del tema y del formato, para ello se analiza en primer lugar el grado de dificultad de los sub-tests. Los valores de los promedios y las desviaciones típicas que se observan en la tabla (V) entrarían dentro de valores normales en un test de carácter objetivo.

Tabla IV. Correlaciones de muestras relacionadas

		1	2	3	4
1. C-test		**--**			
	N	162			
2. 2ª Evaluación		**,723(**)**	--		
	N	161	161		
3. PAAU		**,722(**)**	,575(**)	--	
	N	81	80	81	
4. *Cavemen?*		**,750(**)**	,805(**)	,654(**)	--
	N	162	161	81	162

** La correlación es significativa al nivel 0,01 (bilateral).

Tabla V. Promedios de los sub-tests y del test total

	Media	Desviación típ.
CTEST1	13,26	5,070
CTEST2	15,64	3,818
CTEST3	10,15	5,132
CTEST4	12,07	4,538
CTESTTOTAL	51,12	14,683

* (Los valores correspondientes a los supra-ítems son la media alcanzada en una escala del 0 al 25)
* (Los valores correspondientes a CTESTTOTAL están expresados en una escala del 0 al 100)

En los sub-tests 3 y 4, que introducen el formato no guiado, se obtiene un promedio más bajo (10,15 y 12.07), que en los sub-tests 1 y 2. Estos datos nos indican el aumento del grado de dificultad según los ítems sean guiados o no guiados. El C-test se creó a partir de cuatro textos que se consideraron de un grado de dificultad semejante. Todos coinciden en su carácter divulgativo o periodístico y tratan temas de actualidad e interés general. Si se comparan los resultados de los alumnos en cada uno de los sub-tests según el formato se observan diferencias significativas en los promedios en todos ellos (tabla VI):

Tabla VI. Comparación de medias obtenidas para cada texto según el formato aplicado

Texto base	Formato	Media	N
Road accidents	- - - - -	16,15	81
	_____	12,84	81
Evolution	- - - - -	16,12	81
	_____	12,05	81
American imperialism	- - - - -	10,37	81
	_____	7,47	81
Women doctors	- - - - -	15,15	81
	_____	12,10	81

Esta tabla refleja que las mayores dificultades aparecieron en el texto American imperialism, con eliminaciones no guiadas (modelo A, subtest 3, media = 7,47). Sin embargo, cabe destacar que cuando son guiadas, aunque el promedio mejora sensiblemente (media = 10,37, modelo B subtest 1) sigue siendo el más bajo de los obtenidos en los subtests guiados, probablemente porque se trata de un registro menos común para los alumnos. Se trata de un tema que es frecuente en los medios de comunicación, pero pertenece a la esfera del pensamiento político-social y tiene mayor grado de abstracción que los otros. Estos datos nos llevan a concluir que cuando las eliminaciones no son guiadas baja el número de aciertos y aumentan los valores perdidos, es decir, que un buen número de sujetos ni siquiera intenta resolver los ítems no guiados. Estos resultados subrayan que el formato guiado tiene un efecto motivador que se debe tener presente en el diseño.

El comportamiento de los alumnos ante un ítem cualquiera se puede observar en lass siguientes tablas (VIIa y VIIb). Se tratar de recuperar el término “hunger”

Tabla VIIa.

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Incorrecto	46	56,8	65,7	65,7
	Correcto	24	29,6	34,3	100,0
	Total	70	86,4	100,0	
Perdidos	sin hacer	11	13,6		
Total		81	100,0		

Tabla VIIb.

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Incorrecto	32	39,5	68,1	68,1
	Correcto	15	18,5	31,9	100,0
	Total	47	58,0	100,0	
Perdidos	sin hacer	34	42,0		
Total		81	100,0		

Las diferencias entre las respuestas correctas, incorrectas y los que no lo intentan nos lleva a pensar que el formato tiene incidencia en los resultados. Los datos de la variación y densidad léxicas son muy semejantes en todos los textos del C-Test, aunque en un continuo de facilidad / dificultad Evolution (variación 63,46 y densidad léxica 45,19) y American imperialism (variación 70,37 y densidad léxica 60,18) aparecerían en uno y otro extremo respectivamente. Este último texto también presenta valores más extremos en el carácter de los términos afectados por la mutilación, ya que los términos léxicos afectados suponen el 76% del texto frente al 24% de los términos de función como se observa en la siguiente tabla:

Tabla VIII. El texto: densidad, variación y mutilación

Textos	Variación léxica	Densidad léxica	Términos léxicos afectados	Términos de función afectados
Road Accidents	69.47	59.95	13 52%	12 48%
Evolution	63.46	45.19	10 40%	15 60%
American Imperialism	**70.37**	**60.18**	**19 76%**	**6 24%**
Women Doctors	69.14	58.51	14 56%	11% 44%

Estos factores pueden ser la explicación de que los promedios obtenidos en el subtest de American Imperialism: 10,37 y 7,47, según el formato sea guiado o no, sean tan bajos y tan diferentes. La densidad y la variación afectan a los resultados (Klein-Braley 1985: 91, Dörnyei y Katona 1992: 197, Farhady y Keramati 1996).

Todo ello ratifica la importancia de la selección de los textos y formato en la creación de cualquiera de las variantes del C-tests.

5. EL MODELO DE LA ELECCIÓN MÚLTIPLE DESDE UN ANÁLISIS DE LOS FRACTILES

El test de elección múltiple es un modelo de evaluación que se puede presentar en varios formatos en función de las opciones. Es un test denostado, pero que nuestros alumnos tienen que conocer porque es el referente principal en los tests a gran escala como sucede en los tests en el marco de convergencia europea (Weir,2005, Alderson and Huhta 2005). Es un modelo, lo mismo que el C-Test, mide el dominio que el alumno tiene de lo que se examina por la vía del reconocimiento, de la inferencia y del razonamiento.

Dejaré para otra ocasión la calibración de los ítems: la teoría clásica de los tests (TCT), o la teoría de la respuesta al ítem (TRI) (Zimowski et al. 2003) y me centraré en el análisis de los aciertos y de los errores en las respuestas a través de los fractiles. Esta aproximación nos permitirá no sólo calificar a los alumnos en función de su aciertos sino también obtener información del conocimiento parcial del alumno a través de sus errores (Herrera-Soler 2005).

5. 1. Metodología

5.1.1. Participantes

712 candidatos a la administración pública.

5.1.2 . Material

El test Albion de la editorial EOS que consta de 57 ítems distribuidos en cuatro subtests. En cada uno de ellos se subraya el aspecto semántico, el morfo-sintáctico en ítems individuales o en contexto y el interactivo-comunicativo como se observa en cada uno de los siguientes ejemplos extraídos de cada subtest y que he adaptado al castellano:

Subtest 1. Se le pide al candidato que señale el enunciado que considere equivalente al estímulo.

No llegaremos a tiempo

A. Es muy difícil llegar a tiempo
B. Es casi imposible que lleguemos en el tiempo estipulado
C. Nunca llegaremos
D. No lo conseguiremos

Sub-test 2. Se le pide que elija la opción más apropiada para completar el texto.

No llegues tarde a la reunión, es _____ las ocho p.m.

E. a
F. en
G. dentro de
H. sobre

Sub-test 3. Se le exige lo mismo que en el anterior. La diferencia radica en que no son ítems aislados sino en un contexto. En este caso se trata de una carta comercial.

....Si pudiera enviarnos un catálogo especificando los programas y los _____
le estaríamos muy agradecidos. Pensamos (hacer) un pedido de 25 unidades ....

I. premios
J. indicaciones
K. precios
L. formularios

Sub-test 4. Se le pide que elija la mejor respuesta a la pregunta que se le hace

¿Buenas tardes, Román, ha terminado el informe ya?

M. Sí, lo recibí ayer
N. Sí. lo terminé hace una hora
O. Sí, lo terminaré mañana
P. Sí, mi jefe se olvidó dármelo

5.1. 3. Proceso

Se hace un análisis de la respuesta de los candidatos por medio de fractiles.

Herramientas

Los datos se pueden estudiar con una simple calculadora o con el SPSS. Por supuesto, si es posible, el ordenador nos ahorra mucho trabajo.

5.2. Fractiles

¿Qué entendemos por fractiles?. Es la división en grupos, en este caso en cuatro grupos, en función del nivel de aciertos conseguidos, de los candidatos que hacen el test, i.e. según que su rendimiento sea bajo, aceptable, bueno o excelente en el conjunto de los ítems. Esta división nos permite ver si el ítem es idóneo para todos los candidatos o sólo para un grupo determinado. Éste modelo es sencillo de aplicar y suministra información no sólo de la distribución de las respuestas correctas sino también de la distribución de las respuestas incorrectas, que permiten hacer inferencias sobre el conocimiento parcial que tiene el candidato de algunas de las opciones disuasorias que le presentamos.

El análisis de los fractiles tiene como referente el paradigma de que en cada ítem los aciertos alcancen el 40 – 60 % y las opciones incorrectas estén en torno al 10%. Esta distribución que sería la ideal difícilmente se cumple por alguna de las siguientes razones:

1. inclusión de ítems de motivación que pueden responder prácticamente todos los alumnos,
2. dificultad del ítem en sí,
3. deficiente enunciado del ítem,
4. y, sobre todo, la dificultad de ofrecer opciones incorrectas que tengan el mismo grado de verosimilitud.

Esta categorización nos permite valorar cada uno de los ítems en función de las opciones que han elegido los alumnos y definirlas como:

1. Categoría dominante. Los candidatos pueden ignorar prácticamente todas las opciones incorrectas y elegir sólo la correcta o responder dentro de los porcentajes indicados en una, en dos o en las tres opciones incorrectas.
2. Categoría de verdadero / falso simple cuando los candidatos no consideran dos de las opciones incorrectas porque saben que son incorrectas y sus respuestas se centran en la opción correcta y en una opción incorrecta que les parece más verosímil. En la mixta se añade un tercera opción con un porcentaje bajo.
3. Categoría anómala cuando la frecuencia de candidatos que eligen una opción incorrecta es superior a la de los que eligen la opción correcta.

Este modelo de categorización nos permite además de valorar los aciertos descubrir por una parte, el conocimiento parcial de los candidatos al no elegir determinadas opciones y por otro, las carencias que tienen cuando eligen opciones incorrectas.

Los datos del test en su conjunto, es decir, las respuestas correctas (RCs) y las respuestas incorrectas (RIs), son los indicadores del rendimiento de los candidatos y de la calidad del test.

5.3. Resultados

La presentación de los resultados a través de una serie de figuras y tablas sintetizan el comportamiento de los alumnos de una manera plástica (fig 2). Este análisis del test nos muestra un diseño de elección múltiple poco equilibrado. Los porcentajes de la categoría exclusivamente dominante y la dominante +3 deberían estar invertidos, ya que un 10% de los ítems que prácticamente pueden responder todos los candidatos es aconsejable por razones de motivación. Los porcentajes de las categorías dominantes +1 y +2 deberían reducirse a favor de la categoría dominante +3, que debería absorber también a la categoría verdadero / falso.

Tanto los porcentajes de la categoría dominante + 1 como la de verdadero / falso indican que el diseño del ítem, en el que se supone que las tres opciones incorrectas son verosímiles, no es el apropiado. El candidato identifica a dos de ellas como erróneas sin ninguna dificultad, ya sea porque no guardan relación con el ítem ya sea porque el conocimiento parcial del candidato ha sido un factor que no se ha valorado. El ítem se considera anómalo cuando cualquiera de las opciones erróneas obtiene mayor porcentaje de respuestas que la opción correcta.

La distribución absoluta y relativa de las frecuencias permite hacer inferencias más concretas sobre el conocimiento parcial de los candidatos. El objetivo de este trabajo no es presentar un análisis exhaustivo de los datos sino ilustrar el comportamiento de los candidatos en algunos ítems y subrayar la información que se puede obtener de sus conocimientos a través de los aciertos y de los errores. En la tabla, los números nos permiten, analizar detalladamente los datos y hacer inferencias sobre el origen de las dudas y el conocimiento parcial que los candidatos tienen en el ítem 7. Desde un punto de vista holístico los porcentajes totales nos indican que se trata de un ítem idóneo para todos los candidatos, ya que la RC que la eligen un 60,8% está prácticamente dentro de ese rango del 40-60 que propone Fulcher (1997), y las RIs superaran el porcentaje mínimo del 5%. La información se incrementa si se hace una lectura analítica de los porcentajes según el nivel de los candidatos. Los candidatos de nivel muy bajo,

Tabla IX. Item 7.

DOMINANTE+3					
FRACTIL Intervalos de puntuación		1	2	3	4*
0	35	50 (28,4%)	38 (21,6%)	17 (9,7%)	71 (40,3%)
36	43	38 (21,6%)	33 (18,6%)	12 (6,8%	93 (52,8%)
44	49	24 (11,9%)	36 (17,9%)	10 (5%)	131(65,2%)
50	57	7 (4%)	12 (6,9%)	5 (2,9%)	151(88,2%)
TOTAL	%	16,4 %	16,4 %	6,4 %	60,8 %

*Opción correcta

aquellos que en el test no llegan a los 36 puntos, descartan con cierta facilidad la opción 3 y hasta un 40.3% elige la correcta, un ítem muy apropiado, por tanto, para este nivel si atendemos a los porcentajes en cada una de las opciones: . 28,4 – 21,6 – 9,7 y 40,3. Para aquellos candidatos cuyo nivel se considera aceptable, puntuaciones entre 44 y 49, o muy bueno puntuaciones entre 50-57 el ítem tendría un bajo poder discriminatorio. Estos datos nos llevan a concluir que es un ítem apropiado para aquellos alumnos

que sus puntuaciones en el test han sido inferiores a 43 aciertos.

En los gráficos la información no es tan precisa pero es más directa y rápida. Tan sólo presentará tres para ilustrar las categorías señaladas.

La primera impresión en el ítem 13 (fig.4) es que la tercera opción tendría poco poder disuasorio y el resto estaría dentro de parámetros aceptables. La RC queda al nivel del resto de las incorrectas salvo en el grupo de mayor nivel. El trazado de la línea de la opción correcta invierte su tendencia y alcanza porcentajes de aciertos del 60%. Un buen ítem para los candidatos de nivel superior.

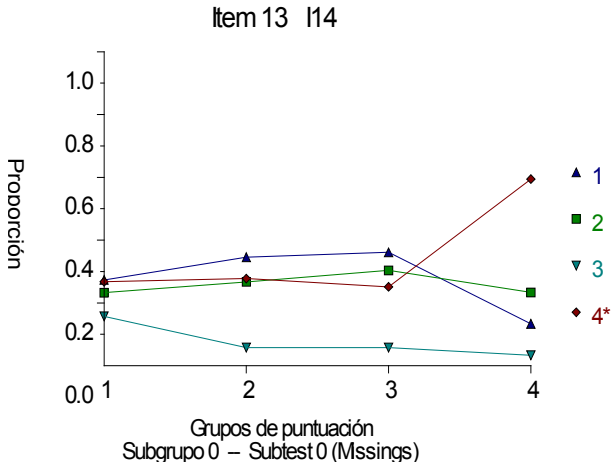

Fig.4

En el siguiente gráfico (fig.5) se observa una tendencia ascendente según los niveles en la RC, las RIs apenas tienen valor disuasorio. Son opciones que el alumno, prácticamente, no considera y que cabría etiquetarlo de poco poder discriminatorio.

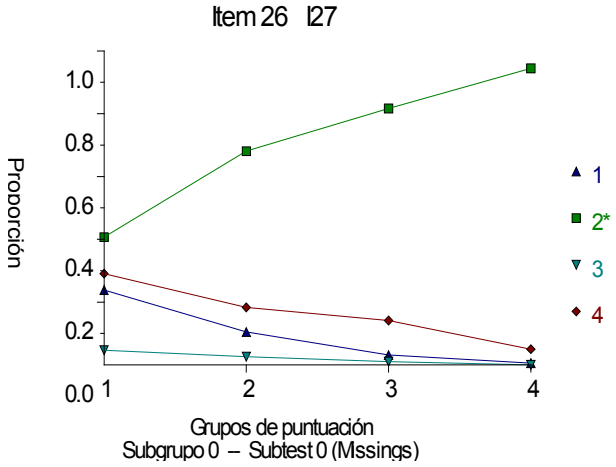

Fig. 5

Finalmente, en la figura 6 se observa como la opción correcta invierte la tendencia de las respuestas con una de las opciones erróneas en el grupo de mayor nivel.

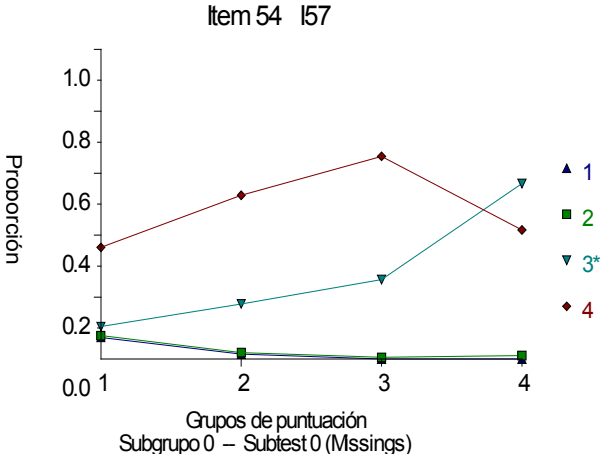

Fig. 6

Los trazados de las opciones 3 y 4 en los que se distribuyen la mayor parte de las respuestas muestran que se trata de un ítem de formato verdadero-falso dentro de un test de elección múltiple.

Las inferencias que se pueden hacer de cada una de estos gráficos tanto desde el punto de vista de la dificultad del ítem y del conocimiento parcial del alumno es de suma importancia. Nos permite conocer los puntos fuertes y débiles de nuestros alumnos y tenerlos en cuenta cuando se diseña un test de elección múltiple.

6. CONCLUSIONES

Este trabajo nos ha permitido avanzar sobre el análisis de la calidad de cada ítem de los distintos modelos de test, la influencia de los diseños, la clasificación de las opciones incorrectas como apropiadas o inapropiadas, los factores que condicionan la calidad del test tales como la familiaridad (Sasaki 2000), la variación y la densidad (Laufer y Nation 1995; Schmitt 2000). Mi objetivo ha sido presentar estos trabajos y a través de invitar a los profesores implicados en la enseñanza de la lengua a que trabajen no sólo con los aciertos sino con el resto de datos. Es un trabajo que merece la pena por la satisfacción personal y por lo que significa para la mejora y calidad de la enseñanza.

7. BIBLIOGRAFÍA

Alderson, J.C. and A. Huhta. 2005. The development of a suite of computer-based diagnostic tests based on the Common European Framework. Language Testing, 22, 3: 3001-319

Babaii, E. y H. Ansary (2001) The C-test. A valid operationalization of reduced redundancy principle? System 29, 209-219.

Babaii, E. y M. J. Moghaddam (2006) On the interplay between test task difficulty and macro-level processing in the C-test. System 34, 586-600.

Eckes, T. y R. Grotjahn (2006) A closer look at the construct validity of C-tests. Language Testing 23 (3) 290-325.

Esteban, M. y H. Herrera (2003) El C-test: instrumento apropiado para la evaluación de la competencia en inglés como lengua extranjera. En G. Luque, A. Bueno y G. Tejada (eds). Las lenguas en un mundo global, pp.71-78. AESLA. Universidad de Jaén.

Esteban García, M. (2005) Niveles de correlación entre el C-test y la prueba de Inglés de Selectividad. En Herrera Soler, H. y J. García Laborda, eds. Estudios y criterios para una Selectividad de calidad en el examen de Inglés, 165-185. Valencia: Ed. Universidad Politécnica de Valencia

Fulcher, G. (1997). An English language placement test: issues in reliability and validity, Language Testing, 14: 113-138.

Herrera-Soler H., M. Amengual, R. Martínez Arias y C. Millar. (2003). Albión: English Evaluation Test. Madrid: EOS.

Herrera Soler, H. 2005. El test de elección múltiple: herramienta básica en la selectividad. En H. Herrera Soler, y J. García Laborda (eds.) Estudios y criterios para una Selectividad de calidad en el examen de Inglés, 165-185. Valencia: Ed. Universidad Politécnica de Valencia.

Jafarpur, A. (1999) Can the C-test be improved with classical item analysis? System 27, 79-89.

Klein-Braley, C. (1997) C-Tests in the context of reduced redundancy testing: an appraisal. Language Testing 14 (1), 47-84.

Laufer, B. (1997) What's in a word that makes it hard or easy: some intralexical factors that affect the learning of words. En N. Schmitt y M. McCarthy (eds). Vocabulary: Description, acquisition and pedagogy, 140-155. Cambridge: Cambridge University Press.

Rashid, S. MD (2002) Validating the C-test amongst Malay ESL Learners. Tunku Mohani Tunku Mohtar, Fatimah Haron y S. Nackeeran, eds. Proceedings of Selected Papers of Fifth Malaysian English Language Teaching Association (MELTA) Biennial International Conference, Petaling Java, Malaysia. [Documento de Internet disponible en www.melta.org.my/modules/sections/12.doc].

Sasaki, M. (2000) Effects of cultural schemata on students' test-taking processes for cloze tests: a multiple data source approach. Language Testing 17 (1), 85-114.

Weir, C.J. 2005. Limitations of the Common European Framework for developing comparable examinations and test.

Language Testing, 22, 3: 281-300.

Zimowski, M.F., Muraki, E., Mislevy, R.J., & Bock, R.D. (2003). BILOG-MG 3.0. Chicago: Scientific Software International

Anexo I

ALTERNATIVE SOURCES OF ENERGY

The world is running out of oil and governments are searching for a suitable alternative, but so far in vain. Many scientists are (1) optimistic that (2) new ways (3) of generating (4) large amounts (5) of energy (6) will be (7) successfully developed (8), but at (9) the same (10) time they (11) fear the (12) consequences. If (13) the world (14) population goes (15) on increasing (16) at its (17) present rate (18), and each (19) individual continue (20) to use (21) more energy (22) every year (23), we will (24) damage the (25) earth´s atmosphere, melt the Artic and Antarctic ice-caps and change the pattern of vegetable and animal life throughout the world – a frightening possibility.

ALTERNATIVE SOURCES OF ENERGY

The world is running out of oil and governments are searching for a suitable alternative, but so far in vain. Many scien_ _ _ _ _ (1) are optim_ _ _ _ _ (2) that n_ _ (3) ways o_ (4) generating la_ _ _ (5) amounts o_ (6) energy wi_ _ (7) be succes_ _ _ _ _ _ (8) developed, b_ _ (9) at t_ _ (10) same ti_ _ (11) they fe_ _ (12) the conseq_ _ _ _ _ _ (13). If t_ _ (14) world popul_ _ _ _ _ (15) goes o_ (16) increasing a_ (17) its pre_ _ _ _ (18) rate, a_ _ (19) each indiv_ _ _ _ _ (20) continues t_ (21) use mo_ _ (22) energy ev_ _ _ (23) year, w_ (24) will dam_ _ _ (25) the earth's atmosphere, melt the Artic and Antarctic ice-caps and change the pattern of vegetable and animal life throughout the world, which is a frightening possibility.

ALTERNATIVE SOURCES OF ENERGY

The world is running out of oil and governments are searching for a suitable alternative, but so far in vain. Many scientists a_ _ (1) optimistic th_ _ (2) new wa_ _ (3) of gener_ _ _ _ _ (4) large amo_ _ _ _ (5) of ene_ _ _ (6) will b_ (7) successfully deve_ _ _ _ _ (8), but a_ (9) the sa_ _ (10) time th_ _ (11) fear t_ _ (12) consequences. I_ (13) the wo_ _ _ (14) population go_ _ (15) on incre_ _ _ _ _ (16) at i_ _ (17) present ra_ _ (18), and ea_ _ (19) individual conti_ _ _ _ (20) to u_ _ (21) more ene_ _ _ (22) every ye_ _ (23), we wi_ _ (24) damage t_ _ (25) earth's atmosphere, melt the Artic and Antarctic ice-caps and change the pattern of vegetable and animal life throughout the world, which is a frightening possibility.

Anexo II

Tutor test

1. I'm glad we had this opp_____ to talk.
2. There are a doz_____ eggs in the basket.
3. Every working person must pay income t_____.
4. The pirates buried the trea_____ on a desert island.
5. Her beauty and ch_____ had a powerful effect on men.
6. La_____ of rain led to a shortage of water in the city.
7. He takes cr_____ and sugar in his coffee.
8. The rich man died and left all his we_____ to his son.
9. Pu_____ must hand in their papers by the end of the week
10. This sweater is too tight. It needs to be stret_____.